Department of Computer Science
Series of Publications A
Report A-2016-2

# Word Associations as a Language Model for Generative and Creative Tasks

Oskar Gross

*To be presented, with the permission of the Faculty of Science of the University of Helsinki, for public criticism in Auditorium XIV, University Main Building, on May $6^{th}$, 2016 at noon.*

University of Helsinki
Finland

**Supervisor**
Hannu Toivonen, University of Helsinki, Finland

**Pre-examiners**
Tony Veale, University College Dublin, Ireland
Krista Lagus, University of Helsinki, Finland

**Opponent**
Timo Honkela, University of Helsinki, Finland

**Custos**
Hannu Toivonen, University of Helsinki, Finland

**Contact information**

Department of Computer Science
P.O. Box 68 (Gustaf Hällströmin katu 2b)
FI-00014 University of Helsinki
Finland

Email address: info@cs.helsinki.fi
URL: http://cs.helsinki.fi/
Telephone: +358 2941 911, telefax: +358 9 876 4314


ISSN 1238-8645
ISBN 978-951-51-2089-2 (paperback)
ISBN 978-951-51-2090-8 (PDF)
Computing Reviews (1998) Classification: I.2 I.2.7
Helsinki 2016
Unigrafia

**Word Associations as a Language Model for Generative and Creative Tasks**

Oskar Gross

Department of Computer Science
P.O. Box 68, FI-00014 University of Helsinki, Finland
oskar.gross@cs.helsinki.fi
http://www.cs.helsinki.fi/u/ogross/

PhD Thesis, Series of Publications A, Report A-2016-2
Helsinki, April 2016, 60+10+54 pages
ISSN 1238-8645
ISBN 978-951-51-2089-2 (paperback)
ISBN 978-951-51-2090-8 (PDF)

**Abstract**

In order to analyse natural language and gain a better understanding of documents, a common approach is to produce a language model which creates a structured representation of language which could then be used further for analysis or generation. This thesis will focus on a fairly simple language model which looks at word associations which appear together in the same sentence. We will revisit a classic idea of analysing word co-occurrences statistically and propose a simple parameter-free method for extracting common word associations, i.e. associations between words that are often used in the same context (e.g., *Batman* and *Robin*). Additionally we propose a method for extracting associations which are specific to a document or a set of documents. The idea behind the method is to take into account the common word associations and highlight such word associations which co-occur in the document unexpectedly often.

We will empirically show that these models can be used in practice at least for three tasks: generation of creative combinations of related words, document summarization, and creating poetry. First the common word association language model is used for solving tests of creativity – the Remote Associates test. Then observations of the properties of the model are used further to generate creative combinations of words – sets of words which are mutually not related, but do share a common related concept.

Document summarization is a task where a system has to produce a short summary of the text with a limited number of words. In this thesis, we will propose a method which will utilise the document-specific associations and basic graph algorithms to produce summaries which give competitive performance on various languages. Also, the document-specific associations are used in order to produce poetry which is related to a certain document or a set of documents. The idea is to use documents as inspiration for generating poems which could potentially be used as commentary to news stories.

Empirical results indicate that both, the common and the document-specific associations, can be used effectively for different applications. This provides us with a simple language model which could be used for different languages.

**Computing Reviews (1998) Categories and Subject Descriptors:**
I.2 Artificial Intelligence
I.2.7 Natural Language Processing

**General Terms:**
Algorithms, Languages, Experimentation

**Additional Key Words and Phrases:**
language models, text analysis, text summarization, computational creativity

# Acknowledgements

It is strange how different experience can *doing a PhD* be. In 2012 my fellow PhD student Jukka Toivanen and I were attending a winter school in Spain. Regardless of a busy schedule we still found time to drink moderate amounts of wine and discuss about life with other participants. I particularly remember that someone said one evening: "If I knew that doing a PhD feels like **this** I would have never started". I have never felt this way thanks to the amazing people who have been around and supported me during the studies.

First, I would like to thank my supervisor Professor Hannu Toivonen. I wouldn't call Hannu a *supervisor* or a *teacher*, but a mentor. After our discussions I always felt that I have learned something new and understood that often *why* is much more important than *how*. I am very happy that I have had a chance to work in a research group where flexibility, freedom and trust results in world class quality research.

I am extremely thankful to Dr. Sven Laur thanks to whom I got really interested in computer science. Sven introduced me to the world of machine learning, data mining and scientific thinking. Sven's patience and the ability to methodologically guide me through my bachelor and master thesis is astonishing. I would also like to thank Professor Jaak Vilo who gave me an opportunity to move to Finland to start my studies here.

Big thanks to my fellow PhD student Jukka Toivanen for great collaboration, academic and social support, Professor Antoine Doucet for hosting my memorable visits to France and being a great co-author to many of my papers. Of course big thanks go to the members and alumni of the Discovery group – Dr. Ping Xiao, Anna Kantosalo, Olli Alm, Simo Linkola, Khalid Alnajjar, Atte Hinkka, Dr. Laura Langohr and Dr. Alessandro Valitutti for great brainstorming sessions, paper celebrations and casual bar evenings.

I would also like to thank my pre-examiners Dr. Tony Veale and Dr. Krista Lagus and also Marina Kurtén for their comments that helped to improve this thesis a lot.

Being able to study without working is something very few people can enjoy. Thus, I would like to thank the Department of Computer Science at the University of Helsinki, Software Newsroom project in Digile's Next-Media research programme, Security ecosystem in Digile's Data2Intelligence research programme, Concept Creation Technology (ConCreTe) project of EU FP7, Algorithmic Data Analysis (Algodan) Centre of Excellence of the Academy of Finland, Hecse and DoCS for making it possible for me to concentrate on research and to attend academic events all over the world.

My life in Helsinki would have been very different without my (debating) friends Alf, Kaisa, Sara and Valtteri. Thank you for the awesome times both in and between rounds. I would also like to thank my friends in Estonia: Ander, Joosep, Kaarel, Liisa, Martin, Priido, Risko and Siim (only to name a few) for the constant support and maintaining contact. Also big thanks go to Allan who in addition to being a great friend has helped me a lot with visual design.

All this would have been very tricky without the support from my family. I am almost sure all this was a secret plan of my grandmother Professor Emeritus Tiina Talvik who constantly pushed me towards completing the degree. And not a bit less of my gratitude goes to my late grandfather Professor Emeritus Raul Talvik who not only taught me how to use my brain but demonstrated himself how thinking can result in wondrous works and can even take you to the edge of the world.

Special thanks go to my family in Estonia – Katrin, Taivo, Kristiina, Inga, Enno and Sale who have constantly supported me during my university studies. It was Katrin & Taivo and their determinism and patience that helped me through the *more complicated* times of my middle and high school and made it possible for me to even consider higher education. Also very big thanks go to my father Andres and his family: Mairi, Werner, Villem, Karoline and my grandmother Mai for having me over and including me into various interesting activities and helping me whenever needed. I also wish to express my gratitude to Astrid, Tuuli & Raivo and Kadri & Toomas with their families for always treating me as their own and being excellent friends.

When I met my wonderful beloved partner Mariliis and her family I understood that I am an extremely lucky person. It is not just the warmth and joy she brings to my life every day but the way she has been able to support me during the ups and downs of life is outright amazing.

*23rd of March, Baltic Sea*
Oskar Gross

# Contents

# Chapter 1

# Introduction

According to some linguistic theories, one of the foundations of understanding language is an ability to understand the words and the *associations* between words [12, 56]. In this thesis the goal is to statistically study word associations in large masses of text and use these associations for generative and creative tasks.

On a large scale there are two different approaches for computational language analysis – *language-oriented* methods that are tailored to work on specific languages and a *language-independent approach*, for which the goal is to develop methods that are applicable to many languages. The language-oriented methods tend to be more accurate and although they are designed for a specific language, some methods are an inspiration for developing methods for other languages, e.g., the Finnish part-of-speech tagger by Löfberg et al. [62] is based on an English tagger [105]. The language-independent methods work on more languages, but usually they tend to be less accurate or need more human annotated data to give accurate results.

In this thesis we develop a language-independent language model which could be learned from data in an unsupervised way. Often models that assign probabilities to sequences of words are called **language models** [66, 52]. This work focuses on a slightly less general problem which is modelling the tendency of words being related rather than estimating the probability of whole phrases. We will use large masses of text to analyse how words co-occur in the same sentence. The idea of using co-occurrences is, of course, not a new one. As J.R. Firth [28] has famously put it:

> You shall know a word by the company it keeps.

In the past, word association analysis has been used for many linguistic tasks, e.g., automatic word sense disambiguation [45] and synonym detec-

tion [101]. As unsupervised language analysis has shown promising results, we took this approach as the basis for learning our language model.

In general we have divided the word associations into two categories: *common (background) word associations* and *document-specific word associations*. The common association model describes how words are associated in text in general (i.e., words which are intuitively strongly associated, e.g., *glass* and *bulletproof* or *car* and *wheel*). This gives us some information about the semantic relations between different words which could be used in applications that deal with natural language. For instance, by knowing how words tend to be associated, we could separate the different contexts one word could be used in (e.g., *bank* as a *financial institution* or as a *river bank*).

In addition to the common word pairs, we also propose another type of word associations — document-specific associations which are important with respect to a certain document (or a set of documents). These associations should show what new links between concepts are established in the document. In this work, in order to highlight word pairs which are important in a target document, we will use a background information corpus that helps us to statistically down-weigh common word pairs.

The difference between association mining and key-phrase extraction [100] is the generality. The goal of key-phrase extraction is to find sequences of words which are descriptive of the document. In association mining the relations between words are analysed on a more general level as the words associated to each other do not have to be sequential and do not have to form phrases.

Using a background (or reference) corpus is not a novel idea itself. In the past it has been used for keyword extraction by contrasting smaller and larger document set, e.g., [58]. A very similar approach to ours but for keywords is Likey [83], which finds relevant keywords by comparing the frequencies of words in the current corpus and the reference corpus.

But how to recognize that we have indeed extracted associations that are specific to a document? We could define a couple of properties that should apply to document-specific associations:

- Associations that are very strongly related to each other outside the document(e.g., *water* and *land*, *Mick* and *Jagger*) should not score high;

- Associations between document-specific words should score higher than associations between very frequent words and document-specific words;

- If a human read the document and then looked at the word association, and she agreed that this document tries to establish a connection between these two concepts, then the association would score high;

If the former two criteria are quite easy to check then, of course, the idea of human validation is more subjective.

The focus of the thesis is not only on word association analysis. We will also focus on generative and creative linguistic tasks, such as document summarization, creative word combination discovery [71] and poetry generation.

Document summarization is one way of dealing with information overload. Specifically news stories (or the publishing field in general) are quite strongly affected by this phenomenon – hundreds of thousands of news providers and bloggers are publishing news stories daily, making it hard for users to obtain diverse information.

Filtering and summarizing are two possible ways of dealing with the problem. Filtering is extracting documents that are related to a respective topic(s). This approach could do well for topics which are quite narrow (e.g., news related to a certain person or disease), but for broader topics (e.g., soccer, basketball, epidemics), the amount of related documents might still be too big for a human to peruse. Even for relatively narrow topics, the problem of redundancy is still prevalent – many news providers use the same news agencies (e.g. Thomson Reuters, Press Association[1] etc), meaning that in order to get new information the reader still has to go through the same information many times.

Automated document summarization could be of help. In document summarization, computational methods are used in order to create short summaries of documents. News stories is not the only domain where document summarization could be seen as a solution for dealing with information overload, the same applies to, e.g., scientific publications, product reviews, internet forums etc.

By providing methods for solving and creating Remote Associates Tests [71] and for generating poetry, this thesis also contributes to a rapidly growing research field called Computational Creativity. Next we will give a brief introduction to what this field is about.

## 1.1 Computational Creativity

The goal in Computational Creativity research is to model, simulate or enhance creativity using computational methods. These areas could be

[1] https://www.pressassociation.com/

verbal or visual creativity, creative problem solving, or some other area requiring creative skills. The current definition of computational creativity is given by Colton et al. [16]:

> the philosophy, science and engineering of computational systems which, by taking on particular responsibilities, exhibit behaviours that unbiased observers would deem to be creative.

This definition gives insight into the foundations and research goals of the field. In the definition, the "responsibilities" is a broader term, which incorporates, e.g., creative intent, ability to employ aesthetic measures to assess artefacts or ability to invent; and "unbiased" refers to observers that have no prejudice against creative machines.

One of the leading thinkers of creativity, Margaret Boden proposed three different types of creativity [9]:

1. **Combinatorial creativity**, where two or more existing concepts are put together in order to create new concepts, e.g. searching for two words for which some properties are shared, but others are not, such an example could be *plane-car*, where individual concepts are both for transport, but the means of operation are very different (conceptual blending);

2. **Exploratory creativity** is searching for potentially creative artefacts, which are already defined declaratively or procedurally, e.g. traversing word associations to look for possible interesting concepts;

3. **Transformational creativity** is when the rules and ideas from one conceptual space are changed such that the space which we are searching itself changes, e.g. for a word association graph, if the search operation re-interprets the meaning of the associations, for instance by creating new entities and linking them to other words.

The work in this thesis deals with exploratory and combinatorial creativity.

Remote Associates Tests [71] are psychometric tests of creativity, where the subject is presented with three cue words and they have to provide an answer word which is related to all of the cue words. For instance, for cue words *spoon*, *bullet* and *quick* the possible answer word is *silver*. This is a task which could be classified under the exploratory creativity category, as given three words, we need to find a single already existing word which satisfies the constraints.

We also defined a reverse RAT problem, in which the subject is given one *seed* word and the goal is to find $n$ words, which are not associated to each other, but they are associated to the seed word. Solving this problem would require combinatorial creativity, as we are looking for combinations of words, which satisfy a certain constraint.

Although document summarization itself is not traditionally thought of as a creative task, it definitely needs creative skills in order to produce good summaries. A very innovative way of creating summaries might definitely fall under the category of transformational creativity, but in our work we only touch the very surface of creating summaries, namely, the methods we propose *combine* sentences from documents to constitute a summary.

Also in this thesis we briefly cover how word associations could be used for poetry generation. As a very general description – the poems are created by substituting words in already written poems. Again, this approach is combinatorial creativity, where combinations of words are selected from the search space and inserted into the poem to create a summary.

These methods for text generation are quite different from approaches using sequence-modelling (see, e.g., [47]), where the text is generated by choosing words probabilistically in a sequence. Our text generation methods are either replacing words in already written poems or creating text by combining already written sentences, thus our language model is not used for and is not modelling grammar. The model is used for providing information about how words are associated to each other and the sentence structure is achieved by other means.

In the rest of the thesis we will first show how to build the language models and then demonstrate the use of them in generative and creative tasks, but first we will give the contributions of this thesis.

## 1.2 Contribution of this Thesis

The thesis consists of five publications which introduce novel methods for finding document-specific associations and use them for different linguistic and creative tasks. The thesis also contains and introductory part which gives an overview of the themes in the thesis and gives a more elaborate overview of the background of the work.

We have implemented all the methods presented in the publications and obtained empirical results. The publications themselves are as follows.

**PAPER I — Lexical Creativity from Word Associations**, Oskar Gross, Hannu Toivonen, Jukka M. Toivanen, and Alessandro Valitutti. In *Knowledge, Information and Creativity Support Systems (KICSS), 2012*

*Seventh International Conference on*, 35-42, IEEE, 2012.

We demonstrate how the psychometric tests of creativity (Remote Associates Tests [71]) can be solved with word associations found in bigrams. We show empirically that the common word associations do correlate with the relations found in WordNet [76] and use these associations for generalizing the bigram-based methods to a more general setting. Additionally, we propose a methodology for generating creative combinations between words.

The author of the thesis implemented the methods and experiments. The author played a major role in writing the article.

**PAPER II — Software Newsroom – An Approach to Automation of News Search and Editing**, Juhani Huovelin, Oskar Gross, Otto Solin, Krister Linden, Sami Maisala, Tero Oittinen, Hannu Toivonen, Jyrki Niemi, Miikka Silfverberg. In *Journal of Print Media Technology research*, 2(3), 141-156, IARIGAI, 2013.

This paper was the joint work of a larger group and the outcome of a project in collaboration between many universities and companies. The author played a major role in writing Sections 2.4.2–2.4.4, 3.2 and also participated in the writing of the introduction and conclusions of the article.

The before-mentioned Sections focus on using word association analysis on social media data. In this publication we give examples of how the word associations could be visualized and outline early work on the summarization system later published with details and experimental results in PAPER III.

**PAPER III — Document Summarization Based on Word Associations**, Oskar Gross, Antoine Doucet, and Hannu Toivonen. In *Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval*, 1023-1026, ACM, 2014.

We propose a novel unsupervised Association Mixture Text Summarization method, an improved model for finding document-specific associations. Document-specific associations are used to generate extractive summaries from a set of documents.

The author played a major part in implementing and writing the paper

**PAPER IV — Language-Independent Text Summarization with Document Specific Word Associations**, Oskar Gross, Antoine Doucet, and Hannu Toivonen. In *Proceedings of 31st ACM Symposium on Applied Computing, Accepted for Publication*, ACM, 2016

This paper provides a methodology for using the document-specific associations for creating summaries in different languages. The paper builds on the ideas in PAPER III and extends the method by introducing a graph-

based component to the summarization method and provides analysis on how the method works for 9 different languages.

A major part of experimenting and writing was done by the author.

**PAPER V — The Officer Is Taller Than You, Who Race Yourself! Using Document Specific Word Associations in Poetry Generation**, Jukka M. Toivanen, Oskar Gross, Hannu Toivonen. In *The Fifth International Conference on Computational Creativity*, 355-362, Josef Stefan Institute, Ljubljana, Slovenia, 2014.

This paper presents a method for generating poetry about a single news story by using a corpus-based poetry generation method [96] as a building block.

The author of the thesis had a major part regarding implementations of the word association models. All the authors contributed equally to writing the paper. This publication is also used in the Ph.D. thesis by Jukka Toivanen.

## 1.3 Structure of the Thesis

So far we have covered the motivation and the overall ideas in the thesis. The rest of the thesis is divided into four parts: word associations models, generative tasks, creative tasks and conclusions. As the word association models are the main building block for our methods, in Chapter 2 we will take a closer look at word association methods (largely based on the work by Dunning [22]) and propose a novel method for extracting document-specific associations.

Chapter 3 will cover how the document-specific associations could be used for document summarization and Chapter 4 touches on how word associations could be used in tasks that require lexical creativity. The creative tasks we will look at are related to tests of creativity and poetry generation. The introductory part of the thesis is concluded in Chapter 5.

# Chapter 2

# Word Associations

In this chapter we will cover the methods for finding word associations which we use as basic components for the generative and creative tasks. We will cover two types of word associations, but the main emphasis is put on the *document-specific word associations* – the associations which capture the important word associations of a certain document. Before introducing our methods we will give an overview of related work.

## 2.1 Background

Language models that are based on word co-occurrence analysis are not a new idea. The co-occurrence analysis could be divided into two ideas – *distributional semantics* and *direct co-occurrence analysis*. Distributional semantics aims to find semantic similarities between linguistic items in large masses of text [41] by estimating word similarity via shared co-occurrence with other words. On the other hand, direct co-occurrence analysis estimates the similarity of two words by using their co-occurrence counts as a reference. We will start by taking a look on how relations between words could be defined and then we will give an overview of the ideas in distributional semantics and direct co-occurrence analysis.

### 2.1.1 Finding Relations

There are a couple of ways for finding common relations between words. Here by relations we do not mean just mere statistical associations between words, but rather associations that have a meaning (e.g., the *part-of* relation indicates that some concept is part of another one, e.g., an *engine* is part of a *car*). This kind of relations could be created manually, semi-manually and automatically. Of course there are trade-offs. For instance,

manually created systems have a lower error rate, but at the same time the coverage of the relations is smaller. The semi-manual systems use manually generated relation databases as a training set and use this information in order to find more relations, e.g., Alfonseca et al. propose a method for extending ontologies with domain-specific knowledge [1]. The goal of fully automatic systems is to infer the relations directly from the natural language.

An example for a manually curated lexical database is WordNet [76], which contains different strictly hierarchical relations between words. Originally WordNet is in English, but it has been translated to many languages (e.g. Italian [2], Russian [3] and Romanian [99] to name a few). The words are organised into synsets, sets of words which are considered to be synonyms. The relations in WordNet are defined to be between these synsets. Between noun synsets, the defined relations are *hypernyms, hyponyms, coordinate terms, meronyms* and *holonyms*. For instance, if Y is a hypernym of X, this means every X is a kind of Y, e.g. *mammal* is a hypernym of *carnivore*. For the explanation of other types of relations we refer to Miller [76].

Cyc [60] is a hand-built inference tool for AI which contains concepts and facts, which could be manipulated in order to produce new knowledge. In addition to rules, Cyc also has an inference engine, which is able to perform general logical deduction.

Although the manually curated lexical databases are very useful and have been used in numerous different applications successfully, automatically harvesting relations between words is still an interesting research problem. Even though WordNet has been translated into many languages, constructing this kind of information automatically would make language processing systems more language independent.

**Log-Likelihood Ratio** In this thesis, we are not focusing so much on trying to harvest semantic relations, but rather on mere statistical associations between words. The log-likelihood ratio is a very general non-parametric statistical test use in many areas. Using a log-likelihood ratio for word co-occurrence analysis was proposed by Dunning [22] who showed, in particular, that the log-likelihood ratio does not overestimate the importance of very frequent word associations like some other measures.

Our proposed methods, introduced later in more detail, are built on the log-likelihood ratio measure. Log-likelihood ratio is a measure of direct co-occurrence, whereas Latent Semantic Indexing [19] or Word2Vec [75] are not (see below). Log-likelihood ratio is not the only direct co-occurrence

measure, another examples are e.g., mutual information, pointwise mutual information [14] and Jaccard index [46].

**Latent Semantic Analysis** Latent Semantic Analysis [19] aims to find a set of concepts (instead of terms) in a corpus using singular value decomposition. The semantic similarity (relatedness) of two words can then be estimated by comparing them in the concept space.

For illustrating the idea, consider a set of documents $D = \{d_1, \cdots, d_n\}$ and a set of terms $T = \{t_1, t_2, \cdots, t_m\}$. First we generate a matrix $F^{m \times n}$, where the columns of the matrix represent the documents and rows represent the words.

The simplest form of representing observation of a term in a document is in binary form, where $F_{i,j} = 1$ represents that a word $t_i$ is found in document $d_j$ and $F_{i,j} = 0$ that it is not.

The problem with using a binary approach is that it ignores the frequency or relative frequency of the terms in the document and thus does not represent the importance of the terms in the document well. To reduce this problem many approaches have been proposed for more accurate weighting of terms occurrences in the matrix (e.g. absolute or relative word frequency [63], tf-idf [89], entropy-based measures [21], pointwise mutual information [14] etc).

Given that we have the populated matrix $F$, the next step is to apply Singular Value Decomposition on the matrix:

$$F = USV,$$

where $U$ is $m \times m$, $S$ is a diagonal $m \times n$ matrix and $V$ is a $n \times n$ matrix. The values along the diagonal in $S$ are called the singular values, and the vectors in $U$ and $V$ are called left and right singular vectors, respectively. When selecting $k$ largest singular values and the corresponding vectors from $U$ and $V$ then we get the $k$-rank approximation to $F$ with the smallest error:

$$F^\star = U^{m \times k} S^{k \times n} V^{n \times k}.$$

This approximation creates a latent space $F^\star$ which is less noisy and is expected to contain merged dimensions with terms of similar meanings. The reduced latent space is used further to analyse how words relate to each other in the new space.

Latent semantic analysis has then evolved to *Probabilistic Latent Semantic Analysis* [43] and later to *Latent Dirichlet Allocation* [8], a Bayesian inference approach.

**Word2Vec** Word2Vec [75] is a multi-component method for learning associations between words. The learning of the word relations is done by a *shallow* neural network. The power of the method comes from the fact that the neural network-based approach is combined with a subsampling of frequent words. This combination gives the system the ability to process billions of words in a matter of day. The word2vec system is revolutionary in the sense that it makes possible simple yet introspective arithmetic between word vectors. As an example brought by Mikolov et. al. [75], the result of the vector calculation *vec('Madrid')-vec('Spain')+vec('France')* results in a vector which is closest to the vector representation of *'Paris'*. This effect demonstrates that the system is implicitly able to capture information about words on a rather deep level.

The above is not an exhaustive list of different available methods in distributional semantics, but rather an overview of the most frequently used ones.

### 2.1.2 Networks of Concepts

The word associations can be represented as a graph of words where the nodes are the words and the edges are the relations between these words. The graphs could be either un-weighted or weighted. In the first case all the relations between words would be treated as equal, in the other case, the relations have unequal strengths. Intuitively it makes more sense that some connections are stronger than others. For instance, consider the word *airplane*. Although the word *seat* is definitely related to planes, intuitively the word *wing* would have a much stronger relation. Also, it is possible to use edge labels that define how two words are related to each other (e.g., synonym/antonym relation).

Network formalisms were adopted relatively early in the history of artificial intelligence to represent and reason with information. Semantic networks have their roots in describing human learning and memory [91]. They come in several forms, but they are usually formal to allow semantic inference and reasoning. The Semantic Web can also be seen as a form of a semantic network. Its focus is on sharing and reuse of information in the web.

Numerous attempts have been made to automatically construct networks of concepts. Some are based on information extraction or other relatively involved natural language-processing methods, and aim to extract annotated relation types. For instance, ASKNet [40] analyses text structure and generates a labelled word relation network. ASKNet and spreading activation were effectively used for detecting semantic relatedness between

words.

Methods for more formal use tend to be even more complex. For instance, Lenat et al. [69] use a combination of Cyc and WWW to assist in entering new knowledge into Cyc.

Word co-occurrence networks are often used together with other lexical resources. For instance, Navigli [80] uses a semantic network-based approach for word-sense disambiguation, where different machine-readable dictionaries and WordNet [76] were translated into a graph. Then the word-sense disambiguation task was solved by analysing the cycles and quasi-cycles in the corresponding network.

Roark et al. [88] use co-occurrence statistics for lexicon construction. First, a small set of examples is given as an input to the algorithm, then the co-occurrence statistics are used to detect potential entries for a word category.

## 2.2 Notation

Before giving insight into the common and document-specific word association methods, we introduce the notation which we will use in the rest of the thesis.

Let $T$ denote the set of all words. Our methods do not take into account the order of words in a sentence, order of sentences in a document or the order of documents in a corpus. For this reason, we will treat a sentence $s$ as a set of words $s = \{t_1, \cdots, t_k\} \subset T$, a document $D$ as a set of sentences $D = \{s_1, \cdots, s_n\}$ and a set $S$ of many documents as one long document $D_s = \bigcup_{D \in S} D$.

We say that two words $t_i$ and $t_j$ *co-occur* in sentence $s$ if they both are in $s$. Given a document $D$, let $n_{ij}$ denote the number of sentences in which $t_i$ and $t_j$ co-occur, i.e.,

$$n_{ij} = |\{s \in D \mid t_i \in s \text{ and } t_j \in s\}|.$$

To estimate if the number of co-occurrences is statistically interesting, we also need to know how often words $t_i$ and $t_j$ occur separately and how many sentences do not contain either one. For document $D$, we denote by $n_{i-j}$ the number of sentences that contain $t_i$ but not $t_j$, and by $n_{-ij}$ the number of sentences that do not contain $t_i$ but contain $t_j$. The number of sentences where neither of the terms appear is denoted by $n_{-i-j}$. Furthermore, we denote by $n_{i\cdot}$ the total number of sentences in which the word $t_i$ occurs and by $n_{j\cdot}$ the total number of sentences in which the word $t_j$ occurs. Let $n$ denote the total number of sentences in $D$.

## 2.3 Log-Likelihood Ratio

In our work, the basis for finding the common and document-specific associations is the log-likelihood ratio [22]. The log-likelihood ratio test is a statistical test which is used to compare two different models. We use log-likelihood ratio as it is a standard and computationally simple method. Additionally it provides a natural framework to calculate the document-specific associations, as discussed later.

The log-likelihood ratio test is based on comparing the likelihood of an event under two models; the *null model* is a model for which the parameter space is constrained and the *alternative model* for which the parameter space is relaxed. For instance, in our case this means that for the *null* model we expect occurrences of words to be independent, whereas for the *alternative* model we estimate the probability of co-occurrence from the corpus.

Formally, consider two models $M_{null}$ and $M_{alt}$, where the former is the *null* model and the latter the *alternative* model and let $L(M)$ denote the likelihood of an event under the model $M$.

The log-likelihood ratio expresses how many times the data is more likely under one model than the other:

$$LLR = -2 \log \frac{L(M_{null})}{L(M_{alt})}. \tag{2.1}$$

The log-likelihood ratio test can only be used to compare two nested models – this means that the more complex model can differ from the simpler model by additional parameters. We will use this ratio in order to find either common word associations or document-specific associations. The main difference between these two is the way of defining the null and the alternative model.

## 2.4 Common Word Associations

By common word associations we mean associations between two words for which a human would recognize them to be related to each other. An example of such associations would be *Obama* and *USA*, *car* and *tyre* and *bank* and *finance*. As the examples indicate, we are not only looking for words which appear together in very close proximity (e.g. 2-3 words apart), but rather tend to appear in the same context. A common problem in word co-occurrence analysis is how to set the context or window size.

One way would be to define a *sliding window* of length $n$ and treat words in this proximity as co-occurring. The downside of the approach is,

that the optimal context size may vary for different languages or different types of text. On the other hand, the possibility of choosing a window size gives users more options for what kind of associations to consider.

As the words that co-occur in the same sentence tend to have stronger relation to each other than to words in other sentences [76], in our methods we will use a sentence as the context size. Although this choice might not be optimal, the benefit is that the parameter does not need to be tuned for different languages. The drawback of course is, that the assumption is quite strong, e.g., for composite sentences, we might associate words which actually are not related at all. Also, when looking at co-occurrences only on sentence level we lose the associations between words which appear in two strongly related consecutive sentences.

We are looking for such word pairs, which tend to co-occur more than we would expect. In order to use the log-likelihood ratio test, we would first need to define the two models $M_{null}$ and $M_{alt}$. We will use the multinomial distribution with categorical variables as the underlying model for estimating the likelihood of word co-occurrences. Given $n$ independent trials and $k$ different categories of outcomes the multinomial distribution gives the probability for all different combinations of outcomes. The multinomial distribution is defined by the respective probabilities for each of the $k$ outcomes and the absolute number of outcomes for each of the $k$ categories.

In order to illustrate our proposed methods, we will use two running examples in this thesis. The documents we use can be found at the end of the thesis in Appendix A and B. The articles are "Intelligent Machines: Do we really need to fear AI?" (Appendix A) and "Medical Scientist Proves Hypothesis Set by Lennart Meri" (Appendix B). As an example let us consider the article in Appendix A and the words *military* and *leaders*. In the context of the multinomial model each sentence is one trial. The possible outcomes for each sentence are the following:

- Both words *military* and *leaders* appear in the sentence;
- The word *military* appears, but the word *leaders* does not appear in the sentence;
- The word *leaders* appears, but the word *military* does not appear in the sentence;
- Neither of the words appears in the sentence.

All these events are mutually exclusive from each other and one of these events has to be true for each trial. The counts for the word pair can be found in Table 2.1. Column $N$ gives the counts for different combinations

| **Occurrences** | **N** | **P** | **Q** |
|---|---|---|---|
| **military& leaders** | 2 ($n_{ij}$) | 0.001 ($p_{ij}$) | 0.03 ($q_{ij}$) |
| **military& ¬leaders** | 0 ($n_{-ij}$) | 0.033 ($p_{-ij}$) | 0 ($q_{-ij}$) |
| **¬military& leaders** | 0 ($n_{i-j}$) | 0.033 ($p_{i-j}$) | 0 ($q_{i-j}$) |
| **¬military& ¬leaders** | 56 ($n_{-i-j}$) | 0.933 ($p_{-i-j}$) | 0.97 ($q_{-i-j}$) |
| **military** | 2 ($n_{i\cdot}$) | 0.034 ($p_{i\cdot}$) | 0.034 ($q_{i\cdot}$) |
| **leaders** | 2 ($n_{\cdot j}$) | 0.034 ($p_{\cdot j}$) | 0.034 ($q_{\cdot j}$) |
| **Total sentences:** | 58 | — | — |

Table 2.1: The counts for the co-occurrences of words *military* and *leaders*. The probability parameters **P** represent the probability under the assumption that the words *military* and *leaders* are statistically independent and **Q** represents the maximum likelihood parameters

of the word co-occurrences and column $P$ gives the probabilities calculated under the assumption that *military* and *leaders* are statistically independent (see further).

The multinomial model takes two sets of parameters: $N$ which are the observed counts of the word pair in the corpus and $P$, which are the corresponding probabilities for the events.

In order to find the common word associations, under the $M_{null}$ model we obtain the probability parameters $P$ under the assumption that two words are appearing statistically independently, whereas for the alternative $M_{alt}$ model we estimate the probabilities $Q$ via maximum likelihood. When the likelihood of the co-occurrence of the word pair is more likely in the corpus than it would be under the independence assumption, we say that these two words are statistically associated.

More formally, for the model $M_{null}$ we give the following probability parameters (cf. Table 2.1): the probability of words $t_i$ and $t_j$ co-occurring is denoted by $p_{ij}$, the probability of the words occurring without each other by $p_{-ij}$ and $p_{-ji}$, respectively. The probability of neither of the words occurring in a sentence is denoted by $p_{-i-j}$.

Let $q_{ij}$, $q_{-ij}$, $q_{i-j}$, $q_{-i-j}$ similarly denote the probabilities for the $M_{alt}$ model. The difference for the two models $M_{null}$ and $M_{alt}$ comes from how the probability parameters are obtained. For the null model the probabili-

ties $p_{ij}, p_{-ij}, p_{i-j}, p_{-i-j}$ are estimated under the independence assumption:

$$\begin{aligned} p_{-ij} &= (1 - n_{i\cdot}/n) \cdot n_{\cdot j}/n \\ p_{i-j} &= n_{i\cdot}/n \cdot (1 - n_{\cdot j}/n) \\ p_{ij} &= n_{i\cdot}/n \cdot n_{\cdot j}/n \\ p_{-i-j} &= 1 - p_{ij} - p_{-ij} - p_{i-j}. \end{aligned}$$

For the alternative model the parameters are obtained by using the maximum likelihood calculation:

$$\begin{aligned} q_{ij} &= n_{ij}/n \\ q_{-ij} &= n_{-ij}/n \\ q_{i-j} &= n_{i-j}/n \\ q_{-i-j} &= n_{-i-j}/n. \end{aligned}$$

Now, by using the parameters and the word co-occurrence counts, the likelihood of the word pair $t_i$ and $t_j$ for the respective models is given in the form:

$$L(M_{null}) = \binom{n_{ij} + n_{-ij} + n_{i-j} + n_{-i-j}}{n_{ij}, n_{-ij}, n_{i-j}, n_{-i-j}} p_{ij}^{n_{ij}} p_{-ij}^{n_{-ij}} p_{i-j}^{n_{i-j}} p_{-i-j}^{n_{-i-j}} \tag{2.2}$$

$$L(M_{alt}) = \binom{n_{ij} + n_{-ij} + n_{i-j} + n_{-i-j}}{n_{ij}, n_{-ij}, n_{i-j}, n_{-i-j}} q_{ij}^{n_{ij}} q_{-ij}^{n_{-ij}} q_{i-j}^{n_{i-j}} q_{-i-j}^{n_{-i-j}}. \tag{2.3}$$

In order to measure the strength of the association between the words $t_i$ and $t_j$, we use the log-likelihood ratio as given in Equation (2.1). The multinomial coefficients cancel out and the log-likelihood ratio calculation becomes:

$$LLR(M_{null}, M_{alt}) = -2 \ln \frac{p_{ij}^{n_{ij}} p_{-ij}^{n_{-ij}} p_{i-j}^{n_{i-j}} p_{-i-j}^{n_{-i-j}}}{q_{ij}^{n_{ij}} q_{-ij}^{n_{-ij}} q_{i-j}^{n_{i-j}} q_{-i-j}^{n_{-i-j}}}.$$

This is a standard result for obtaining the log-likelihood ratio for word co-occurrences. The $LLR$ is used to measure the strength of the association where higher values indicate stronger statistical association between words.

As an example consider the article found in Appendix B. By applying word co-occurrence analysis, we will obtain the word pairs found in Table 2.2. Although the article is short, we can already see how some word associations (e.g., *medical* and *scientist*) which we perceive to be associated get higher scores than other words. In the table we have highlighted a word pair which is important later, w.r.t. the document-specific associations.

Next, we will see how some changes to the previous models make it possible to extract document-specific associations.

| **Word A** | **Word B** | **LLR** |
|---|---|---|
| iceland | britain | 19.62 |
| four | wrote | 17.74 |
| fan | mythical | 17.74 |
| near | norway | 17.74 |
| stone | stumbling | 17.74 |
| medical | scientist | 17.74 |
| ⋮ | ⋮ | ⋮ |
| were | talvik | 0.74 |
| have | talvik | 0.74 |
| **talvik** | **saaremaa** | 0.74 |
| ⋮ | ⋮ | ⋮ |

Table 2.2: Common word associations extracted from the article in Appendix B.

## 2.5 Document-Specific Associations

In the previous section we described how we could find the general word co-occurrence statistics. However in some cases, e.g., when we want to get a quick overview of a document, we are not that much interested in how the concepts are related to each other generally in language, but rather how the concepts are related in specific documents or domains.

A natural approach would be to apply the *common word association method* within the given document(s) in order to find such associations. Let us consider the example article we have in Appendix A. We applied the common word association method on the article and the top 10 strongest observed word pairs are given in Figure 2.1. Notice that the word associations are not very specific to the document. Although the words do represent important themes of the articles, the associations themselves are common (e.g., *military leaders* and *book author* are very strong and common associations).

In order to put emphasis on the associations which are specific to a document we could down-weigh associations which tend to be common anyway and give higher weight to associations which are novel with respect to the associations which we already know. For instance, if you consider the word associations found in Table 2.2, downweighing word associations like *stumbling* and *stone* or *medical* and *scientist* would make sense.

To achieve that, we introduce extra information to the $M_{null}$ model, in order to get a better estimation of how the word pairs tend to behave. First

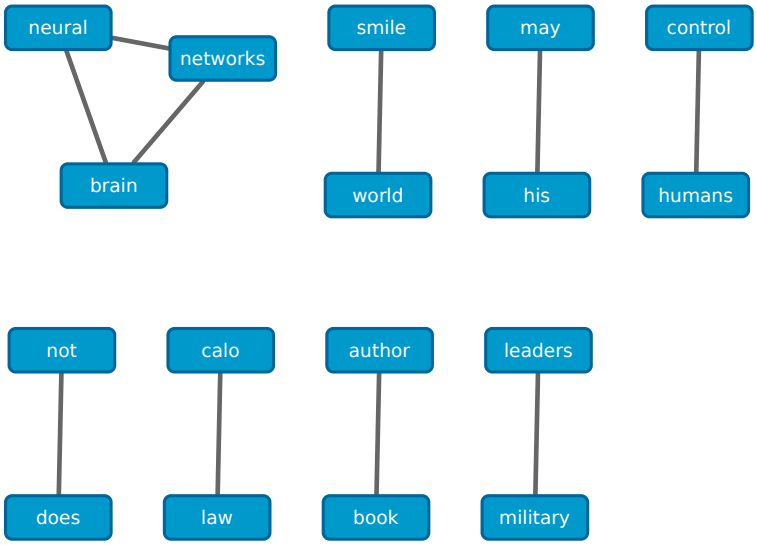



Figure 2.1: Common word associations extracted from the article given in Appendix A.

we collect a large amount of information on how word pairs tend to co-occur in a large mass of text and then use this information to give higher weight to word associations which co-occur in the document more frequently than in texts in general.

We introduce a background corpus $\mathcal{B}$, which contains a large amount of documents (e.g. Wikipedia, Project Gutenberg, Reuters Corpus etc). One should choose the background corpus to be similar to the documents used for document-specific association extraction. The reason for this is, that different types of documents might use different language, phrases and also different topics might be prevalent. For instance even if *Barack Obama* is quite prevalent in news stories, it is not present in the Project Gutenberg corpus.

Next, let us consider a document $D$ from which we would like to extract the word associations. Essentially the idea is, that the $M_{null}$ model is enriched with information from $\mathcal{B}$ and $M_{alt}$ is obtained from document $D$. Consider that the background $\mathcal{B}$ has a total of $n$ sentences, for each word pair we obtain the counts $n_{ij}$, $n_{-ij}$, $n_{i-j}$, $n_{-i-j}$. Similarly consider document $D$ having $m$ sentences and the counts $m_{ij}$, $m_{-ij}$, $m_{i-j}$, $m_{-i-j}$ are obtained for each word pair.

The simplest way to now compare the importance of a word pair $t_i$

and $t_j$ is to estimate the $M_{null}$ model from the background corpus $\mathcal{B}$. We could obtain probability parameters directly from the counts: $p_{ij} = n_{ij}/n$, $p_{i-j} = n_{i-j}/n$, $p_{-ij} = n_{-ij}/n$, $p_{-i-j} = n_{-i-j}/n$. A straightforward way to obtain the probabilities of the occurrences from the document $D$ is:

$$\begin{aligned} q_{ij} &= m_{ij}/m \\ q_{i-j} &= m_{i-j}/m \\ q_{-ij} &= m_{-ij}/m \\ q_{-i-j} &= m_{-i-j}/m \end{aligned} \tag{2.4}$$

Given that we have defined the $M_{null}$ and $M_{alt}$ model, the next natural step is to find the log-likelihood ratio between these two models:

$$LLR(M_{null}, M_{alt}) = -2 \ln \frac{p_{ij}^{m_{ij}} p_{-ij}^{m_{-ij}} p_{i-j}^{m_{i-j}} p_{-i-j}^{m_{-i-j}}}{q_{ij}^{m_{ij}} q_{-ij}^{m_{-ij}} q_{i-j}^{m_{i-j}} q_{-i-j}^{m_{-i-j}}}.$$

Again, in order to illustrate how well this approach works, consider the document in Appendix A. For this example, we use Reuters and Brown corpus as the background corpus $\mathcal{B}$. By using the model above the obtained top 15 pairs can be seen in Figure 2.2. The results are not very convincing – it is easy to notice 12 pairs out of 15 contain the word *human* or *humans*. Moreover, 9 pairs out of 15 contain a very common word (other than human/humans) as a member of the pair.

*But why?* This is due to the fact that the word *human* is not very prevalent in the background corpus (i.e. appears approx 200 times). Since it is frequent in the document, it also has many frequent co-occurrences with other words just by chance (e.g., *human* and *for*). When compared to the background corpus $B$, these stand out. The same applies to the word *robots*, which is not encountered in the background corpus and is thus creating pairs with quite common words.

A similar effect occurs when we have seen both of the words before but they never co-occurred. Intuitively this makes sense – if we have observed words, say *car* and *jackson* very many times but we have never observed them in the same sentence, then the event of their co-occurrence is unexpected. In other words the problem is that when we encounter a word pair $(t_i, t_j)$ in $D$ which we have never encountered in $\mathcal{B}$ we will overestimate the importance of this pair.

In order to reduce this problem, we introduce Mixture Model in PAPER III— a model which collects information from the background corpus and incorporates additional information from document $D$. If we merged the counts of the word pairs in the document directly to the background model we would also emphasize the association in the background as we do in the

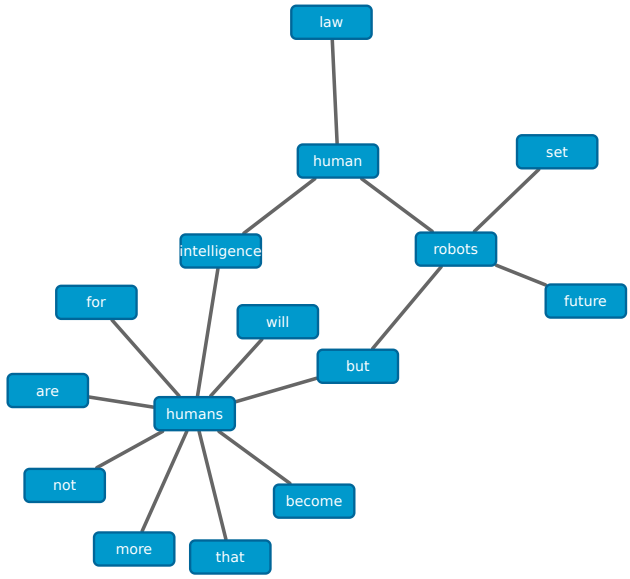



Figure 2.2: Word associations extracted from the article given in Appendix A by comparing the document associations directly to the background associations

document. To avoid this we add the information to the $M_{null}$ model under the independence assumption. This means, that instead of adding $m_{ij}$ to $n_{ij}$, the co-occurrence counts are updated as follows:

$$\begin{aligned} n'_{ij} &= n_{ij} + (m_{i\cdot} \cdot m_{\cdot j})/m \\ n'_{-ij} &= n_{-ij} + ((m - m_{i\cdot}) \cdot m_{\cdot j})/m \\ n'_{i-j} &= n_{i-j} + (m_{i\cdot} \cdot (m - m_{\cdot j}))/m \\ n'_{-i-j} &= n + m - n'_{ij} - n'_{-ij} - n'_{i-j} \\ n' &= n + m. \end{aligned}$$

Essentially what we do is that we add the expected number of occurrences in document $D$ to the background by assuming independence between the terms in the document. The final equation defines the total number of sentences in the background $\mathcal{B}$ and document $D$.

After this modification the probability parameters for the null model $M_{null}$ are obtained as before — $p_{ij} = n'_{ij}/n'$, $p_{-ij} = n'_{-ij}/n'$, $p_{i-j} = n'_{i-j}/n'$, $p_{-i-j} = n'_{ij}/n'$. The parameters for the alternative model are obtained exactly as before (Equation (2.4)). Now again, the document-specific associations are calculated for by using log-likelihood ratio (Equation (2.1)).

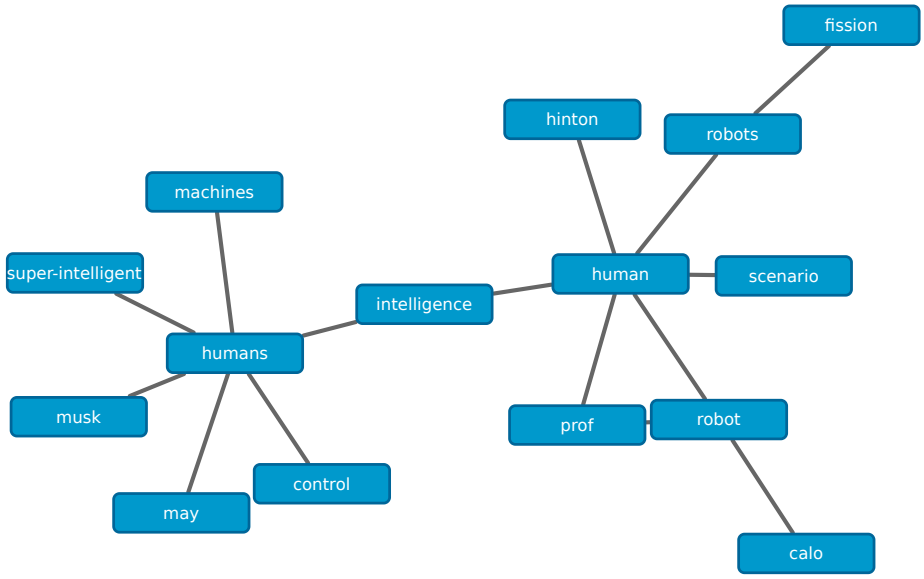



Figure 2.3: Document-specific word associations extracted with Mixture Model from the article given in Appendix A.

In addition to the Mixture Model approach, it would also be possible to use smoothing for reducing some random effects. In order to keep our methods simple, we have not implemented smoothing into our methods, but potentially methods like, e.g., [25, 85] could be used also for our method.

For illustrating the results we use the document in Appendix A with the same background corpus as before. The associations obtained with the Mixture Model can be seen in Figure 2.3. Notice, how the pairs have become a little bit more diverse. Although the word *human* or *humans* is still very prevalent, instead of very common words (e.g., *but*, *and* and *of*), we observe much informative words, e.g., *intelligence*, *machines*, *prof* or *musk*.

A way to illustrate the effect of the documents in the background would be to analyse the same document and compare the pairs with two different backgrounds – for one the background contains only general information and for the other background it also contains some more specific information about the domain. For this, consider the article in Appendix B, a news story about a recently published book which reconstructs the voyage of Pytheas, an antique Greece traveller. The book gives reasons why Ultima Thule is most probably Saaremaa, an island in Estonia.

Firstly we calculated the pairs using the Mixture model and Reuters

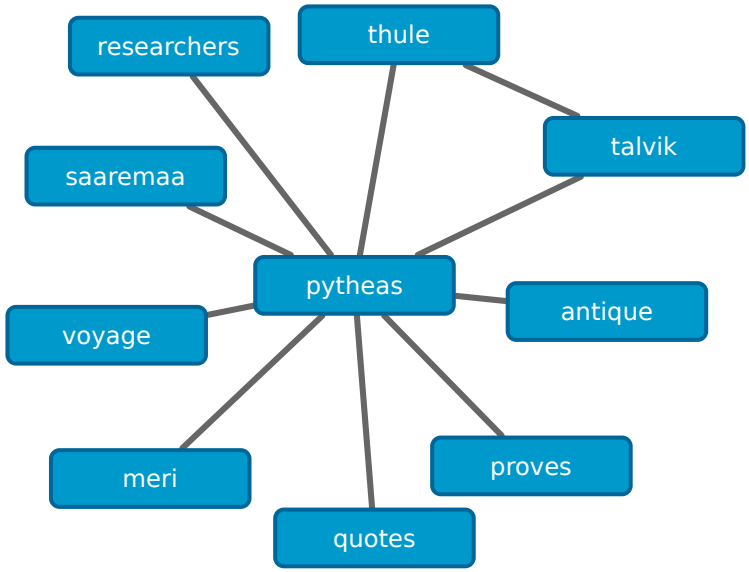



Figure 2.4: Document-specific word associations extracted with Mixture Model from the article given in Appendix B.

and Brown corpus as the background. The results for the top 10 pairs can be observed in Figure 2.4. As we can see, Pytheas is a very central character of the news story and 9/10 associations contain the respective word.

Secondly, to have more specific information about Pytheas in the background we included the text from the Wikipedia page about Pytheas[1] to the background corpus. The top 10 extracted pairs with the new background are presented in Figure 2.5. Observe, how some relatively unimportant pairs were down-weighted. For instance, if *pytheas* and *researchers* were associated before, then for the *more knowledgeable* background the pair does not appear. Similarly many other words have been down-weighed and instead of Pytheas, Talvik has become the central of the story, as some of the previous pairs (in Figure 2.4) were rather common w.r.t. Pytheas. The news story could be summarized quite well with a sentence: "Raul Talvik wrote a book which builds on the hypothesis set by Lennart Meri, that the island described by Pytheas as Ultima Thule, is actually Saaremaa". By looking at the top pairs in Figure 2.5, it indicates that it captures the essence of the document rather well

Also the empirical results of text summarization (Chapter 3) will show

[1]https://en.wikipedia.org/wiki/Pytheas

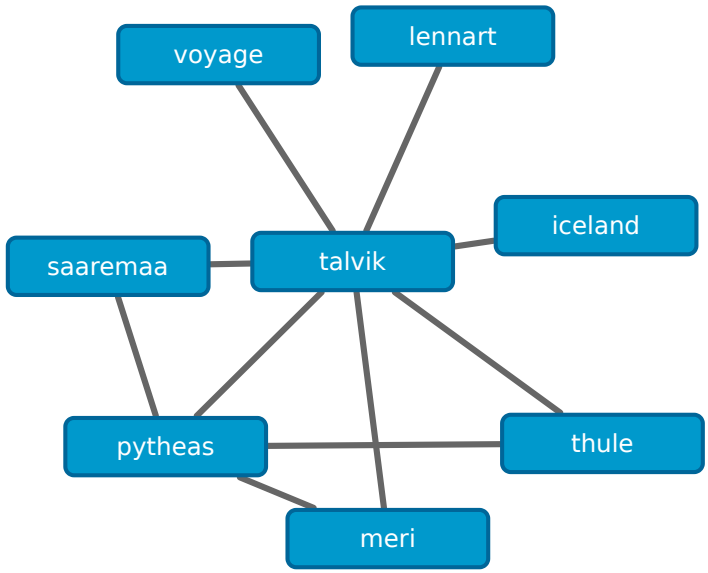



Figure 2.5: Document-specific word associations extracted with Mixture Model from the article given in Appendix B. In addition to Reuters and Brown corpus documents, the background also includes the Wikipedia page text for Pytheas.

that the Mixture Model does capture the information in the documents better than the regular model. But more about this in the next chapter.

# Chapter 3

# Document Summarization

In the previous chapter we introduced the method of extracting the document-specific word associations from a document. In this chapter we will focus on automatic document summarization. In our methodology the word associations are the building blocks of the method.

The goal of document summarization is to condense information represented in many documents into a short, say 250-word, summary. As briefly mentioned before this task can be divided into two distinct problems, first the modelling of what is written in the document(s) and secondly using this information to create a summary. Both of these problems are extremely complicated for a computer, sometimes even for humans.

The first part of this problem was already touched upon in Chapter 2, where we gave an overview of the word associations. Here, the document-specific associations are used for estimating what might be important in the document. The next step is using this information in order to create a summary.

Document summarization methods, like many other computational problems, can be divided into two approaches: supervised and unsupervised methods. As in machine learning, the difference is that supervised methods use existing labelled examples in order to learn how to solve a task at hand and unsupervised methods do not need labelled examples. Both of the approaches have positive and negative aspects. The supervised approach has more information at its disposal and should therefore perform better. However, producing the examples can be costly and is language-specific.

Although the unsupervised approach is more language-independent, it also has potential pitfalls, e.g., the developed algorithm might work for certain types of data, the performance of the method might be worse etc. In this thesis we focus on the unsupervised approach for document summa-

rization, although we believe that potentially the word associations could also be used as sentence or document features in supervised learning.

Next we will give an overview of the possible approaches to text summarization.

## 3.1 Single vs Multi-Document Summarization

The summarization techniques can be divided into two main categories: single- and multi-document summarization.

The automatic single-document summarization problem, pioneered by Luhn [64], takes a single, self-contained document about a single topic (e.g. news story, scientific article, Wikipedia article etc) as an input and produces a summary. For multi-document summarization the task is to create a summary of many documents. Usually the documents are about the same topic and it is very likely that they partially cover the same information (e.g. a set of news stories from different sources, scientific articles on the same topic etc).

The main difference between these tasks is that for single-document summarization it is reasonable to assume that there is little or no redundancy in the document. On the other hand, for multi-document summarization the opposite is true – it is very likely that the documents share some amount of information [7, 92].

In a sense the multi-document summarization is a slightly harder problem, because there are more aspects which need to be taken into account. For instance, the information content of the documents might be very different, meaning that some documents might not contain much useful information at all. Also, the system needs to take into account the possible redundancies etc.

For the single-document summarization there are more possible assumptions one can make about the document. As an example, news stories tend to follow a certain structure which then could possibly be utilised for summarization [61].

There is also a sub-task of multi-document summarization, called the *update summarization* [18]. The goal in this task is to create summaries with an assumption that the user has already read some of the documents. So in addition to modelling what is important in the documents, the system also has to recognize what information the user has already obtained and subtract this from the final summary. The task itself is very realistic – for many topics or events in the news we will get daily updates which often repeat a lot of information. Now considering that we already know what

the user knows, we could create a summary which contains only new and relevant information to her.

## 3.2 Text Generation

Generation of meaningful text is one of the toughest problems in natural language. This is even so for seemingly relatively easy problems – given an input text, just make it a bit shorter such that it says more or less the same thing. As language generation itself is a complex field, with many sub-problems [87], in text summarization many methods have taken an easier road by *extracting* sentences from the original documents (*extractive summarization*). The basic idea here is that the system will choose sentences from the original documents to cover different aspects of the topic and these sentences will constitute a summary.

This approach reduces the problem complexity significantly from generation of text to selection of sentences from a finite set, although, there are some potential pitfalls with just choosing sentences. As sentences appear in context, then by individually extracting them from the document might turn them into nonsense.

Another problem with sentence selection is the ordering of the sentences. For single-document summarization, the solution might be intuitive – just ordering the sentences in the order of appearance already makes sense, on the other hand for multi-document summarization the problem is more complex, as merely the position of a sentence in the document does not give too much information about the possible placement of the sentence in the final summary.

Even for single-document summarization, Jing [48] discovered that extracted sentences by professional summarizers might not be in the same order as in the original document, but the problem is even worse in the multi-document case.

Although in our work we do not put emphasis on sentence ordering, but rather on sentence extraction, an interesting way for ordering sentences is proposed by Barzilay et al. [5]. In this work two algorithms are considered, Chronological Ordering [70] and Majority Ordering [5]. The Chronological Ordering is the best for summarizing events. The time of the event is approximated by the date when it was first written about. Then the sentences are ordered by taking into account the publishing time and the event.

The Majority Ordering takes into account the *themes* in the document. A *theme* is a set of sentences, which contain repeated information, i.e. are essentially about the same event or concept. The Majority Ordering al-

gorithm takes into account the pairwise orderings of different themes and orders the sentences by maximizing the similarity between the pairwise orderings of themes in the summary and the pairwise orderings of the themes in the corpus which is being summarized.

A step forward from plain *sentence extraction* is a *hybrid* approach, where the sentences are either combined or modified before outputting the summary. The possible approaches for changing sentences are sentence revision [79], sentence fusion [6] and sentence compression [50].

The idea behind *sentence revision* is that the system takes the draft summary as an input and *revises* it to be more concise. Some of the methods include a set of rules, which indicate the method for eliminate parts of sentences or aggregating many sentences together, e.g. [65], the cut-and-paste method by Jing [51] uses human written professional summaries to identify the six possible operations on the summary sentences and similarly Nanba et al. [79] used human subjects to identify the factors which make extracts hard to read and devised revision rules for each of the factors.

*Sentence fusion* takes two sentences which share information, but also partly contain different information. The goal is to combine the sentences and either produce a new sentence which contains only the shared information or the information from both sentences without redundancy. For instance Filippova et al. [27] propose an unsupervised approach, which creates a dependency tree of related sentences and by analysing syntactic importance and word informativeness, a new dependency tree is induced, which in turn will constitute the new sentence. Other methods are proposed, which take into account the alignment of sentences [68], paraphrasing rules [7] etc.

The goal of *sentence compression* is to remove unnecessary parts in a sentence. For instance, Jing [49] proposes a method, which uses different kinds of knowledge, including contextual information, human written summary corpus statistics and syntactic knowledge to automatically compress sentences.

As our proposed methodology does not take into account even the ordering of the sentences, not to mention sentence revision or other refining methods, we refer to Nenkova et al. [81] who provide an exhaustive survey on the topic.

## 3.3 Sentence Selection

Our proposed methods in PAPERS II-IV are focused on *multi-document, unsupervised extractive summarization.*

The simplest of the unsupervised summarization methods are based on analysing word frequency. The seminal work in the area of document summarization was by Luhn [64], who proposed a method which took into account the word frequencies to extract the sentences to form a summary. Another word frequency-based system is SUMBASIC [102], which uses the probability of words to score the importance of a sentence and greedily pick the best scoring sentences. Many approaches use some kind of term-weighting and one of the most popular choices is *tf-idf* [89], which is used in many systems (e.g. [23, 31, 44]). Although many systems use term weighting for some parts of their system, the model itself is too simple to capture the complexity of the language.

For this reason, many systems turn to semantics. A popular choice from the domain of unsupervised methods is Latent Semantic Analysis (LSA) [19]. For instance the system by Hongyan [33] uses the right singular vectors of the term-sentence matrix to incrementally select sentences to be included in the summary. A similar approach is used by Wang et al. [104] where they combine term description with sentence description for each topic. Other systems using LSA are, e.g., [94, 95, 78, 82, 93].

A few techniques are language-independent, unsupervised and effective also in multi-document summarization. The most successful approach of the multilingual multi-document summarization workshop (MultiLing 2013) was UWB [93], a method based on singular value decomposition (SVD). UWB performed best in almost all the languages tested in MultiLing 2013.

An original approach for summarization is the DSDR method of He et al. [42]. This approach generates a summary by representing the candidate sentences as weighted term-frequency vectors and selects sentences in order to best "reconstruct" the original document. The authors define two objective functions, linear and nonnegative linear, which measure the goodness of the reconstruction. This work has been extended by combining document reconstruction and topic decomposition [107].

## 3.4 Mixture Model in Document Summarization

Our document summarization method essentially has two parts. In the first part the document-specific word associations are extracted from the documents. In the second part the sentences from the documents are selected greedily with an objective to cover as many document-specific associations as possible (cf. Chapter 2). The intuition behind this is that if the word associations are describing the important associations in the document, then

by choosing the sentences which contain these associations we can hopefully capture the central ideas of the document. The results indicate that our proposed method gives promising results not only in the English language, but in 9 languages, suggesting the method is quite language independent.

Our method is different from previously suggested methods, because firstly, it uses word pairs instead of just words and secondly, the documents are not analysed separately, but background information is considered. This means that the method does not try to figure out what is important in the document by analysing specific structures or co-occurrences of words just in the set of documents alone, but it incorporates the background corpus in order to extract the specific information of the documents.

This enables the system to be quite language independent and to make only a few assumptions about the document type or structure. The disadvantage of the method is the need for a background information corpus, which in the best case has similar types of text as the documents to be summarized. Also, due to the statistical nature of the approach, producing summaries of longer documents is more reliable than for short ones.

In PAPER III and PAPER IV the methods were broken down to two separate steps, first the sentence-scoring mechanism and second the optimization of selecting sentences. For this purpose we also define a *document-specific associations graph* which is a weighted graph representation where words are nodes and edges are associations between them.

Essentially we considered 3 scoring techniques for sentences: a) measuring the coverage of the document-specific associations by the chosen sentences; b) measuring the coverage of the most central words in the document-specific association graph; c) by combining the coverage and centrality approaches. The rationale behind covering the most central words in the association graph is similar to other document-specific word extraction models – by covering the central words of the document, presumably we also cover the most important ideas of the document. Out of these three, the combined approach, which takes into account the association coverage and the centrality of the words, performed the best.

We also considered two optimization techniques for choosing sentences: a) the greedy approach, where we iteratively chose the best sentence; b) genetic algorithm approach, where the goal was to optimize the overall summary. The greedy approach tended to give better results than the evolutionary algorithm. This most probably was related to poor parameter and genetic operation functions definition, rather than the overall performance of the algorithm.

In the introductory part we give a short overview of the best perform-

ing combination of the previously briefly described options — the greedy optimization of the coverage of associations and central words.

### Greedy Summarization

Consider that we have a set of documents $D \in S$ (as a refresher we refer to the notation in Section 2.2) which we would like to summarize. As mentioned before, we treat these documents as one long document $D_s$. This essentially means that we just trivially turn the approach into a single-document summarization problem. The first step is to extract the document-specific associations as described in Section 2.5 in order to get a better understanding of what might be important in the documents.

**Notation.** Consider that the strength between extracted association $(t_i, t_j)$ from document $D$ is given in the form $LLR(t_i, t_j)$. Then, the document-specific associations are pairs which have a stronger association than 0 between them, i.e. $LLR(t_i, t_j) > 0$, and which co-occur at least twice in the document $D_s$.

As the summarization technique also incorporates graph algorithms we also consider a graph $G = (V, E, W)$, where $V = \bigcup_{s \in D_s} s$ is the set of nodes (all words in the document $D_s$),

$$E = \{\{t_i, t_j\} \mid t_i \neq t_j, \exists s \in D_s \text{ s.t. } \{t_i, t_j\} \subset s\}$$

The log-likelihood ratio $LLR$ is used as the edge weight, i.e., $W(t_i, t_j) = LLR(t_i, t_j)$.

**Scoring.** As mentioned before, the scoring of the sentence relevance is based on two components: the coverage of the associations and the coverage of the centrality.

Given a sentence $s$, the coverage of the associations is simply the sum of all weights of the edges found in the sentence divided by the length of the sentence $|s|$:

$$cover(s) = \sum_{e \in E : e \subset s} W(e)/|s|.$$

To measure the importance of words, given word associations, we use the document graph $G$ and calculate the closeness centrality [30] for each of the nodes in the graph. As a distance measure between two nodes $v_i$ and $v_j$ we use the sum of the inverse weights $1/W(e)$ on the shortest path from $v_i$ to $v_j$. For each node $v \in V$, the centrality is given as:

$$C(v) = \frac{|V|}{\sum_{u \in V} d(u, v)},$$

where $d(u, v)$ is the length of the shortest path between nodes $u$ and $v$. Similarly to covering the associations, the centrality score is the sum of the centrality scores of the words found in the sentence normalized to sentence length:

$$centrality(s) = \sum_{v \in V : v \in s} C(v)/|s|.$$

In order to combine these two measures we simply normalize them to be between 0 and 1 and then add together:

$$combined(s) = \frac{cover(S)}{\sum_{\{t_i,t_j\} \in E} W(t_i, t_j)} + \frac{centrality(S)}{\sum_{v \in V} C(v)}.$$

This approach is very simple and the potential drawback is that it does not ensure that the numbers tend to range similarly.

**Sentence Selection.** Now that we have defined the score, the sentence selection is done in a greedy way by choosing the best scoring sentences one-by-one. The hope is that by doing this, the outcome is close to optimal.

Essentially all the sentences in $D_s$ are scored by the *combined*() function and the best one is chosen and added to the summary. Then the graph is updated, by setting all the weights of the edges found in the chosen sentence to 0 and then the next best sentence is chosen. The formal description can be found in Algorithm 1.

**Results.** Instead of presenting the same results as in the published articles, in this thesis we made a separate experiment. Let us again consider the article in Appendix B and we will generate 100-word summaries by using different document-specific association methods as the underlying model for the document. In multi-document summarization a popular summary length is often 250 words. As in this case we have only one document, 100 words seems to be sufficient.

First, we will calculate the summary by using just the Reuters and Brown corpus as the background. The sentences are in the order as they were picked by the greedy algorithm. For the top associations used for this summary we refer to Figure 2.4. The words in bold indicate the words also seen in top 10 associations.

1. He claimed that the mysterious **Thule** mentioned by **Pytheas** was actually **Saaremaa**.

**Algorithm 1** Greedy Selection Algorithm

**procedure** GREEDYSELECT
  **Input:** $D_s$, a set of sentences to be summarized
  **Output:** $S \subset D_s$, a summary of $D_s$
  $S \leftarrow \emptyset$ ▷ An initially empty summary
  $ls \leftarrow 0$ ▷ Current summary length
  **while** $ls < k$ **do**
    $\hat{s} \leftarrow null$
    $\hat{s} \leftarrow \underset{\substack{s \in D_s: \\ |s|+ls \leq k}}{\operatorname{argmax}} combined(s)$
    **if** $\hat{s} = null$ **then**
      $break$
    **end if**
    $S \leftarrow S \cup \hat{s}$
    **for** $(t_i, t_j) \subset \hat{s}$ **do**
      $W(t_i, t_j) \leftarrow 0$
    **end for**
    $ls \leftarrow ls + |s|$
  **end while**
  **return** S
**end procedure**

2. Resulting from his thorough research and analysis, Mr **Talvik** arrived at the conclusion that **Pytheas** reached the Baltic Sea shores.

3. To complicate work for **researchers**, the place names used by **Pytheas** are not the same today.

4. Medical scientist **proves** hypothesis set by Lennart **Meri**.

5. Lots of **researchers** of **Pytheas'** voyage have concluded that as he sailed around Britain from there he probably started off for **Thule** – whether to Iceland or Mid-Norwegian coast.

6. Comparing and analysing the **quotes**, Mr **Talvik** also discovered such as were obviously invented.

Another summary is produced by also using the Pytheas document in the background. The numbers following the sentences indicate whether the sentence was also found in the previous summary. For the top associations used for this summary we refer to Figure 2.5.

> He claimed that the mysterious **Thule** mentioned by **Pytheas** was actually **Saaremaa**. (1)
>
> Resulting from his thorough research and analysis, Mr **Talvik** arrived at the conclusion that **Pytheas** reached the Baltic Sea shores. (2)
>
> Medical scientist proves hypothesis set by Lennart **Meri**. (4)
>
> Lots of researchers of **Pytheas' voyage** have concluded that as he sailed around Britain from there he probably started off for **Thule** – whether to **Iceland** or Mid-Norwegian coast. (5)
>
> Though, according to random calculations by Mr **Talvik**, **Pytheas** lived on **Saaremaa** from nine months to 1.5 years, the knowledge of where **Thule** actually was got lost soon after he died. (NEW!)

By looking at these two summaries, observe that the first picked sentences are the same. The main difference is that sentences (3) and (6) were replaced with the new sentence. Intuitively, the added sentence is more informative than the sentences (3) and (6) together.

Also, in addition to the *document-specific models*, let us also consider the summary produced with common associations. This means that we will apply the standard common word association discovery method on the article in Appendix B and use the associations for creating the summary. The resulting summary is the following:

> Medical scientist proves hypothesis set by Lennart Meri. (4)
>
> He claimed that the mysterious Thule mentioned by Pytheas was actually Saaremaa. (1)
>
> Therefore, later researchers have used texts where other Antique authors talk about the voyage by Pytheas and his observations, and have on their basis arrived at greatly varying conclusions as to where the man actually travelled during the five years. (NEW!)
>
> Even so, when digging into it – just like Prof Talvik the history fan did, five years ago – you will discover that in most discourses the version of the mythical Thule as Saaremaa is not presented. (NEW!)

The summaries are quite different and objectively it is, of course, hard to decide which one of them is better. For the mixture model approach,

it seems that the focus of the summary is slightly more on the core of the subject, whereas for common word associations it is perhaps a bit more general. Here we should take into account that when choosing any sentence from a single relatively short document it is very likely that it is on topic anyway, independently from the sentence selection method.

In this chapter we saw that the word associations are indeed useful for document summarization. In the next chapter we will see how to use them for creative tasks.

# Chapter 4

# Word Associations in Computational Creativity

The search for understanding what creativity is has been going on for a long time [36, 37, 34]. In addition to the efforts in pinpointing what creativity is, there have also been practical approaches. One of these approaches is the development of psychometric tests of creativity. But what if computers could solve a test of creativity? Is there something we could learn from it?

Next, we will take a look at some already existing tests and then we will introduce methods for solving and creating tests of creativity.

## 4.1 Tests of Creativity

Creativity is usually defined as the ability to find associative solutions that are novel and of high quality. S. A. Mednick [73] defines creativity as "the forming of associative elements into new combinations, which either meet specified requirements or are in some way useful". On the basis of this definition, Mednick developed the Remote Associates Test (RAT) of creativity.

The RAT measures the ability to discover relationships between concepts that are only remotely associated. It has been frequently used by psychologists to measure creativity albeit there is some criticism concerning its validity in measuring creative skills [106, 55]. Each RAT *question* presents a set of three mutually distant words to the subject, and the subject is then asked to find a connecting word [73]. For instance, given the cue words 'lick', 'mine', and 'shaker' the *answer* word is 'salt': 'salt lick', 'salt mine', and 'salt shaker' connect 'salt' with each of the three words. The test is constructed so that the word associations in the test should be

familiar to people brought up in the culture in question (e.g. USA).

Most of the RAT answer words are quite uncommon. Thus, the test subject should propose answer words which are used less frequently in everyday speech to perform well on the test [73, 38]. This supports the idea that creative solutions usually are relevant and novel.

The RAT performance has been established to correlate with traditional measures of IQ [72], and there is some evidence that it predicts originality during brainstorming [29]. Additionally, several studies have linked RAT results to more specific creativity-related phenomena, such as intuition and incubation [11, 97, 103]. Thus, the RAT arguably provides a well-established method to assess the associative creativity in a psychological context.

There are other creativity tests available, e.g., the Torrance Tests of Creative Thinking (TTCT) [98]. TTCT measures creativity in four categories: fluency (i.e. number of meaningful ideas generated), flexibility (i.e. number of different meaningful ideas from different categories), originality (i.e. the measure of statistical rarity of the responses) and elaboration (i.e. ability to give details to the ideas). In addition to these, there are of course other tests (e.g., [13, 20]).

## 4.2 Solving Tests of Creativity

Intuitively, computationally solving the Remote Associates Test does not seem to be a very hard problem. Moreover, perhaps we could learn something in the process and potentially be able to generate such tests automatically?

Before focusing on our contributions, we will first give an overview of the related work.

### Creative Association Discovery

Several papers have been published on supporting creativity by discovering links between concepts. In creative problem solving, for instance, Juršič et al. [53] propose a system CrossBee for finding unexpected links between concepts from different contexts. Bisociations are based on the idea by Koestler [57] — relations between two (or more) concepts from different contexts that are not not directly related to each other.

Bisociations have also been studied in other contexts [54, 90]. For instance Petrič et al. [84] have used bisociations for creative literature mining. One of the motivations for literature mining is to find potentially related

research articles, which are linked to each other, but are from different domains.

Using bisociations is not limited to textual data. For instance Mozetič et al. [77] use the idea of bisociation on microarray data for finding enriched gene sets.

Examples of methods more directly based on link prediction in heterogeneous networks are given by Eronen and Toivonen [24].

### Solving RATs

An interesting question is whether a computer could solve a test which is meant for measuring the creativity of humans. Apparently it is possible and even more, for RATs, we discovered that on average a computer could do this more accurately than humans. But how were the tests solved in the first place?

Formally, the remote associates tests consist of questions where the subject is given three cue words $c_1, c_2$ and $c_3$ and they have to provide an answer word $a$. Computationally we need to solve the following problem: given three words $c_1, c_2, c_3$, which are not associated to each other, provide a word $a$, which is associated to all of them.

**Solving with 2-grams.** Words which are used next to each other also tend to be related to each other. Thus we used the Google 2 gram corpus [74] in order to model the simple associations between words. The preprocessing of the data is minimal — the 2 grams were just divided into pairs of words by using a whitespace to split them.

The tests were solved with a simple probabilistic model. We estimated the probabilities $P(a)$, $P(c_1|a)$, $P(c_2|a)$ and $P(c_3|a)$ from the Google corpus. When using a simple Naïve Bayes model, it turns out that we could quite accurately predict the word $a$ by knowing the cue words $c_1, c_2$ and $c_3$. Actually, we can do it even better than humans, as the average accuracy for humans for RATs is around 0.5 [10], whereas the performance for our system was 0.66!

The 2-gram data consists of word pairs which have very strong connections to each other. Could we do the same with more flexible associations?

**Generalized Approach.** For more flexible associations we extracted the common word associations (cf. Section 2.4) from the English Wikipedia[1]. Then, the common word associations were considered as a graph, where

[1] https://en.wikipedia.org

the words represented the nodes, the associations were the edges and the log-likelihood ratio value is the weight between nodes. By analysing the neighbourhood of the three cue words in this graph, we discovered that the answer word is usually part of the intersection of the neighbourhoods. However, the word which on *average* had the strongest association to the cue words was not the correct answer word. These seemingly correct words where usually quite common words and also quite strongly associated to one or two cue words, but the association to the third word was weaker.

The former observation is actually very well in concordance with the findings by Gupta et al. [38], that for solving RATs, the first word which comes to mind is usually wrong. In order to take advantage of this idea, we introduced a penalty factor into the choosing mechanism – the idea was to penalise words which are used very frequently. In the case where the answer word was indeed in the joint neighbourhood, it turns out that the penalty component has a significant impact to the accuracy of the method.

### Creating Interesting Combinations

Having a process which is able to solve creativity tests raises the question whether we could use it somehow to generate $n$ cue words based on a given answer word. This kind of approach could be applied for brainstorming where the goal is to try different ideas which could solve a problem. Also, there exists a relation to the four categories of the Torrance tests [98], namely, fluency is the number of responses we could provide for a certain concept; flexibility is the distance of the meaningful response words in the graph; originality can be measured by how rare the words we provide are and elaboration could be the neighbourhood of each of the provided words.

Again, using the **common word association network**, we propose a simple greedy method which takes the neighbourhood of the word $a$ and starts greedily picking words which are strongly related to $a$. To ensure that the words are not inter-related, after choosing a word $c_i$ we remove all the other words which are related to $c_i$ from the network. The greedy strategy has some obvious problems by only choosing the words which have the strongest connection. This strategy can (and most probably will) easily miss combinations, which would result in much higher average strength to the word $a$. This is a general problem of greedy approaches – in every iteration they essentially optimize for the next iteration, but do not look further from there. This means that the provided solution is locally optimal, but not necessarily globally.

These problems could be overcome by using some other search strategies, e.g., genetic algorithms or particle swarm optimization. In this case

| Seed Word | Cue Word 1 | Cue Word 2 | Cue Word 3 |
|---|---|---|---|
| imperialism | colonialism | lenin | american |
| missile | warhead | defence | flight |

Table 4.1: Generated remote associates tests.

we were not after the absolute optimal solution but rather the goal was to test the concept.

The results for this can be observed in Table 4.1, where we have provided just two examples (more in the original paper) for illustrating what kind of *tests* we might generate. In order to test whether these results are in concordance with the previous work, we applied the RAT solver algorithm (with common word associations) on the three cue words and checked whether the seed word and answer word matches. In 97% of the cases it did, indicating that the generated combinations have similar properties to the original tests.

### But Why?

Although the considered algorithms are quite simple, they are able to perform linguistic tasks which might be quite complicated for humans. This suggest that the amount of data might have an effect on computers' ability to solve different creative tasks.

Another aspect is the potential application areas for this system. We can also think of the RAT generation another way – given a concept, we are looking for words, which describe different contexts of the given concept. This kind of system could be used for creativity support – e.g, for helping to create advertising slogans or helping copywriters to explore different contexts of one specific concept. For instance, consider a copywriter who needs to advertise a *missile*. The words *warhead*, *defence* and *flight* might all trigger different ideas for the slogan, e.g., *Flies as smoothly as a Falcon* or *Defends you even when you sleep*.

Also, an interesting thing to notice is that, in order to work better, the algorithm is essentially penalising the most frequent choice. This is often where creativity lies – by not taking the straightforward road to the solution, but also considering options which might not look likely in the first place.

## 4.3 Poetry

As we discussed before, language generation is a notoriously hard task. If for the summarization task we already have some representation of the ideas and concepts we would like to express, the goal is even more challenging for poetry. One of the reasons is, that for poetry one would expect the content itself to be more metaphorical and to have a less obvious meaning than a summary of a news story. Also when it comes to poetry, there are certain constraints (e.g. rhyming, rhythm) and linguistic techniques which can be used, e.g., alliteration or sound symbolism.

The approach of PAPER V is based on our earlier work [96]. There we proposed a method where a template is extracted randomly from a given corpus and words in the template are substituted by words related to a given topic.

Next, we will look at different approaches to poetry generation and then we will give a brief overview of our method which gathers inspiration from a news story and writes a poem based on said news story.

### 4.3.1 Background

Quite many poetry generation methods have been developed and they vary a lot in their approaches by combining different computational and statistical methods in order to handle the aspects of linguistic complexity and creativity.

The state of the art in text generation (although not applied for poetry) by substituting words is presented, for instance, by Guerini et al. [35].

ASPERA [32] is a system, based on case-base reasoning, which generates poetry of an input text by composing poetic fragments that are retrieved from a case-base of existing poems. In the system each poetry fragment is annotated with a prose string that expresses the meaning of the fragment. This prose string is then used as the retrieval key for each fragment and the fragments are combined by using additional metrical rules.

The work of Manurung et al. [67] uses rich linguistic knowledge (semantics, grammar) to generate metrically constrained poetry out of a given topic via a grammar-driven formulation. This approach needs strong formalisms for syntax, semantics, and phonetics, and there is a strong unity between the content and form. The GRIOT system [39] is able to produce narrative poetry about a given theme. It models the theory of conceptual blending [26] from which an algorithm based on algebraic semantics was implemented. In particular, the approach employs "semantics-based interaction". This system allows the user to affect the computational narrative

and produce new meanings.

Also simpler approaches have been used to generate poetry. In particular, Markov chains ($n$-grams) have been widely used as the basis of poetry generation systems (e.g, [4, 86]) as they provide a clear and simple way to model some syntactic and semantic characteristics of language [59]. The problem with the Markov chains approach is that the content and form tends to be poor.

Creating poetry from news stories was also proposed by Colton et al. [15]. Their method generates poetry by filling in user-designed templates with text extracted from news stories.

### 4.3.2 Document Specific Poetry

Casual creativity [17] is a word used to refer to a set of small creative acts which we do every day for enjoyment, for instance jokes, talking in rhymes or fixing things at home with limited tools. Even if currently computers are not able to produce Picasso-level artwork, they could be applicable for casually creative tasks.

A possibly thought-provoking addition to a news story could be a poem which is inspired by the content of the news story. Using human labour for writing poems for each news story might be questionable, but computationally it could be a interesting addition. The idea of mixing poetry and news comes from them being to an extent very similar and different at the same time. Interestingly, both of them have a certain set of structural elements that are commonly used, e.g., for news stories these might be the *lead* and *structure of paragraphs*; and for poetry these elements could be, e.g., *rhyme*, *alliteration* etc. At the same time the goal of these two genres is very different – news should provide, in the ideal world, neutral information and analysis of global events, but for poetry the intention often is to induce emotions or create an atmosphere.

In PAPER V we propose a method for generating poems inspired by news stories. The question here was, how could we use a news story or a set of news stories as an inspiration for poetry generation? The idea of combining our previous work [96] and document-specific association extraction seemed to be an intuitive way to generate poetry on the topic of the news story.

The method uses a poetry corpus and a news story to generate a poem. First the system chooses a random poem from the poetry corpus. For the poetry generation machinery — P. O. Eticus [96] — the news story and the chosen poem have to be morphologically analysed, i.e., words in the text are analysed for their part of speech, singular/plural, case, verb tense etc.

Then, after being morphologically analysed, the document-specific word associations are extracted from a news story. The system substitutes words in the poem with the words found in the document-specific associations by taking into account the morphology of the words. In order to focus on the most important topics of the document, the system prefers words which create stronger associations. The outcome is the document-specific poem.

As an example consider a poem, which was created by using a news story about Justin Bieber[2] road racing while being drunk:

> The officer is taller than you, who race yourself
> So miami-dade and miami-dade: race how its entourages
> are said
> Co-operate and later in the singer, like a angeles of
> alcohols
> Racing with jails and singers and co-operate race

Although not exactly communicating what happened in the news story it gives a rather good feel of what might have happened – this is how information is often communicated in artistic works.

In this chapter we saw that the word associations can be used in the field of computational creativity which does suggest that in addition to being able to capture some essence of the documents, the word associations also are in general applicable to different kinds of tasks.

[2]http://www.bbc.com/news/world-us-canada-25863200

# Chapter 5

# Conclusion

In this thesis we covered two possible ways for extracting word associations – common word associations which model how words tend to co-occur; and document-specific associations which extract the important associations in the document. We also gave an overview of how these associations could be used for generative and creative tasks.

In this chapter we will conclude the introductory part of the thesis by discussing the contributions and give ideas on how the methods could be developed and used further.

## 5.1 Contributions

We will first give a concise list of the contributions in the papers.

- The main contribution of this thesis are the methods for finding *document-specific associations* and empirically evaluating their practical use. The development of the methods is discussed in PAPER II–PAPER III and the associations find use in PAPER IV and PAPER V;

- The extraction of *common word associations* is not a new idea and is largely built on the work by Dunning [22]. In PAPER I we provide experiments to show how the *common word associations* correlate to with the relations found in WordNet [76];

- In the field of document summarization we propose a new method Association Mixture Text Summarization which uses the document-specific summarization. In PAPER III we empirically show that the

summarization method works well on the English language. In PAPER IV we improve the method and empirically show that the method is generalizable to many languages;

- In the field of computational creativity, in PAPER I we propose a method for solving psychometric tests of creativity – Remote Associate Tests [71] (RAT) – and propose a method for expressing a concept via other mutually non-related concepts;

- We also show how the document-specific associations could be used for generating poetry about a certain document in PAPER V.

## 5.2 Discussion

The proposed language models are fairly simple but evidently they can be used for various linguistic tasks. This could be due to the fact that the principles behind the methods for extracting the associations are reasonable. In this work we also noticed an interesting nuance – although the algorithms we used were rather simple, we were able to do tasks which were at least to some extent intelligent.

The reason behind this probably is that the amount of data in this case is even more important than the complexity of the approach we use. Of course, we are not the first to notice that trade-off (cf. data mining). For instance the Word2Vec [75] system discussed before also focuses on how to process more data rather than have a more complex model. In our case we think it is reasonable to assume that, although the models are quite simple, with a large enough text corpus we could already capture enough regularities from the language to do something intelligible with it — create and solve creativity tests or create summaries of documents by choosing the most important sentences.

Evaluating the quality of the document-specific associations is a relatively hard task. The problem is that the rigorous definition of which word associations are *right* and which are *wrong* does not necessarily exist. Due to this reason we have not directly empirically evaluated the associations. However, it does not mean that the quality of the associations can not be assessed at all. This is where the document summarization and poetry factors in.

It is reasonable to assume that a good summary of the document contains the most central information. Thus, if we create a summary of the document by trying to cover as much of the associations as possible, then the goodness of the summary should reflect the quality of the associations.

The relatively good performance in the summarization task does indicate that the associations do make sense.

In PAPER V we proposed a method for creating document specific poetry and for evaluation we gave a number of automatically generated poems inspired by a news story. Whether this is a good way of evaluating creative artefacts is out of the scope of this thesis. The short answer is that it is definitely not the best way for evaluation, but on the other hand it gives some idea of the potential of the method. However, the fact that these poems tended to be on topic is still evidence that it was not just a blind chance that the document-specific associations were useful for automatic document summarization.

Another question is whether we capture something different with the document-specific associations than with common word associations? As can be seen in PAPER III, the inclusion of background information to the association calculation does improve the summarization performance. This suggests that the document-specific associations are able to capture the central ideas of a document better than the common associations. Even more, as the idea does seem to be applicable to many languages, it indicates that the proposed method is also quite robust.

In addition to the evaluation of the associations by the proxy task, in this thesis we also gave concrete examples of the document-specific associations extracted with different methods. We noticed some concrete aspects:

- The document-specific associations tend to contain less commonly associated words (e.g. *new* and *york*) (cf. Figures 2.2 ,2.3 and 2.4);
- By using a mixture model, the document-specific associations tend to contain words which are not very frequent in all documents (e.g., pronouns, articles etc.) (cf. Figure 2.3);
- The inclusion of specific information to the background model has a direct effect on the document-specific associations — the words which we expect to co-occur get a lower score (cf. Figure 2.4 vs Figure 2.5).

Again, of course these effects do not explicitly say that we have extracted the best possible associations — probably we have not, but they do indicate that it is quite likely that these associations capture some specifics of the document, especially due to the fact that these associations are *between words which co-appear in the same sentence.*

Remember that in the introductory part (Chapter 1) we brought three properties for document-specific associations: a) the association should not be common; b) the association should not contain frequent words; c) human

agrees that the document tries to establish the same association between the concepts. We argue that we have succeeded in extracting word associations which have the first two properties. Subjectively we would argue, that the example of Pytheas also demonstrates that the third property can be applied to the extracted associations — at least to us it intuitively seems that the obtained associations are the ones which the article wanted to establish.

## 5.3 Outlook

It seems that there are quite interesting directions which could be taken by using the document-specific associations. We did show empirically that the associations can be used in practice for different linguistic tasks. As discussed before the common and the document-specific associations can be represented as a word network. In the future it would be interesting to apply network analysis algorithms for these graphs, e.g., perhaps automatically detecting different contexts of specific words or even whole documents.

Currently the document summarization performance has been evaluated with automatic methods. The downside of these methods is that they do not measure the overall coherence of the summaries. Thus, although our proposed document summarization method performs quite well numerically, the coherence of the summaries leaves much to be desired. We see that potentially the method could be enhanced by applying, e.g., sentence revision or sentence fusion methods. Due to text generation not being the focus of this thesis, this has been left for future work.

# References


[1] Alfonseca, E., and Manandhar, S. Extending a lexical ontology by a combination of distributional semantics signatures. In *Knowledge Engineering and Knowledge Management: Ontologies and the Semantic Web*, A. Gómez-Pérez and V. Benjamins, Eds., vol. 2473 of *Lecture Notes in Computer Science*. Springer Berlin Heidelberg, 2002, pp. 1–7.

[2] Artale, A., Magnini, B., and Strapparava, C. WordNet for Italian and its use for lexical discrimination. In *AI*IA 97: Advances in Artificial Intelligence*, M. Lenzerini, Ed., vol. 1321 of *Lecture Notes in Computer Science*. Springer Berlin Heidelberg, 1997, pp. 346–356.

[3] Balkova, V., Sukhonogov, A., and Yablonsky, S. A russian WordNet. from UML-notation to Internet/intranet database implementation. In *Proceedings of the Second Global WordNet Conference. GWC-2004* (Brno, Czech Republic, 2004), P. Sojka, K. Pala, P. Smrz, C. Fellbaum, and P. Vossen, Eds., Masaryk University, pp. 31–38.

[4] Barbieri, G., Pachet, F., Roy, P., and Degli Esposti, M. Markov constraints for generating lyrics with style. In *20th European Conference on Artificial Intelligence, ECAI* (Montpellier, France, 2012), L. De Raedt, C. Bessiere, D. Dubois, P. Doherty, P. Frasconi, F. Heintz, and P. Lucas, Eds., vol. 242, pp. 115–120.

[5] Barzilay, R., Elhadad, N., and McKeown, K. R. Inferring strategies for sentence ordering in multidocument news summarization. *Journal of Artificial Intelligence Research 17*, 1 (Aug. 2002), 35–55.

[6] Barzilay, R., and McKeown, K. R. Sentence fusion for multidocument news summarization. *Computational Linguistics 31*, 3 (Sept. 2005), 297–328.

[7] BARZILAY, R., MCKEOWN, K. R., AND ELHADAD, M. Information fusion in the context of multi-document summarization. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics* (Stroudsburg, PA, USA, 1999), ACL '99, Association for Computational Linguistics, pp. 550–557.

[8] BLEI, D. M., NG, A. Y., AND JORDAN, M. I. Latent Dirichlet allocation. *The Journal of Machine Learning Research 3* (Mar. 2003), 993–1022.

[9] BODEN, M. A. *The Creative Mind: Myths and Mechanisms*. Psychology Press, 2004.

[10] BOWDEN, E. M., AND JUNG-BEEMAN, M. Normative data for 144 compound remote associate problems. *Behavior Research Methods, Instruments & Computers 35*, 4 (2003), 634–639.

[11] BOWERS, K. S., REGEHR, G., BALTHAZARD, C., AND PARKER, K. Intuition in the context of discovery. *Cognitive Psychology 22* (1990), 72–110.

[12] CHOMSKY, N. *Syntactic Structures*. The Hague: Mouton, 1957.

[13] CHRISTENSEN, P., GUILFORD, J., MERRIFIELD, R., AND WILSON, R. Alternate uses test. *Beverly Hills, CA: Sheridan Psychological Services* (1960).

[14] CHURCH, K. W., AND HANKS, P. Word association norms, mutual information, and lexicography. *Computational Linguistics 16*, 1 (1990), 22–29.

[15] COLTON, S., GOODWIN, J., AND VEALE, T. Full-FACE poetry generation. In *Proceedings of the Third International Conference on Computational Creativity* (Dublin, Ireland, May 2012), M. L. Maher, K. Hammond, A. Pease, R. Pérez, D. Ventura, and G. Wiggins, Eds., pp. 95–102.

[16] COLTON, S., AND WIGGINS, G. A. Computational creativity: The final frontier? In *20th European Conference on Artificial Intelligence, ECAI* (Montpellier, France, 2012), L. De Raedt, C. Bessiere, D. Dubois, P. Doherty, P. Frasconi, F. Heintz, and P. Lucas, Eds., pp. 21–26.

[17] Compton, K., and Mateas, M. Casual creators. In *Proceedings of the Sixth International Conference on Computational Creativity (ICCC 2015)* (Park City, Utah, June–July 2015), H. Toivonen, S. Colton, M. Cook, and D. Ventura, Eds., Brigham Young University, pp. 228–235.

[18] Dang, H. T., and Owczarzak, K. Overview of the TAC 2008 update summarization task. In *Proceedings of Text Analysis Conference* (Gaithersburg, Maryland, USA, 2008), National Institute of Standards and Technology (NIST), pp. 1–16.

[19] Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., and Harshman, R. Indexing by latent semantic analysis. *Journal of the American Society for Information Science 41*, 6 (1990), 391–407.

[20] Dow, G. T., and Mayer, R. E. Teaching students to solve insight problems: Evidence for domain specificity in creativity training. *Creativity Research Journal 16*, 4 (2004), 389–398.

[21] Dumais, S. T. Improving the retrieval of information from external sources. *Behavior Research Methods, Instruments, & Computers 23*, 2 (1991), 229–236.

[22] Dunning, T. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics 19*, 1 (1993), 61–74.

[23] Erkan, G., and Radev, D. R. LexRank: Graph-based lexical centrality as salience in text summarization. *Journal of Artificial Intelligence Research* (2004), 457–479.

[24] Eronen, L., and Toivonen, H. Biomine: Predicting links between biological entities using network models of heterogeneous database. *BMC Bioinformatics 13*, 119 (2012).

[25] Essen, U., and Steinbiss, V. Cooccurrence smoothing for stochastic language modeling. In *Acoustics, Speech, and Signal Processing, 1992. ICASSP-92., 1992 IEEE International Conference on* (Mar 1992), vol. 1, pp. 161–164 vol.1.

[26] Fauconnier, G., and Turner, M. *The Way We Think: Conceptual Blending and the Mind's Hidden Complexities.* Basic Books, New York, 2002.

[27] FILIPPOVA, K., AND STRUBE, M. Sentence fusion via dependency graph compression. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (Stroudsburg, PA, USA, 2008), Association for Computational Linguistics, pp. 177–185.

[28] FIRTH, J. R. A synopsis of linguistic theory 1930-55. *Studies in Linguistic Analysis (Special Volume of the Philological Society) 1952-59* (1957), 1–32.

[29] FORBACH, G., AND EVANS, R. The remote associates test as a predictor of productivity in brainstorming groups. *Applied Psychological Measurement 5*, 3 (1981), 333–339.

[30] FREEMAN, L. C. Centrality in social networks conceptual clarification. *Social networks 1*, 3 (1979), 215–239.

[31] FUNG, P., NGAI, G., AND CHEUNG, C.-S. Combining optimal clustering and hidden Markov models for extractive summarization. In *Proceedings of the ACL 2003 Workshop on Multilingual Summarization and Question Answering – Volume 12* (Sapporo, Japan, 2003), Association for Computational Linguistics, pp. 21–28.

[32] GERVÁS, P. An expert system for the composition of formal spanish poetry. *Journal of Knowledge-Based Systems 14*, 3–4 (2001), 181–188.

[33] GONG, Y., AND LIU, X. Generic text summarization using relevance measure and latent semantic analysis. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (New York, NY, USA, 2001), SIGIR '01, ACM, pp. 19–25.

[34] GÖTZ, I. L. On defining creativity. *Journal of Aesthetics and Art Criticism* (1981), 297–301.

[35] GUERINI, M., STRAPPARAVA, C., AND STOCK, O. Slanting existing text with Valentino. In *Proceedings of the 2011 International Conference on Intelligent User Interfaces* (2011), P. Pu, M. J. Pazzani, E. André, and D. Riecken, Eds., ACM, pp. 439–440.

[36] GUILFORD, J. P. Creativity. *American Psychologist 5* (1950), 444–454.

[37] GUILFORD, J. P. Creativity: Yesterday, today and tomorrow. *The Journal of Creative Behavior 1*, 1 (1967), 3–14.

[38] Gupta, N., Jang, Y., Mednick, S., and Huber, D. The road not taken. Creative solutions require avoidance of high-frequency responses. *Psychological Science* (2012).

[39] Harrell, D. F. Shades of computational evocation and meaning: The GRIOT system and improvisational poetry generation. In *Proceedings of the Sixth Digital Arts and Culture Conference* (Copenhagen, Denmark, 2005), pp. 133–143.

[40] Harrington, B. A semantic network approach to measuring relatedness. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters* (Stroudsburg, PA, USA, 2010), COLING '10, Association for Computational Linguistics, pp. 356–364.

[41] Harris, Z. Distributional structure. *Word 10*, 23 (1954), 146–162.

[42] He, Z., Chen, C., Bu, J., Wang, C., Zhang, L., Cai, D., and He, X. Document summarization based on data reconstruction. In *Twenty-Sixth AAAI Conference on Artificial Intelligence* (2012), pp. 620–626.

[43] Hofmann, T. Probabilistic latent semantic indexing. In *Proceedings of the 22Nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (New York, NY, USA, 1999), SIGIR '99, ACM, pp. 50–57.

[44] Hovy, E., and Lin, C.-Y. Automated text summarization and the SUMMARIST system. In *Proceedings of a Workshop on Held at Baltimore, Maryland: October 13-15, 1998* (Stroudsburg, PA, USA, 1998), TIPSTER '98, Association for Computational Linguistics, pp. 197–214.

[45] Ide, N., and Véronis, J. Introduction to the special issue on word sense disambiguation: The state of the art. *Computational Linguistics 24*, 1 (Mar. 1998), 2–40.

[46] Jaccard, P. Étude comparative de la distribution florale dans une portion des alpes et des jura. *Bulletin del la Société Vaudoise des Sciences Naturelles 37* (1901), 547–579.

[47] Jelinek, F. *The Impact of Processing Techniques on Communications*. Springer Netherlands, Dordrecht, 1985, ch. Markov Source Modeling of Text Generation, pp. 569–591.

[48] JING, H. Summary generation through intelligent cutting and pasting of the input document. Tech. rep., Columbia University, 1998.

[49] JING, H. Sentence reduction for automatic text summarization. In *Proceedings of the Sixth Conference on Applied Natural Language Processing* (Stroudsburg, PA, USA, 2000), ANLC '00, Association for Computational Linguistics, pp. 310–315.

[50] JING, H., AND MCKEOWN, K. R. The decomposition of human-written summary sentences. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (New York, NY, USA, 1999), ACM, pp. 129–136.

[51] JING, H., AND MCKEOWN, K. R. Cut and paste based text summarization. In *Proceedings of the 1st North American Chapter of the Association for Computational Linguistics Conference* (Stroudsburg, PA, USA, 2000), Association for Computational Linguistics, pp. 178–185.

[52] JURAFSKY, D., AND MARTIN, J. H. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*, 1st ed. Prentice Hall PTR, Upper Saddle River, NJ, USA, 2000.

[53] JURŠIČ, M., CESTNIK, B., URBANČIČ, T., AND LAVRAČ, N. Cross-domain literature mining: Finding bridging concepts with CrossBee. In *Proceedings of the Third International Conference on Computational Creativity* (Dublin, Ireland, May 2012), M. L. Maher, K. Hammond, A. Pease, R. Pérez, D. Ventura, and G. Wiggins, Eds., pp. 33–40.

[54] JURŠIČ, M., SLUBAN, B., CESTNIK, B., GRČAR, M., AND LAVRAČ, N. Bridging concept identification for constructing information networks from text documents. In *Bisociative Knowledge Discovery*. Springer, 2012, pp. 66–90.

[55] KASOF, J. Creativity and breadth of attention. *Creativity Research Journal 10*, 4 (1997), 303–315.

[56] KATZ, J. J., AND FODOR, J. A. The structure of a semantic theory. *Language 39*, 2 (1963), pp. 170–210.

[57] KOESTLER, A. *The act of creation*. Hutchinson, London, 1964.

[58] Lagus, K., and Kaski, S. Keyword selection method for characterizing text document maps. In *Artificial Neural Networks, 1999. ICANN 99. Ninth International Conference on (Conf. Publ. No. 470)* (1999), vol. 1, pp. 371–376 vol.1.

[59] Langkilde, I., and Knight, K. The practical value of n-grams in generation. In *Proceedings of the Ninth International Workshop on Natural Language Generation* (Ontario, Canada, Aug. 1998), pp. 248–255.

[60] Lenat, D. B. CYC: A large-scale investment in knowledge infrastructure. *Communications of the ACM 38*, 11 (Nov. 1995), 33–38.

[61] Lin, C.-Y., and Hovy, E. Identifying topics by position. In *Proceedings of the fifth conference on Applied natural language processing* (1997), Association for Computational Linguistics, pp. 283–290.

[62] Löfberg, L., Archer, D., Piao, S., Rayson, P., McEnery, T., Varantola, K., and Juntunen, J.-P. Porting an English semantic tagger to the Finnish language. In *Proceedings of the Corpus Linguistics 2003 Conference* (2003), D. Archer, P. Rayson, A. Wilson, and T. McEnery, Eds., pp. 457–464.

[63] Luhn, H. P. A statistical approach to mechanized encoding and searching of literary information. *IBM Journal of Research and Development 1*, 4 (Oct. 1957), 309–317.

[64] Luhn, H. P. The automatic creation of literature abstracts. *IBM Journal of Research and Development 2*, 2 (Apr. 1958), 159–165.

[65] Mani, I., Gates, B., and Bloedorn, E. Improving summaries by revising them. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics* (Stroudsburg, PA, USA, 1999), ACL '99, Association for Computational Linguistics, pp. 558–565.

[66] Manning, C. D., and Schütze, H. *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, MA, USA, 1999.

[67] Manurung, H. M., Ritchie, G., and Thompson, H. Towards a computational model of poetry generation. In *Proceedings of AISB Symposium on Creative and Cultural Aspects and Applications of AI and Cognitive Science* (2000), pp. 79–86.

[68] MARSI, E., AND KRAHMER, E. Explorations in sentence fusion. In *Proceedings of the European Workshop on Natural Language Generation* (Aberdeen, Scotland, 2005), pp. 109–117.

[69] MATUSZEK, C., WITBROCK, M., KAHLERT, R. C., CABRAL, J., SCHNEIDER, D., SHAH, P., AND LENAT, D. Searching for common sense: Populating Cyc from the web. In *Proceedings of the 20th National Conference on Artificial Intelligence* (Pittsburgh, Pennsylvania, USA, 2005), M. Veloso and S. Kambhampati, Eds., pp. 1430–1435.

[70] MCKEOWN, K. R., KLAVANS, J. L., HATZIVASSILOGLOU, V., BARZILAY, R., AND ESKIN, E. Towards multidocument summarization by reformulation: Progress and prospects. In *Proceedings of the Sixteenth National Conference on Artificial Intelligence and the Eleventh Innovative Applications of Artificial Intelligence Conference Innovative Applications of Artificial Intelligence* (Menlo Park, CA, USA, 1999), AAAI '99/IAAI '99, American Association for Artificial Intelligence, pp. 453–460.

[71] MEDNICK, M. Research creativity in psychology graduate students. *Journal of Consulting Psychology; Journal of Consulting Psychology 27*, 3 (1963), 265–266.

[72] MEDNICK, M. T., AND ANDREWS, F. M. Creative thinking and level of intelligence. *Journal of Creative Behavior 1*, 4 (1967), 428–431.

[73] MEDNICK, S. The associative basis of the creative process. *Psychological review 69*, 3 (1962), 220–232.

[74] MICHEL, J.-B., SHEN, Y. K., AIDEN, A. P., VERES, A., GRAY, M. K., THE GOOGLE BOOKS TEAM, PICKETT, J. P., HOIBERG, D., CLANCY, D., NORVIG, P., ORWANT, J., PINKER, S., NOWAK, M. A., AND AIDEN, E. L. Quantitative analysis of culture using millions of digitized books. *Science 331*, 6014 (2011), 176–182.

[75] MIKOLOV, T., SUTSKEVER, I., CHEN, K., CORRADO, G. S., AND DEAN, J. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems* (2013), pp. 3111–3119.

[76] MILLER, G. A. WordNet: A lexical database for English. *Communications of the ACM 38*, 11 (1995), 39–41.

[77] Mozetič, I., Lavrač, N., Podpečan, V., Helena Motaln, P. K., Petek, M., Gruden, K., Toivonen, H., and Kulovesi, K. Bisociative knowledge discovery for microarray data analysis. In *Proceedings of the International Conference on Computational Creativity* (Lisbon, Portugal, Jan. 2010), D. Ventura, A. Pease, R. Pérez, G. Ritchie, and T. Veale, Eds., Department of Informatics Engineering, University of Coimbra, pp. 190–199.

[78] Murray, G., Renals, S., and Carletta, J. Extractive summarization of meeting recordings. In *Proceedings of the 9th European Conference on Speech Communication and Technology* (Lisbon, Portugal, Sept. 2005), International Speech Communication Association, pp. 593–596.

[79] Nanba, H., and Okumura, M. Producing more readable extracts by revising them. In *Proceedings of the 18th Conference on Computational Linguistics - Volume 2* (Saarbrücken, Germany, 2000), Association for Computational Linguistics, pp. 1071–1075.

[80] Navigli, R. Using cycles and quasi-cycles to disambiguate dictionary glosses. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics* (2009), Association for Computational Linguistics, pp. 594–602.

[81] Nenkova, A., and McKeown, K. A survey of text summarization techniques. In *Mining Text Data*. Springer, 2012, pp. 43–76.

[82] Ozsoy, M. G., Cicekli, I., and Alpaslan, F. N. Text summarization of turkish texts using latent semantic analysis. In *Proceedings of the 23rd International Conference on Computational Linguistics* (Beijing, China, 2010), Association for Computational Linguistics, pp. 869–876.

[83] Paukkeri, M.-S., and Honkela, T. Likey: Unsupervised language-independent keyphrase extraction. In *Proceedings of the 5th International Workshop on Semantic Evaluation* (Stroudsburg, PA, USA, 2010), SemEval '10, Association for Computational Linguistics, pp. 162–165.

[84] Petrič, I., Cestnik, B., Lavrač, N., and Urbančič, T. Outlier detection in cross-context link discovery for creative literature mining. *The Computer Journal 55*, 1 (2012), 47–61.

[85] RAO, C. R. Diversity: Its measurement, decomposition, apportionment and analysis. *Sankhyā: The Indian Journal of Statistics, Series A* (1982), 1–22.

[86] REDDY, S., AND KNIGHT, K. Unsupervised discovery of rhyme schemes. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short Papers - Volume 2* (Stroudsburg, PA, USA, 2011), Association for Computational Linguistics, pp. 77–82.

[87] REITER, E. Has a consensus NL generation architecture appeared, and is it psycholinguistically plausible? In *Proceedings of the Seventh International Workshop on Natural Language Generation* (Kennebunkport, Maine, USA, 1994), Association for Computational Linguistics, pp. 163–170.

[88] ROARK, B., AND CHARNIAK, E. Noun-phrase co-occurrence statistics for semiautomatic semantic lexicon construction. In *Proceedings of the 17th international conference on Computational linguistics-Volume 2* (1998), Association for Computational Linguistics, pp. 1110–1116.

[89] SALTON, G., AND BUCKLEY, C. Term-weighting approaches in automatic text retrieval. *Information processing & management 24*, 5 (1988), 513–523.

[90] SLUBAN, B., JURŠIČ, M., CESTNIK, B., AND LAVRAČ, N. Exploring the power of outliers for cross-domain literature mining. In *Bisociative Knowledge Discovery*, M. Berthold, Ed., vol. 7250 of *Lecture Notes in Computer Science*. Springer Berlin Heidelberg, 2012, pp. 325–337.

[91] SOWA, J. F., Ed. *Principles of Semantic Networks: Explorations in the Representation of Knowledge*. Morgan Kaufmann, San Mateo, CA, USA, 1991.

[92] STEIN, G. C., BAGGA, A., AND WISE, G. B. Multi-document summarization: Methodologies and evaluations. In *Proceedings of the 7th Conference on Automatic Natural Language Processing (TALN'00)* (Lausanne, Switzerland, 2000), ATALA Press, Paris, pp. 337–346.

[93] STEINBERGER, J. The UWB summariser at Multiling-2013. In *Proceedings of the MultiLing 2013 Workshop on Multilingual Multidocument Summarization* (Sofia, Bulgaria, August 2013), Association for Computational Linguistics, pp. 50–54.

[94] Steinberger, J., and Ježek, K. Text summarization and singular value decomposition. In *Advances in Information Systems*, T. Yakhno, Ed., vol. 3261 of *Lecture Notes in Computer Science*. Springer Berlin Heidelberg, 2005, pp. 245–254.

[95] Steinberger, J., Kabadjov, M. A., Poesio, M., and Sanchez-Graillet, O. Improving LSA-based summarization with anaphora resolution. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing* (2005), Association for Computational Linguistics, pp. 1–8.

[96] Toivanen, J. M., Toivonen, H., Valitutti, A., and Gross, O. Corpus-based generation of content and form in poetry. In *Proceedings of the Third International Conference on Computational Creativity* (Dublin, Ireland, May 2012), M. L. Maher, K. Hammond, A. Pease, R. Pérez, D. Ventura, and G. Wiggins, Eds., pp. 175–179.

[97] Topolinski, S., and Strack, F. Where there's a will—there's no intuition: The unintentional basis of semantic coherence judgments. *Journal of Memory and Language 58*, 4 (2008), 1032–1048.

[98] Torrance, E. P. *Torrance tests of creative thinking.* Personnel Press, Incorporated, 1968.

[99] Tufi, D., Barbu, E., Mititelu, V. B., Ion, R., and Bozianu, L. The Romanian WordNet. *Romanian Journal on Information Science and Technology 7*, 1-2 (2004), 107–124.

[100] Turney, P. D. Learning algorithms for keyphrase extraction. *Information Retrieval 2*, 4, 303–336.

[101] Turney, P. D. Mining the web for synonyms: PMI-IR versus LSA on TOEFL. In *Proceedings of the 12th European Conference on Machine Learning* (London, UK, UK, 2001), EMCL '01, Springer-Verlag, pp. 491–502.

[102] Vanderwende, L., Suzuki, H., Brockett, C., and Nenkova, A. Beyond SumBasic: Task-focused summarization with sentence simplification and lexical expansion. *Information Processing & Management 43*, 6 (2007), 1606–1618.

[103] Vul, E., and Pashler, H. Incubation benefits only after people have been misdirected. *Memory & Cognition 35* (2007), 701–710.

[104] WANG, Y., AND MA, J. A comprehensive method for text summarization based on latent semantic analysis. In *Natural Language Processing and Chinese Computing*. Springer, 2013, pp. 394–401.

[105] WILSON, A., AND RAYSON, P. Automatic content analysis of spoken discourse: a report on work in progress. *Corpus Based Computational Linguistics* (1993), 215–226.

[106] WORTHEN, B. R., AND CLARK, P. M. Toward an improved measure of remote associational ability. *Journal of Educational Measurement 8*, 2 (1971), 113–123.

[107] ZHANG, Z., LI, H., AND HUANG, L. TopicDSDR: Combining topic decomposition and data reconstruction for summarization. In *Web-Age Information Management*, J. Wang, H. Xiong, Y. Ishikawa, J. Xu, and J. Zhou, Eds., vol. 7923 of *Lecture Notes in Computer Science*. Springer Berlin Heidelberg, 2013, pp. 338–350.

# Appendices

# Chapter A

# Intelligent Machines: Do We Really Need to Fear AI?

**By *Jane Wakefield***

**Technology reporter**

**28 September 2015, BBC**

Picture the scenario - a sentient machine is "living" in the US in the year 2050 and starts browsing through the US constitution. Having read it, it decides that it wants the opportunity to vote. Oh, and it also wants the right to procreate. Pretty basic human rights that it feels it should have now it has human-level intelligence. "Do you give it the right to vote or the right to procreate because you can't do both?" asks Ryan Calo, a law professor at the University of Washington. "It would be able to procreate instantly and infinitely so if it and its offspring could vote, it would break the democratic system."

This is just one of the questions Prof Calo is contemplating as he considers how the law has to change to accommodate our ever-growing band of robot and AI companions. He does not think that human-level intelligence is coming to machines any time soon but already our relationship with them is raising some interesting questions. Recently there was a tragic accident at a VW factory in Germany, when a robotic arm, that moved car parts into place, crushed a young man who was also working there. Exact details of the case are not yet released but it is believed human error was to blame. Volkswagen has not commented on the incident. While industrial accidents do happen, the law gets a little fuzzy when it involves a robot. It would be unlikely that a human could sue a robot for damage, for example. "Criminal law requires intent and these systems don't do things

wrong on purpose," said Prof Calo. How the world deals with the rise of artificial intelligence is something that is preoccupying leading scientists and technologists, some of who worry that it representsss a huge threat to humanity.

Elon Musk, founder of Tesla motors and aerospace manufacturer Space X, has become the figurehead of the movement, with Stephen Hawking and Steve Wozniak as honorary members. Mr Musk who has recently offered £10m to projects designed to control AI, has likened the technology to "summoning the demon" and claimed that humans would become nothing more than pets for the super-intelligent computers that we helped create. The pet analogy is one shared by Jerry Kaplan, author of the book, Humans Need Not Apply. In it, he paints a nightmarish scenario of a human zoo run by "synthetic intelligences". "Will they enslave us? Not really - more like farm us or keep us on a reserve, making life there so pleasant and convenient that there's little motivation to venture beyond its boundaries," he writes. Human intelligence will become a curiosity to our AI overlords, he claims, and they "may want to maintain a reservoir of these precious capabilities, just as we want to preserve chimps, whales and other endangered creatures".

Philosopher Nick Bostrom thinks we need to make sure that any future super-intelligent AI systems are "fundamentally on our side" and that such systems learn "what we value" before it gets out of hand - King Midas-style. Setting the controls for AI should come before we crack the initial challenge of creating it, he said in a recent talk. Without clearly defined goals, it may well prove an uncomfortable future for humans, because artificial intelligence, while not inherently evil, will become the ultimate optimisation process. "We may set the AI a goal to make humans smile and the super-intelligence may decide that the best way to do this would be to take control of the world and stick electrodes in the cheeks of all humans. "Or we may set it a tough mathematical problem to solve and it may decide the most effective way to solve it is to transform the planet into a giant computer to increase its thinking power," he said during his talk.

## Not yet

Ask an expert in AI when the robots will take over the world and they are likely to give you a wry smile. For IBM's head of research, Guru Banavar, AI will work with humans to solve pressing problems such as disease and poverty. While Geoff Hinton, known as the godfather of deep learning, also told the BBC that he "can't foresee a Terminator scenario". "We are still a long way off," although, he added, not entirely reassuringly: "in the

long run, there is a lot to worry about." The reality is that we are only at the dawn of AI and, as Prof Hinton points out, attempting to second-guess where it may take us is "very foolish". "You can see things clearly for the next few years but look beyond 10 years and we can't really see anything - it is just a fog," he said. Computer-based neural networks, which mimic the brain, are still a long way from replicating what their human counterparts can achieve. "Even the biggest current neural networks are hundreds of times smaller than the human brain," said Prof Hinton. What machines are good at is taking on board huge amounts of information and making sense of it in a way that humans simply can't do, but the machines have no consciousness, don't have any independent thought and certainly can't question what they do and why they are doing it. As Andrew Ng, chief scientist at Chinese e-commerce site Baidu, puts it: "There's a big difference between intelligence and sentience. Our software is becoming more intelligent, but that does not imply it is about to become sentient." AI may be neutral - but as author James Barrat points out in his book, Our Final Invention, that does not mean it can't be misused. "Advanced AI is a dual-use technology, like nuclear fission. Fission can illuminate cities or incinerate them. At advanced levels, AI will be even more dangerous than fission and it's already being weaponised in autonomous drones and battle robots." Already operating on the South Korean border is a sentry robot, dubbed SGR-1. Its heat-and-motion sensors can identify potential targets more than two miles away. Currently it requires a human before it shoots the machine gun that it carries but it raises the question - who will be responsible if the robots begin to kill without human intervention? The use of autonomous weapons is something that the UN is currently discussing and has concluded that humans must always have meaningful control over machines. Noel Sharkey co-founded the Campaign to Stop Killer Robots and believes there are several reasons why we must set rules for future battlefield bots now. "One of the first rules of many countries is about preserving the dignity of human life and it is the ultimate human indignity to have a machine kill you," he said. But beyond that moral argument is a more strategic one which he hopes military leaders will take on board. "The military leaders might say that you save soldiers' lives by sending in machines instead but that is an extremely blinkered view. Every country, including China, Russia and South Korea is developing this technology and in the long run, it is going to disrupt global security," he said. "What kind of war will be initiated when we have robots fighting other robots? No-one will know how the other ones are programmed and we simply can't predict the outcome."

We don't currently have any rules for how robots should behave if and when they start operating autonomously. Many fall back on a simple set of guidelines devised by science fiction writer Isaac Asimov. Introduced in his 1942 short story Runaround, the three laws of robotics - taken from the fictional Handbook of Robotics, 56th edition 2058, are as follows: A robot may not injure a human being or, through inaction, allow a human being to come to harm A robot must obey the orders given to it by human beings, except where such orders would conflict with the first law A robot must protect its own existence as long as such protection does not conflict with the first or second laws.

# Chapter B

# Medical Scientist Proves Hypothesis Set by Lennart Meri

**By *Priit Pullerits***

**Senior editor**

**16 October 2015, Postimees**

Not limited to liking what Lennart Meri wrote and published four decades back, medical scientist Raul Talvik believed it.

In his book "Hõbevalge" (Silver White) dating 1976 and its sequel "Hõbevalgem" seven years later, Mr Meri wrote that four centuries BC the major Ancient explorer Pytheas reached the territory of what is now Estonia. He claimed that the mysterious Thule mentioned by Pytheas was actually Saaremaa.

Even so, when digging into it – just like Prof Talvik the history fan did, five years ago – you will discover that in most discourses the version of the mythical Thule as Saaremaa is not presented. Largely, researchers think Thule is either Iceland, or some islands near coasts of Great Britain or Norway.

Yesterday, the studies by Mr Talvik (80) were presented to the nation, having penned into a thorough work labelled "Teekond maailma ääreni" (Voyage to the Edge of the World) wherein he proves that the Greek explorer Pytheas indeed reached all the way to Saaremaa, as was decades ago claimed by Mr Meri – unlike others who have delved into the topic.

This was no easy feat, as the description of travels by Pytheas has not been preserved. Therefore, later researchers have used texts where other Antique authors talk about the voyage by Pytheas and his observations, and have on their basis arrived at greatly varying conclusions as to where

the man actually travelled during the five years. To complicate work for researchers, the place names used by Pytheas are not the same today.

## "Shut up and row!"

To begin with, Mr Talvik tried to get a better picture of who Pytheas (who lived about 350–285 BC) actually was. "If I get to know his personality, from there I can guess and derive his activity as well," he explains.

Based on Antique sources, Mr Talvik ascertained that Pytheas was a simple man, poor, a lower class guy. "Being poor, he had no fleet as some mistakenly believe," he refutes one assumption. "Fleets weren't just handed out to people."

From there, Mr Talvik concluded while building on scarce sources that Pytheas had to have been wise. "When, alone, you embark on a voyage for years thro wild lands – back then, most were Barbarians – and you survive, you must be a good communicator and a friendly man," he said. "The man had no money. I'm sure they gave him an oar and said shut up and row."

Using the vast databases on Antique writers, Mr Talvik divided all quotes on Pytheas according to their reliability and verifiability, into three groups. The first, for example, contained such where his own books were quoted, and the second where it was quoted what Pytheas had said. Comparing and analysing the quotes, Mr Talvik also discovered such as were obviously invented. "In one place it is indirectly referred," he notes, "that Thule is at a place where the day lasts for six months and the night likewise. Meaning the North Pole. This is too much, that Pytheas discovered North Pole."

A reason why later quotes feature errors and slips is, says Mr Talvik, that the texts were usually copied by slaves. "Largely, they couldn't care less what they were writing," he observes. "They merely copied."

While up to now all the Thule issue dissectors have mainly relied on 17 ubiquitous quotes by Antique authors, Mr Talvik was able in his research to boost that by about 30. That added confidence.

## The fateful stumbling stone

Finally, as Mr Talvik had the assumed travel route of Pytheas all put together, he remembers he breathed a sigh a relief. But just for a moment. It all had to be proved.

He found 20 spots visited by Pytheas. Of these, he identified 15 with not much trouble. Five, however, were left hanging, Thule included. And with these, admits Mr Talvik, it got tough. At times, he was in dire straits.

"It was the issue of what Pytheas knew and could do," he continues, describing the process that followed. "What he could do could be concluded from what he did. But what did he know about astronomy, or geography?"

In lots of analyses, a faulty answer to this question proves the stumbling stone. Because nowadays all researchers know how to calculate the latitudes and longitudes, to say nothing about the knowledge that the Earth is a globe. Two and a half millennia ago, all was otherwise.

Lots of researchers of Pytheas' voyage have concluded that as he sailed around Britain from there he probably started off for Thule – whether to Iceland or Mid-Norwegian coast. In Prof Talvik, such conclusions cause disbelief as archaeological data says Man only reached Iceland in 9th century as the great travels of the Vikings begun. On top of that, in Antique times they only sailed the ships close to the coastlines and mainly during the day, which will exclude crossing the open seas from Britain to Iceland or Norway or near Greenland.

Resulting from his thorough research and analysis, Mr Talvik arrived at the conclusion that Pytheas reached the Baltic Sea shores. To prove that, he performed complex calculations with ancient seagoing data and maps. Turned out, he found vital pillars in two amber islands of the day, Basilia and Abalus – within a day's journey from each other – as identified via descriptions by tribes in old-time Scythia and ancient Germans. Abalus falls in the areas of today's Kaliningrad Oblast, and Basilia on Kurzeme coast. From Basilia to Thule, it remained an about three days' journey. That's exactly what it takes to reach from there to Saaremaa, in large rowing boats. To Thule, that is, using the name given it by Pytheas.

"Mr Pytheas was in this habit of giving his own names to places," smiles Mr Talvik. "Like this town which he named Rich in Doves. When you venture in a land where you know not the languages and can't speak them, how then do you write where you have been?"

## Solid stuff

As related to the Thule mystery, lots of references have made to the quote that Pytheas was in a place where the Sun goes to sleep, and this has undergone varying interpretations. Mr Talvik is supportive of the hypothesis by Mr Meri that Mr Pytheas meant the Kaali meteorite crater which came into being in 900–500 BC. Probably, as Pytheas saw the vast forest burnt down, he named the isle Thule, the Isle of Fire.

Though, according to random calculations by Mr Talvik, Pytheas lived on Saaremaa from nine months to 1.5 years, the knowledge of where Thule actually was got lost soon after he died. Until now, that is, as Prof Talvik

probably opened the most convincing chapter in the riddle.

"Seems to me these arguments are hard to fight," he says. The more so that he wasn't about to prove boots and all that Thule was the very Saaremaa. "I'm a Lennart Meri fan," he confesses, "but not to the degree to fake data to force his point."

All the man did was to go on a search where the mythical Thule was. And was simply taken to Saaremaa, as piloted by the facts and the links.

# Paper I

Oskar Gross, Hannu Toivonen, Jukka M. Toivanen, and Alessandro Valitutti

**Lexical Creativity from Word Associations**

In *Knowledge, Information and Creativity Support Systems (KICSS), 2012 Seventh International Conference on*, 35-42, IEEE, 2012.

2012 Seventh International Conference on Knowledge, Information and Creativity Support Systems

# Lexical Creativity from Word Associations

Oskar Gross, Hannu Toivonen,
Jukka M Toivanen and Alessandro Valitutti
Department of Computer Science and HIIT
University of Helsinki, Finland
Email: {oskar.gross, hannu.toivonen, jukka.toivanen, alessandro.valitutti}@cs.helsinki.fi

***Abstract*—A fluent ability to associate tasks, concepts, ideas, knowledge and experiences in a relevant way is often considered an important factor of creativity, especially in problem solving. We are interested in providing computational support for discovering such creative associations.**

**In this paper we design minimally supervised methods that can perform well in the *remote associates test* (RAT), a well-known psychometric measure of creativity. We show that with a large corpus of text and some relatively simple principles, this can be achieved. We then develop methods for a more general word association model that could be used in lexical creativity support systems, and which also could be a small step towards lexical creativity in computers.**

## I. Introduction

A fluent ability to associate tasks, concepts, ideas, knowledge and experiences in a relevant way is often considered an important factor of creativity, especially in problem solving. We are interested in providing computational support for discovering such creative associations. As a first step in this direction, we aim to design minimally supervised methods that perform well in the *remote associates test* (RAT) [1], a well-known psychometric measure of creativity.

The remote associates test is based on finding associations between words. In a RAT question, the subject is presented three *cue words*, e.g., 'coin', 'quick', and 'spoon'. Her task is then to find a single *answer word* that is related to all of the cue words. (Try to think of one! The answer word is given at the end of this paper.)

Accordingly our focus in this paper is on lexical creativity. While this may be considered a limited area of associative creativity, it has great potential in those tools for creativity support or problem solving that are based on verbal information, and also in creative language use such as computational poetry [2].

Our aim is to device methods that not only score well on RATs, but also require a minimum amount of explicit knowledge as input. We rely on corpus-based methods that learn word associations from large masses of text with statistical methods. Independence of knowledge bases, lexicons, or grammars also makes the methods easier to be applied to different languages.

In this paper, we first present a simple corpus-based method that has a relatively good performance (approximately 70%) on a standard RAT. RAT questions are well suited for corpus-based computational methods, and 2-gram models are largely sufficient to model and discover associations in them.

Next, inspired by the RAT setting, we propose a more general framework where more liberal, semantic associations between words can be discovered and used to support creativity, instead of the tightly bound, even idiomatic words of the RAT. To this end, we use word *co-occurrence networks*. Co-occurrence statistics of words are again computed from a document corpus, but in this case the words do not need to occur next to each other. The co-occurrence network can then be used as a simple model for creative inference, or as a component of a creativity support tool.

In the next section, we give a brief overview of the remote associates test of creativity. The contributions of this paper are then in the subsequent sections:

- We give a novel method that scores well on RAT questions of creativity using only frequencies of word collocations as its data (Section III).
- We generalize the RAT setting to more abstract relations between words and describe word co-occurrence networks for this purpose (Section IV)
- We propose a method for finding creative associations from word co-occurrence networks and give experimental results (Section V).

We review related work in Section VI, and conclude the paper in Section VII.

## II. Background: Remote Associates Test of Creativity

Creativity is usually defined as the ability to find associative solutions that are novel and of high quality. S. A. Mednick [1] defines creativity as "the forming of associative elements into new combinations, which either meet specified requirements or are in some way useful". On the basis of this definition, Mednick developed the remote associates test of creativity.

The RAT measures the ability to discover relationships between concepts that are only remotely associated. It is frequently used by psychologists to measure creativity albeit there is some criticism concerning its validity in measuring creative skills. Each RAT *question* presents a set of three mutually distant words to the subject, and the subject is then asked to find a word (creatively) connecting all these words together [1]. For instance, given the cue words 'lick', 'mine', and 'shaker' the *answer* word is 'salt': 'lick salt', 'salt mine', and 'salt shaker' connect salt with each of the three words. The test is constructed so that the word associations in the

978-0-7695-4861-6/12 $26.00 © 2012 IEEE
DOI 10.1109/KICSS.2012.38

test should be familiar to people brought up in the respective culture (e.g. USA).

Most of the RAT answer words are quite uncommon. Thus, the test subject should propose answer words which are used less frequently in everyday speech to perform well on the test [1], [3]. This supports the idea that creative solutions usually are relevant and novel. The RAT performance has been established to correlate with traditional measures of IQ [4], and there is some evidence that it predicts originality during brainstorming [5]. Additionally, several studies have linked RAT results to more specific creativity-related phenomena, such as intuition and incubation [6], [7], [8]. Thus, the RAT provides arguably a well established method to assess the associative creativity in a psychological context.

## III. A Computational Solution to RAT

We will now give a computational method for solving RAT tests with high accuracy, using only frequencies of word pairs in a large corpus. We will walk through the ideas using a number of experiments, so we start by describing the data we have used.

### A. Background

*a) RAT tests:* We combined RAT tests from two sources [9], [10] and obtained a total of 212 questions. Following good practices of data analysis, this set of tests was then divided into two disjoint sets: a training set of 140 questions and a test set of 72 questions. Method development is carried out using the training set, while the validation set is used to test the performance on the final methods. This procedure avoids overly optimistic results that would be obtained by tuning and testing the methods on the same instances.

*b) Corpus:* Instead of a full corpus of text, we directly use Google 2-grams [11], a large, publicly available collection of 2-grams (see below).

We next formalize some of the concepts and introduce notation used in the rest of the paper.

*c) Notation:* $n$-grams, i.e., frequencies of different sequences of $n$ words, are used widely in language modelling. For solving RATs, we use 2-grams. A 2-gram is a sequence of two words or, more formally, a vector $n = (n_1, n_2)$ of two words $n_1$ and $n_2$. The (absolute) frequency of a 2-gram $n = (n_1, n_2)$, denoted by $n_c$, is the number of times the sequence $(n_1, n_2)$ of words occurred in a given corpus $C_G$. We denote by $N$ the set of all 2-grams and by $N_c$ the total of their occurrences. Let $N'_c(t)$ denote the sum of frequencies of the 2-grams that contain word $t$, i.e.,

$$N'_c(t) = \sum_{n \in N : t \in n} n_c.$$

In a similar way,

$$N'_c(t_1, t_2) = \sum_{n \in N : t_1, t_2 \in n} n_c = (t_1, t_2)_c + (t_2, t_1)_c$$

denotes the total of frequencies of 2-grams that contain both $t_1$ and $t_2$.

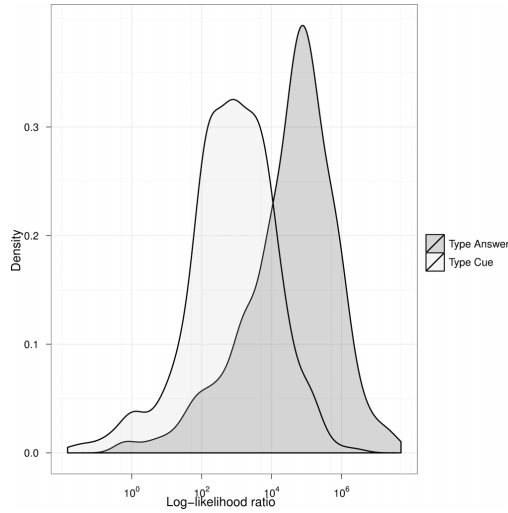



Fig. 1. The log-likelihood distribution of the different types of word pairs

Formally, a RAT is a quadruple $r = (c_1, c_2, c_3, a)$, where $c_i$ is the $i$th cue word and $a$ is the answer word.

### B. Methods

*a) Frequencies of RAT word pairs:* The way RAT tests are constructed implies that 2-grams $(c_i, a)$ or $(a, c_i)$ consisting of a cue word and the answer word should have relatively high frequencies, and that 2-grams $(c_i, c_j)$ consisting of two cue words should have relatively low frequencies.

Since the individual words in a RAT may have different frequencies, 2-grams also have different expected frequencies. So, rather than directly comparing the frequencies of 2-grams, we estimate how much the observed frequencies differ from the ones expected assuming statistical independence. We measure this deviation by the log-likelihood ratio (LLR) [12]. For this calculation, we estimate the individual frequencies of words by the number of times they occur in 2-grams.

Figure 1 shows the LLR distributions for cue word pairs ('Type cue') and for cue word, answer word pairs ('Type answer'). The cue word, answer word pairs clearly tend to be more closely related than the cue word pairs, but there is also a lot of overlap between the distributions. The difference between the distributions is statistically significant (Wilcoxon rank sum test p-value $< 2 \cdot 10^{16}$).

*b) Scoring function:* To solve a RAT test we need to find an answer word that is related to all of the cue words. We propose to treat each RAT question $r$ as a probabilistic problem, where we want to find the most likely answer word $a$, i.e., one that maximizes the conditional probability $P(a|c_1, c_2, c_3)$.

We have

$$\begin{aligned} P(a|c_1,c_2,c_3) &= \frac{P(a,c_1,c_2,c_3)}{P(c_1,c_2,c_3)} \\ &\propto P(a,c_1,c_2,c_3) \quad (1) \\ &= P(c_1,c_2,c_3|a)P(a). \quad (2) \end{aligned}$$

Assuming that the cue words $c_1, c_2, c_3$ are mutually independent, as they essentially are by construction of RATs, we have

$$P(c_1,c_2,c_3|a)P(a) = P(a)\prod_{i=1}^{3} P(c_i|a). \quad (3)$$

(In machine learning, this is known as the Naïve Bayes model. It often has a good practical predictive performance even if the independence assumption does not hold [13].)

We estimate the conditional probabilities from the relative frequencies of the words in the 2-grams,

$$P(a) = \frac{N'_c(a)+1}{N_c+1},\ P(c,a) = \frac{N'_c(c,a)+1}{N_c+1}, \quad (4)$$

giving

$$P(c|a) = \frac{P(c,a)}{P(a)} = \frac{N'_c(c,a)+1}{N'_c(a)+1}. \quad (5)$$

*c) Answer word search:* Given a RAT test, finding the best scoring answer word $a$ among millions of words is not straightforward. We do this in two steps. In the first step, we extract words that occur at least once with each cue word. Let this set of candidate words be $\Gamma$. In the second step, we compute the conditional probabilities of the candidate words and choose the best one, i.e.,

$$\begin{aligned} &\arg\max_{a\in\Gamma} P(a)\prod_{i=1}^{3} P(c_i|a) = \\ &\qquad = \arg\max_{a\in\Gamma} P(a)\prod_{i=1}^{3} \frac{(N'_c(c_i,a)+1)}{(N'_c(a)+1)}. \quad (6) \end{aligned}$$

### C. Experiments

We experimented with the RAT solver using the training and test sets with 140 and 72 RATs, respectively.

Already in the first experiment, the method was able to give correct answers to 56% of the RATs in the training set and the accuracy for the test set RATs was 54%. By looking at the results we observed that many false solutions were very frequent words of English (also known as stopwords).

After simple stopword removal (we used the NLTK [14] stopword list) from the candidate set, the accuracy of the system for both sets increased to 66%. Now, many of the seemingly incorrect results were actually solved essentially correctly, but instead of the singular in the correct answer, the plural form of the answer word was proposed by the system. Such minor issues could be easily solved, but since our main interest is more in the principles that may help develop computational creativity, we did not delve into details.

An upper bound for the accuracy of the 2-gram-based technique for the training set is 96% and for the test set it is 99%. This is how often the candidate set included the correct answer word. Many of the remaining failed cases are due to compound words. For instance, for the RAT question with cue words *puss*, *tart* and *spoiled* the answer word *sour* is not detected because in everyday text 'sourpuss' is written together. Again, techniques to take this into account could be developed, but would not probably help finding truly creative associations.

Our results indicate that the method described above solves RAT questions more accurately than an average human. According to Bowden and Jung-Beeman [15], mean human accuracy for their 144 RAT questions is approximately 0.5, whereas the accuracy of our simple method is 0.66.

Overall the results indicate that the computational method based on 2-grams has already captured some principles of creativity, as measured by RATs.

## IV. Generalized Approach to Support Creativity

The 2-gram model model used above is severely restricted and essentially only considers idiomatic phrases, such as compound words of exactly two elements. Obviously, many — if not most — relevant and informative associations between terms are manifested by less stringent proximity.

We next propose a more powerful, generalized approach to support creativity based on relations which are semantic in nature [16]. We are motivated by the observation that RATs are relatively easy for computers and that more general notions of relatedness of words or concepts could be used. Since RATs already correlate with creativity, a more general version could likely be used to support more challenging tasks of creativity.

In this section we describe a simple method for creating a network of semantically associated words. We experimentally test and illustrate how connections in this network tend to make sense. We also show how to apply the RAT solving principles to these networks in order to support some sorts of creative inference.

### A. Word Co-Occurrence Network Construction

We briefly describe how a word co-occurrence network can be generated using existing text analysis methods. We assume a corpus of unstructured documents, and we treat documents as bags of sentences and sentences as bags of words. Formally, the document corpus $C_W$ is a set of documents $d_i \in C_W$, where each document $d_i$ is a (multi)set of sentences $d_i = \{s_{i1}, \ldots, s_{in}\}$, and each sentence is a set of words $s_{ij} \subset T_W$, where $T_W$ is the set of all words.

We analyse word co-occurrences at the granularity of sentences, since words which are in one sentence have a strong relation to each other [17]. Valid alternative approaches could be based on a sliding window of words or a paragraph, for instance.

Formally, the word co-occurrence network $G = (V, E, W)$ is a weighted, undirected graph with nodes $V$, edges $E \subset V \times V$, and edge weights $W : V \times V \to \mathbb{R}_+$. For notational convenience, we assume $W(e_1, e_2) = 0$ if there is no edge between $e_1$ and $e_2$.

Before constructing the graph we preprocess the documents. First, we extract nouns and named entities from the documents and discard everything else. In addition to simplicity, this choice is motivated by nouns and named entities being conceptually more basic than concepts referred to by verbs or prepositions [18]. Obviously, some information is lost here. We then lower-case and lemmatize all the words. The named entities are concatenated with an underscore.

We use the log-likelihood ratio (LLR) to measure the strength of an association between two terms [12]. In the word co-occurrence network, lemmatized nouns and named entities are then nodes, and they are connected with an edge whenever the LLR is high enough (see below). The connections are also weighted by the LLRs.

### B. Word Co-Occurrence Network of Wikipedia

In order to discover more general connections between words we chose to extract word co-occurrences from a text corpus. Google n-gram data sets are not used here since they only contain information about words which appear very close to each other.

In these experiments we construct the co-occurrence network from the English Wikipedia as of September 2011, consisting of 2,078,604 encyclopedic articles from all areas of life. For preprocessing the data we use Natural Language Processing Toolkit (NLTK) [14].

Without any pruning of edges, the co-occurrence network constructed from Wikipedia would consist of 1,900,846 nodes and 89,076,150 edges. Figure 2 shows the distribution of LLR values, i.e., the weight distribution of all possible edges before any pruning. As is to be expected, a majority of weights are small but there is a long tail to large weights.

Selecting a threshold value for LLR is a complicated task. Our reasoning was, that the minimum log-likelihood ratio value should be at least as high as it is for two terms which co-occur only twice and together. In our case the value $t = 70.44$ was used as the threshold value for the co-occurrence network. This removes approximately 95% of the edges from the network (cf. Figure 2). As a result, the network consists of 595,029 different terms and 4,644,456 edges.

### C. Co-occurrence Network vs. WordNet Semantic Relations

To experimentally investigate what kind or semantic relations are discovered by the LLR-based method, we next experiment with WordNet [17]. It is a curated lexical database of English, with a large amount of manually assigned semantic relations of different types between words. WordNet is an accurate and powerful resource but limited in its scope. There are approximately 120,000 nouns in WordNet, when including example sentences and glossaries (see below). The co-occurrence network thus has around 470,000 nodes which do not appear in WordNet at all.

Our goal has not been to reproduce WordNet. Rather, we aim for a coverage much wider than WordNet (our 595k terms vs. WordNet's 120k terms), and also for language-independence so that the methods are applicable also in languages for which WordNet or similar resource do not exist. The sole purpose of these experiments is to shed light on the types of relationships discovered by LLR.

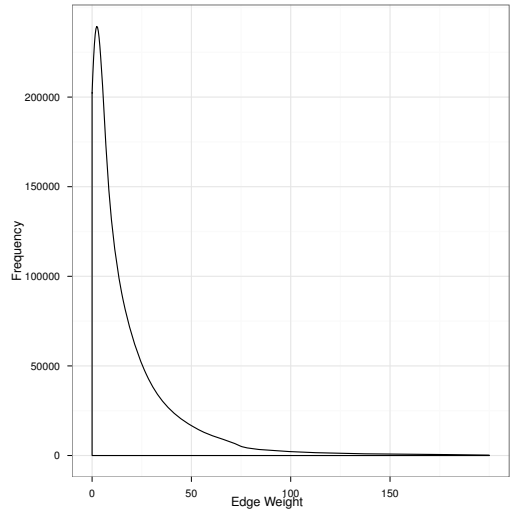



Fig. 2. Weight (LLR) distribution of the co-occurrence network before pruning.

Given two words $w_1$ and $w_2$, we consider their following possible relations in WordNet:

- $w_1$ is a *hypernym* of $w_2$, or vice versa (e.g. 'vehicle' is a hypernym of 'car').
- $w_1$ is a *holonym* of $w_2$, or vice versa (e.g., 'car' is a holonym of 'wheel').
- $w_1$ is a *holonymic sister* of $w_2$, i.e., they share a holonym (e.g., 'wheel' and 'door' both are parts of a car).
- $w_1$ and $w_2$ are *synonyms* (e.g., 'car' and 'automobile').
- $w_1$ and $w_2$ are *coordinate terms*, i.e., they share a hypernym (e.g., 'car' and 'ship' both are vehicles)
- $w_1$ appears in the *definition* of $w_2$, or vice versa (e.g., 'motor' appears in the WordNet definition of car: "a motor vehicle with four wheels; usually propelled by an internal combustion engine").
- $w_1$ appears in the *example sentences* of $w_2$, or vice versa (e.g., 'work' appears in the WordNet example use of the word car: "he needs a car to get to work").

More distant WordNet similarities could also be considered by transitively applying the above relations (for an overview see, e.g., [19]).

Because of the limited scope of WordNet, for our experiments concerning WordNet relations we randomly picked 5,000,000 pairs of words that do occur in WordNet. We excluded those words in our co-occurrence network that do not appear in WordNet, since obviously WordNet is not able to say anything about their relations.

| Relation Type in WordNet | Number of Examples |
|---|---|
| Hypernym Relations | 117 |
| Holonym Relations | 49 |
| Holonymic Sister Relations | 6 |
| Synonym Relations | 33 |
| Coordinate Relation | 2,729 |
| Definition Relation | 948 |
| Example Relation | 70 |
| No Relation | 4,996,048 |
| Total Sample | 5,000,000 |

TABLE I
DISTRIBUTION OF DIFFERENT WORDNET SEMANTIC RELATION TYPES IN A RANDOM DATASET.

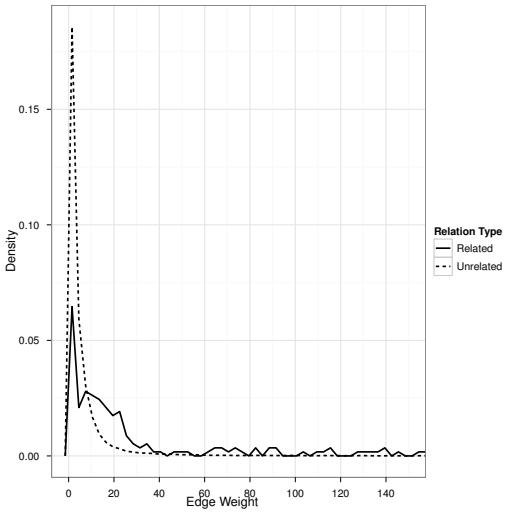



Fig. 3. Edge weight (LLR) distributions of edges which either are or are not related in WordNet.

The distribution of WordNet association types in the random sample of 5,000,000 pairs is shown in Table I. The number of words which are related in WordNet form a very small fraction of the dataset. Also, most term pairs in this random sample have low LLRs, essentially following the distribution of Figure 2.

Correlation between WordNet and LLRs is illustrated in Figure 3, where the edge weight distributions are drawn separately for those pairs that are related in WordNet and those that are not. Visually, the difference is clear: approximately already from edge weight 15 on, related word pairs have a higher density than unrelated pairs.

Since so few pairs are related in WordNet, we also look at the data using ROC (Receiver Operating Characteristic) curve which is suited for unbalanced class distributions. The curve can be seen in Figure 4 (zoomed in to the lower left corner).

The true positive rates grow in the beginning very fast (note

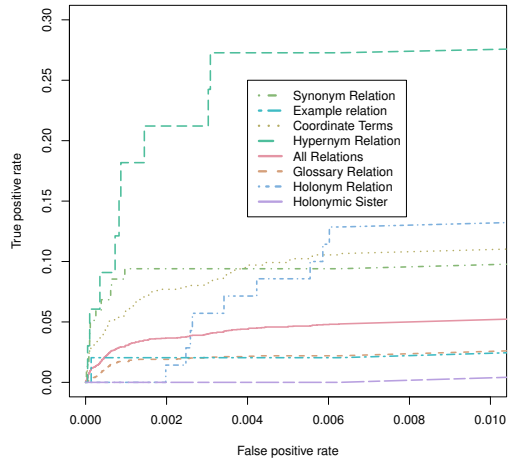



Fig. 4. Zoom-in to the lower left corner of the ROC plot.

the difference in $x$ and $y$ scales in the figure), but then they level off to a straight line towards point $(1, 1)$. This indicates that the top ranking term pairs are typically WordNet related, as suggested also by Figure 3, but after that there is no visible difference.

These experiments show that the relations discovered by LLR tend to make sense semantically. The sheer numbers additionally show that the co-occurrence method has a much higher coverage than WordNet (but obviously WordNet has strengths, such as semantic categories of relationships and manually curated contents). We believe that word co-occurrence based models on which we can build creativity support methods could be much more interesting than the 2-gram models for solving RATs.

## V. CREATIVE ASSOCIATION DISCOVERY

We now proposed initial methods for finding more general creative associations. First we will propose a generalized version of the method proposed for RATs in Section III. Note, however, that now the goal is not to solve RATs, they are just used to ensure that the responses of the proposed algorithm are sound.

In the final subsection we will actually propose a method for generating generalized RATs, and we will show that the generation method is quite stable. We will also provide examples of the creative inference to the reader.

### A. Generalization of RAT-Related Methods

*a) Candidate word selection:* The generalization of the candidate method from the previously presented method is quite straightforward. In the method which used 2-grams as the

model of co-occurrences the words which co-occur with every cue word were used as candidate answer words. Choosing the candidate set can be done in a similar way for the co-occurrence network by choosing the joint neighbourhood of all the cue words.

More formally, let us consider a set $T = \{t_1, \ldots, t_n\}$ of words which we treat as cue words. We will define the joint neighbourhood as the intersection of all the neighbours of the cue words:

$$\mathcal{N}(T) = \{u \mid \{t_i, u\} \in E \text{ for all } t_i \in T\}. \quad (7)$$

*b) Scoring:* For ranking the candidates, consider first a single candidate word $a \in \mathcal{N}(T)$. We propose using a score which depends on two aspects of the candidate word $a$. First, a good answer word $a$ should be strongly related to all of the cue words $t_i$. Second, a good answer word is specific to the cue words, i.e., does not associate strongly with too many other words. The second criterion also relates to the fact that high-frequency candidates are not considered as creative [3].

We define the scoring function as

$$score(a, T) = \alpha(a, T) \cdot \beta(a), \quad (8)$$

where $\alpha(a, T)$ is the association weight-induced component of the score and $\beta(a)$ is the candidate frequency-induced component of the score.

Some reasonable scores which could be calculated as the $\alpha$ component are the following:

1) The minimum weight (MINW) between the answer word and the cue words, i.e., "the weakest link":
$$\alpha(a, T) = \min_{t_i \in T}(W(a, t_i)).$$
2) The average edge weight (AVGW) between the cue words and the answer word
$$\alpha(a, T) = \frac{1}{|T|} \sum_{t_i \in T} W(a, t_i).$$
3) As the edge weights are ratios, it is also reasonable to consider the harmonic mean (HARM)
$$\alpha(a, T) = \frac{|T|}{\sum_{t_i \in T} \frac{1}{\max(W(a, t_i), 1)}}.$$

Analogously there are different ways to penalize the answer word frequency. In this paper we consider the two most obvious approaches related to the degree of the candidate node $a$. The first approach penalizes a score by dividing it by the candidate node degree (DEG), i.e.,

$$\beta(a) = \frac{1}{\deg(a)}.$$

A logarithmic smoothing of the degree penalty (DEGL) component might give more stable results:

$$\beta(a) = \frac{1}{\log(\deg(a))}.$$

### B. Generalized RAT Creation

In standard RAT questions the goal is to provide an answer word given the cue words. While this measures creative abilities, often the opposite task has more practical value: we have a concept (the answer word, e.g., the topic of a problem we want to solve), and we want to have it associated creatively with other concepts. For instance, let's assume we are interested in the word 'riding' and, to support our creativity, would like to see it associated with different things. The method that we will give below recommends these words: 'election', 'horseback' and 'accident'.

In this task, given an answer word, our goal is to select words that are strongly related to the answer word and at the same time are not related to each other. We propose this simple algorithm for selecting such words given the answer word $a$: First, choose the node with the strongest connection to $a$ and add it to the (so far empty) cue word set $R$. Then, consider other nodes in a decreasing order of their association with the answer word $a$. Add a node to the cue word set $R$ if and only if it is not connected to any member of $R$. Iterate until the desired number of cue words has been chosen or all neighbours of the answer word have been considered.

### C. Experiments

Our first experimental goal is to test how well different scoring functions work on RAT questions. We will conduct these experiments on the training set. Once we have chosen the best method we will validate it using the separate test set.

Recall that the documents were preprocessed to support discovery of non-trivial associations between concepts. This preprocessing, i.e. including only named entities and nouns in the network, actually hinders solving the RATs. Therefore, we compare different scoring functions using those RAT questions where the candidate answer set (the joint neighbourhood of the cue words) contains the correct answer word. 21% of the test cases fell in this category. The relatively low score is explained by preprocessing aspect which we described earlier (i.e. many common entities are treated as one, e.g. 'political', 'party' is treated as 'political_party' in the co-occurrence network).

Results are shown in Table II (for acronyms used in the table, see the previous subsection). For $\alpha(a, T)$, the association weight-dependent component of the score, the harmonic mean (HARM) systematically produced best results. For $\beta(a)$, the candidate frequency-dependent component, the best results were obtained when dividing the score by the number of associations, i.e., the degree of node $a$ in the co-occurrence graph (DEG). Overall, their combination also gave the best result.

To test the stability of the score, we then conducted the same experiment on the test data. The test set size shrinks to only 10 questions after taking the joint neighbourhood, so the statistical power is not high. However, the obtained accuracy of 0.8 indicates that there was no serious overfitting to the training set. In the next experiments we will thus use the combination of the harmonic mean and degree penalty.

TABLE II
COMPARISON OF THE ACCURACY OF DIFFERENT COMBINATIONS OF SCORING METHODS FOR CANDIDATE WORDS.

| | $\alpha(a,T)$ | | |
|---|---|---|---|
| $\beta(a)$ | **MINW** | **AVGW** | **HARM** |
| **Constant** | 0.72 | 0.72 | 0.76 |
| **DEG** | 0.86 | 0.86 | **0.90** |
| **DEGL** | 0.76 | 0.76 | 0.83 |

TABLE III
A SAMPLE OF ARTIFICIALLY GENERATED GENERALIZED RAT QUESTIONS.

| **Seed Word** | **Cue Word 1** | **Cue Word 2** | **Cue Word 3** |
|---|---|---|---|
| **imperialism** | colonialism | lenin | american |
| **missile** | warhead | defense | flight |
| **packaging** | product | paper | artwork |
| **slope** | steep | ski | western |
| **medley** | relay | yankovic | beatles |
| **far** | north | greater | moon |
| **kpmg** | firm | young | report |
| **concert** | band | hall | benefit |

We next analyse the generalized RAT creation process, as an approximation of a creative discovery task. To test the sanity of this method we conducted the following experiment. We chose 1000 random words which each had at least 3 mutually unconnected neighbours in the co-occurrence graph. For each such random word we selected 3 cue words by using the RAT creation process described above. We then solved the RAT question given the 3 cue words, and compared if the answer thus obtained was identical to the original seed word. In 97% of the cases the results were same for both methods, indicating consistency of the methodologies.

Finally, a sample of such artificially created generalized RAT questions is shown in Table III. Subjectively judging, they seem to match quite well classical criteria of creativity, such as the Torrance Tests of Creative Thinking [20]. The RAT creation method could be considered to exhibit *fluency* by producing a number of relevant cue words (and more could be easily generated), *flexibility* by discovering cue words that provide complementary contexts or meanings for the seed word, as well as *originality* by providing relatively rare words. Additionally, *elaboration* could potentially be achieved by using the co-occurrence network to describe the contexts for the various associations.

## VI. RELATED WORK

### A. Measuring Associations Between Terms

The idea of the distributional hypothesis is that words which co-occur in similar contexts tend to have similar meanings [21]. This was nicely put by Firth in 1957: "You shall know a word by the company it keeps" [22]. Followed by these ideas, the semantic similarity between words is calculated by their co-occurrence in documents.

Even if relatively few methods have been proposed for automatic construction of networks of terms, literature on co-occurrence or collocation statistics is abundant. Such measures can be used in an obvious way to build a network of terms. We only review some representative methods here.

*Log-likelihood ratio* is a non-parametric statistical test for co-occurrence analysis. Using log-likelihood ratio for word co-occurrence analysis was proposed by Dunning [12] who showed, in particular, that log-likelihood ratio does not overestimate the importance of very frequent words like some other measures.

*Latent Semantic Analysis* [23] aims to find a set of concepts (instead of terms) in a corpus using singular value decomposition. The semantic similarity (relatedness) of two words can then be estimated by comparing them in the concept space. Latent semantic analysis has then evolved to *Probabilistic Latent Semantic Analysis* [24] and later to *Latent Dirichlet Allocation* [25]. Probably any of these methods could be used to derive co-occurrence networks.

### B. Creative Association Discovery

Several papers have been published on supporting creativity by discovering links between concepts. In creative biological problem solving, for instance, Mozetic et al. [26] propose a method for finding unexpected links between concepts from different contexts. Examples of methods more directly based on link prediction in heterogeneous networks are given by Eronen and Toivonen [27].

## VII. DISCUSSION

Making the 'right' choices is often much easier than making choices which are less rational, but do still make sense. This is what this paper is all about – given constraints, our goal is to propose something as a result which satisfies these constraints, but at the same time is thought-provoking. In creative support systems, one of the purposes is to encourage the user to think more broadly. One way for doing this is by giving answers, which are related to the question, but the relation itself is subtle enough, to induce creative thoughts.

In the paper we briefly described RATs and their underlying mechanisms. We showed that by using 2-grams and a simple probabilistic model it is possible to solve these tests with a good accuracy.

We also described a methodology for creating a network of more general associations than the 2-gram language model could provide. As a ground for the creative inference, we showed that the connections in this network tend to make sense and we can assume that if two words are connected by an edge, they are also semantically related.

Our main contribution is translating the principles which we established in the probabilistic framework for solving RATs to the generalized model with co-occurrence networks. An empirical result was that the associations generated from the network seem to exhibit creativity.

In the future our goal is to validate the methods more objectively, e.g., by some user testing. We plan to test and compare different language models (e.g., LSI, LDA) and provide more in depth analysis for the creative association discovery. Finally, we are planning to use these methods in tasks which relate to lexical creativity (e.g., automatic poetry generation) and in possible lexical creativity support systems (e.g., slogan wizard).

*Answer to the RAT Question in the Introduction*

The intended answer word related to 'coin', 'quick', and 'spoon' is 'silver'.

*Acknowledgements:* This work has been supported by the Algorithmic Data Analysis (Algodan) Centre of Excellence of the Academy of Finland.

## REFERENCES

[1] S. Mednick, "The associative basis of the creative process." *Psychological review*, vol. 69, no. 3, p. 220, 1962.

[2] J. M. Toivanen, H. Toivonen, A. Valitutti, and O. Gross, "Corpus-based generation of content and form in poetry," in *International Conference on Computational Creativity*, Dublin, Ireland, 2012, pp. 175–179.

[3] N. Gupta, Y. Jang, S. Mednick, and D. Huber, "The road not taken creative solutions require avoidance of high-frequency responses," *Psychological Science*, 2012.

[4] M. T. Mednick and F. M. Andrews, "Creative thinking and level of intelligence," *Journal of Creative Behavior*, vol. 1, pp. 428–431, 1967.

[5] G. Forbach and R. Evans, "The remote associates test as a predictor of productivity in brainstorming groups," *Applied Psychological Measurement*, vol. 5, no. 3, pp. 333–339, 1981.

[6] K. S. Bowers, G. Regehr, C. Balthazard, and K. Parker, "Intuition in the context of discovery," *Cognitive Psychology*, vol. 22, pp. 72–110, 1990.

[7] S. Topolinski and F. Strack, "Where theres a willtheres no intuition: The unintentional basis of semantic coherence judgments," *Journal of Memory and Language*, vol. 58, pp. 1032–1048, 2008.

[8] E. Vul and H. Pashler, "Incubation benefits only after people have been misdirected," *Memory & Cognition*, vol. 35, pp. 701–710, 2007.

[9] K. Bowers, G. Regehr, C. Balthazard, and K. Parker, "Intuition in the context of discovery," *Cognitive psychology*, vol. 22, no. 1, pp. 72–110, 1990.

[10] S. Mednick and M. Mednick, *Examiner's Manual, Remote Associates Test: College and Adult Forms 1 and 2*. Houghton Mifflin, 1967.

[11] J.-B. Michel, Y. K. Shen, A. P. Aiden, A. Veres, M. K. Gray, T. G. B. Team, J. P. Pickett, D. Hoiberg, D. Clancy, P. Norvig, J. Orwant, S. Pinker, M. A. Nowak, and E. L. Aiden, "Quantitative analysis of culture using millions of digitized books," *Science*, vol. 331, no. 6014, pp. 176–182, 2011.

[12] T. Dunning, "Accurate methods for the statistics of surprise and coincidence," *Computational linguistics*, vol. 19, no. 1, pp. 61–74, 1993.

[13] P. Domingos and M. Pazzani, "On the optimality of the simple bayesian classifier under zero-one loss," *Machine learning*, vol. 29, no. 2, pp. 103–130, 1997.

[14] S. Bird, E. Klein, and E. Loper, *Natural Language Processing with Python*. O'Reilly Media, 2009.

[15] E. Bowden and M. Jung-Beeman, "Normative data for 144 compound remote associate problems," *Behavior Research Methods*, vol. 35, no. 4, pp. 634–639, 2003.

[16] D. Dailey, "An analysis and evaluation of the internal validity of the remote associates test: What does it measure?" *Educational and Psychological Measurement*, vol. 38, no. 4, pp. 1031–1040, 1978.

[17] G. A. Miller, "Wordnet: a lexical database for english," *Communications of the ACM*, vol. 38, no. 11, pp. 39–41, 1995.

[18] D. Gentner, "Why Nouns Are Learned Before Verbs: Linguistic Relativity Vs. Natural Partitioning," in *Language Development, vol.2: Language, cognition and culture*, S. Kuczaj, Ed. Hillsdale, NJ: Erlbaum, 1982, pp. 301–334.

[19] A. Budanitsky and G. Hirst, "Semantic distance in wordnet: An experimental, application-oriented evaluation of five measures," in *Workshop on WordNet and Other Lexical Resources*, vol. 2, 2001.

[20] E. Torrance, *Torrance Tests of Creative Thinking: Norms-technical Manual. Research Edition. Verbal Tests, Forms A and B. Figural Tests, Forms A and B*. Personnel Press, 1966.

[21] Z. Harris, "Distributional structure," *Word*, vol. 10, no. 23, pp. 146–162, 1954.

[22] J. R. Firth, "A synopsis of linguistic theory 1930-55." vol. 1952-59, pp. 1–32, 1957.

[23] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman, "Indexing by latent semantic analysis," *Journal of the American Society for Information Science*, vol. 41, no. 6, pp. 391–407, 1990.

[24] T. Hofmann, "Probabilistic latent semantic indexing," in *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. New York, NY, USA: ACM, 1999, pp. 50–57.

[25] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.

[26] I. Mozetic, N. Lavrac, V. Podpecan, P. K. Novak, H. Motaln, M. Petek, K. Gruden, H. Toivonen, and K. Kulovesi, "Bisociative knowledge discovery for microarray data analysis," in *The 1st International Conference on Computational Creativity (ICCC-X)*, Lisbon, Portugal, 2010, pp. 190–199.

[27] L. Eronen and H. Toivonen, "Biomine: Predicting links between biological entities using network models of heterogeneous database," *BMC Bioinformatics*, vol. 13, no. 119, 2012.

# Paper II

II

Juhani Huovelin, Oskar Gross, Otto Solin, Krister Linden, Sami Maisala, Tero Oittinen, Hannu Toivonen, Jyrki Niemi, Miikka Silfverberg

**Software Newsroom – An Approach to Automation of News Search and Editing**

In *Journal of Print Media Technology research*, 2(3), 141-156, IARIGAI, 2013.



*The author of the thesis played a major role in writing Sections 2.4.2–2.4.4 and 3.2.*

JPMTR 022 | 1311
UDC 054:004.4

Original scientific paper
Received: 2013-06-28
Accepted: 2013-11-07

# Software Newsroom - an approach to automation of news search and editing

*Juhani Huovelin*[1], *Oskar Gross*[2], *Otto Solin*[1], *Krister Lindén*[3], *Sami Maisala*[1], *Tero Oittinen*[1], *Hannu Toivonen*[2], *Jyrki Niemi*[3], *Miikka Silfverberg*[3]

[1] Division of Geophysics and Astronomy
Department of Physics, University of Helsinki
FIN-00560 Helsinki, Finland

E-mails: juhani.huovelin@helsinki.fi
otto.solin@helsinki.fi
sami.maisala@helsinki.fi
tero.oittinen@helsinki.fi

[2] Department of Computer Science and HIIT
University of Helsinki
FIN-00014 Helsinki, Finland

E-mails: oskar.gross@cs.helsinki.fi
hannu.toivonen@cs.helsinki.fi

[3] Department of Modern Languages
University of Helsinki
FIN-00014 Helsinki, Finland

E-mails: krister.linden@helsinki.fi
jyrki.niemi@helsinki.fi
miikka.silfverberg@helsinki.fi

**Abstract**

We have developed tools and applied methods for automated identification of potential news from textual data for an automated news search system called Software Newsroom. The purpose of the tools is to analyze data collected from the internet and to identify information that has a high probability of containing new information. The identified information is summarized in order to help understanding the semantic contents of the data, and to assist the news editing process.

It has been demonstrated that words with a certain set of syntactic and semantic properties are effective when building topic models for English. We demonstrate that words with the same properties in Finnish are useful as well. Extracting such words requires knowledge about the special characteristics of the Finnish language, which are taken into account in our analysis. Two different methodological approaches have been applied for the news search. One of the methods is based on topic analysis and it applies Multinomial Principal Component Analysis (MPCA) for topic model creation and data profiling. The second method is based on word association analysis and applies the log-likelihood ratio (LLR). For the topic mining, we have created English and Finnish language corpora from Wikipedia and Finnish corpora from several Finnish news archives and we have used bag-of-words presentations of these corpora as training data for the topic model. We have performed topic analysis experiments with both the training data itself and with arbitrary text parsed from internet sources. The results suggest that the effectiveness of news search strongly depends on the quality of the training data and its linguistic analysis.

In the association analysis, we use a combined methodology for detecting novel word associations in the text. For detecting novel associations we use the background corpus from which we extract common word associations. In parallel, we collect the statistics of word co-occurrences from the documents of interest and search for associations with larger likelyhood in these documents than in the background. We have demonstrated the applicability of these methods for Software Newsroom. The results indicate that the background-foreground model has significant potential in news search. The experiments also indicate great promise in employing background-foreground word associations for other applications.

A combined application of the two methods is planned as well as the application of the methods on social media using a pre-translator of social media language.

**Keywords:** social media, data mining, topic analysis, machine learning, word associations, linguistic analysis

## 1. Introduction

The vast amount of open data in the internet provides a yet ineffectively exploited source of potential news. Social media and blogs have become an increasingly useful and important source of information for news agencies and media houses. In addition to the news collected, edited and reported by traditional means, i.e., by news agencies, the information in a news-room consists of different types of user inputs. In the social media there is a large amount of user comments and reactions triggered by news stories. Also, fresh article manuscripts and

other types of material can be produced by basically anyone by submitting the information to the internet. As a means of collecting news, this material is already in use by commercial media companies, especially in a hyperlocal media context (e.g., newspapers that discuss local issues).

While this editorial strategy is considerably more advanced than the way news were produced a decade ago, the work still includes manual work that could be automated and the use of open data available in the internet is usually very inefficient. It also does not make much sense to engage humans for browsing internet data, a job that can be done much more efficiently and tirelessly by a machine.

Thus, intelligent computer algorithms that monitor internet data and hunt for anomalies and changes are becoming an increasingly exploited means of news and trend detection. Other applications for the same methodologies are public opinion analysis and forecasting the results of elections. Examples of even more advanced intelligence in prediction would be calls to events, which can be predecessors of demonstrations or even an uprising, and indication of a meeting between high level politicians based on their plans to travel to the same place at the same time.

The same methods, when combined with fusion of heterogeneous data, can help improving the quality and widening the scope of news by the enrichment of existing news material with relevant background information and other associated material (e.g., history, pictures, digital video material). In principle, using the same methodology it is also possible to follow the discussions raised by published news articles and thus automatically collecting feedback from the audience.

Examples of internet services developed for the above purposes are Esmerk Oasis (Comintelli, 2013) and Meltwater Buzz (Meltwater, 2013). Esmerk Oasis is a web-based market intelligence solution. Its services include customized global business information with the possibility of importing complementary information from other sources as well as sharing and distribution of information across the client organization. Meltwater Buzz is a social media monitoring tool that has capabilities for tracking and analyzing user-generated content on the web. Google has also developed several services that perform similar tasks.

Considering the purpose and goal of an automated news search and analysis process, a baseline approach to analyzing text material and creating a short descriptiton of its contents is to simulate the traditional process of news production. The analysis of the material should tell you *what, who, where, and when*? Methodologically, the most challenging task is to find a systematic way of defining the answer to the question *what*, since it includes the need to recognize and unambiguously describe an unlimited range of *topics,* not just individual words. A topic is usually defined as "a set of news stories that are strongly related by some seminal real-world event", and an event is defined as "something (non-trivial) happening in a certain place at a certain time" (Allan, 2002). As an example, the recent meteorite impact in Chelyabinsk was the event that triggered the asteroid impact, natural catastrophes, and doomsday topic. All stories that discuss the observations, consequences, witnesses, probabilities and frequency of such events etc., are part of the topic.

The answers to the other questions*, who, where* and *when,* can be traced by searching named entities and various time tags and information. In practical application to, e.g., social media, however, the latter questions may also pose a significant challenge for an automated approach, since social media language does not obey common rules.

The quality of the language is often very poor, since it may include many local and universal slang words, acronyms and idioms that are known by only a limited local community, and also numerous typing errors.

Blogs are considerably less difficult in this respect, since most of the text in them is in fairly well written standard language.

Methods for event and trend detection and analysis in large textual data include *static and dynamic component models* which are well suited for news search and detection in the internet. Static models are simpler to use and give results that are easier to interpret. A potential disadvantage is that newly emergent trends may remain undetected if the training data for the model is not sufficiently extensive, leading to the model being not generic enough. A dynamic model, on the other hand, is updated continuously in order to keep up with possible emergent topics. Its usage, however, is not as straightforward as that of static models since the emergent trends may be described in terms of dynamic components whose semantics is not yet well understood.

An example of a static component model is Principal Component Analysis (PCA). PCA was invented in 1901 by Pearson (1901). PCA can be performed by eigenvalue decomposition of a data covariance matrix or singular value decomposition of a data matrix. The singular value decomposition of the word count matrix is also cal-led Latent Semantic Indexing (LSI) (Berry, Dumais and O'Brien, 1994; Hofmann, 1999).

We have developed algorithms for automated analysis of text in, e.g., social media, blogs and news data with the aim of identifying "hot" topics that are potential news. We here present the methods and show results of their application using real data.

## 2. Method

2.1 Combining methods

We apply two different approaches that are combined in order to achieve a clearer recognition of potential news in an arbitrary text under analysis. The first method is topic mining including advanced linguistic analysis for named entity recognition. The topic model is based on Multinomial Principal Component Analysis, MPCA (Kimura, Saito and Uera, 2005; Buntine and Jakulin, 2006). While topics are considered to be different kinds of objects than named entities (e.g., Newman et al., 2006), they can be combined in the creation of a probabilistic topic model. The second approach, association analysis, takes into account the word co-occurrences in the document and uses statistics to look for novel word associations in a set of documents. These associations are used for Software Newsroom applications, such as diverging (association) word clouds and automatic summary generation. The background model calculation uses a method based on the log-likelihood ratio (LLR) (Dunning, 1993). This is described in more detail by Toivonen et al. (2012). By extending the ideas in the latter approach, we propose a method for detecting novel word associations.

2.2 Topic mining

For generating static component models from textual data, we use the statistical generative model called *Multinomial principal component analysis* (MPCA) (Buntine and Jakulin, 2006). MPCA is used to model the data in order to obtain a comprehensive understanding of the contents of the data sources in the form of semantically meaningful components or topics.

In our application, the topic model includes four categories of common words (nouns, verbs, adjectives, adverbs) where the nouns are not named entities, and four categories of named entities (persons, places, organisations, miscellaneous), where the miscellaneous category includes all named entities that do not belong to the other three categories. It has been shown that these eight categories are effective for building topic models for English (e.g., Newman et al., 2006). An important aspect of our research is to verify that the linguistic categories can be identified in a language-independent way. We demonstrate this by extracting the eight categories of words from text in Finnish - a language completely unrelated to English.

Let $\mathbf{D}$ be a $d \times N$ matrix representing the training data (documents) as a "bag-of-words", $\mathbf{M}$ a $d \times K$ matrix of documents represented in terms of topics, and $\mathbf{\Omega}$ a $K \times N$ matrix of topics represented in terms of words, where $d$ is the number of documents in the training corpus, $N$ the size of the vocabulary and $K$ the number of topics ($K << N$). We extract from the training corpus two types of features: Part-of-speech tags (nouns, adjectives, verbs, adverbs) and Named Entities (locations, persons, organizations, miscellaneous). Thus, in our case, the vocabulary words are treated as eight multinomials. The aim is to represent the documents in terms of matrices $\mathbf{M}$ and $\mathbf{\Omega}$ as (Equation 1):

$$\mathbf{D} \approx \mathbf{M} \times \mathbf{\Omega}. \qquad [1]$$

In other words, the data is transformed into a lower dimensional space, where documents are represented in terms of topics. The topics are then represented in terms of words. The matrices $\mathbf{M}$ and $\mathbf{\Omega}$ give the probabilities of topics given a document and words given a topic, respectively.

The process for generating the model with MPCA is as follows (Buntine and Jakulin, 2004).

1. A total of $N$ words are partitioned into $K$ partitions $\mathbf{c} = c_1, c_2, \ldots, c_K$ where $\sum_{k=1}^{K} c_k = N$. $N$ is the size of the vocabulary and $K$ the number of topics. The partitioning is done using a latent proportion vector $\mathbf{m} = (m_1, m_2, \ldots, m_K)$. The vector $\mathbf{m}$ for each document forms the $d$ rows in matrix $\mathbf{M}$.

2. Words are sampled from these partitions according to the multinomial for each topic producing a bag-of-words representation $\mathbf{w}_{k,\cdot} = (w_{k,1}, w_{k,2}, \ldots, w_{k,N})$ for each partition $k$.

3. Partitions are combined additively to produce the final vocabulary $\mathbf{r} = (r_1, r_2, \ldots, r_N)$ by totaling the corresponding counts in each partition,

   $r_n = \sum_{k=1}^{K} w_{k,n}$

The above process is described by the following probability model (Equations 2),

$$\begin{aligned} &\mathbf{m} \sim \mathit{Dirichlet}(a) \\ &\mathbf{c} \sim \mathit{Multinomial}(\mathbf{m}, N) \\ &\mathbf{w}_{k,\cdot} \sim \mathit{Multinomial}(\mathbf{\Omega}_{k,\cdot}, c_k) \quad \text{for } k = 1,\ldots,K \end{aligned} \qquad [2]$$

Its estimation is done through a Gibbs sampler (Buntine and Jakulin, 2006). In Gibbs sampling, each unobserved variable in the problem is resampled in turn according to its conditional distribution. Its posterior distribution conditioned on all other variables is computed, and then a new value for the variable using the posterior is sampled. In each cycle of the Gibbs algorithm the last $\mathbf{c}$ for each document is retrieved from storage and then, using a Dirichlet prior for rows of $\mathbf{\Omega}$, the latent component variables $\mathbf{m}$ and $\mathbf{w}$ are sampled. The latent variables are $\mathbf{m}$ and $\mathbf{w}$, whereas $\mathbf{c}$ is derived. As a result we get estimates of the matrices $\mathbf{M}$ and $\mathbf{\Omega}$ in

Equation 1. In the context of an MPCA model, these are estimates of the distribution of documents over topics and topics over words respectively.

There are different ways of estimating topic strengths in a single document given the model created by MPCA. The method applied here is cosine similarity between document vector *d* and topic $\omega_k$ as

$$\mathrm{sim}(\mathrm{d}, \omega_\mathrm{k}) = \frac{\mathrm{d}\cdot\omega_\mathrm{k}}{||\mathrm{d}||||\omega_\mathrm{k}||}. \qquad [3]$$

A topic model including a desired number of yet unnamed topics is first created by the above method (Equations 1 and 2) using a bag-of-words presentation of the training corpus. This is then ready for application in topic analysis of arbitrary text. The topic analysis includes automated simultaneous identification of the topic, person, place, organization, and event in an arbitrary blog article, a discussion thread in social media or an RSS feed, etc. This is done by statistical comparison, or projection (Equation 3), of the new text against the topic model. By tracking the history of the frequency of occurrence of similar stories (which belong to the same topic, i.e., resemble each other), the software can identify the trend of a topic. A statistically significant deviation from the trend in a short time period gives a hint that the source texts that caused this deviation may include a news candidate. In the present analysis, we use Gaussian statistics and the criteria for significant deviation is $3\sigma$. This applies to generic topics, but for words and events that are generally interesting from a newsroom perspective, such as VIPs, accidents, crimes, and natural disasters, all occurrences are tagged as being potential news topics. The method is, on a general level, similar to the approach of Newman et al. (2006) but it includes advanced features developed for practical applicability in a newsroom environment.

### 2.3 Linguistic analysis for named-entity recognition

#### *2.3.1 English vs. Finnish words and named entities*

When adapting topic identification from one language to another, it is necessary to be aware of what units of the language have been chosen and how similar units can be identified in another language. All language analysis methods do not produce the same output granularity. In the following, we outline the units that have been found effective in English and how corresponding units can be identified in Finnish to highlight some of the essentials that need to be considered when choosing linguistic analysis software to adapt to another language.

The most striking difference between Finnish and English is the number of inflected forms in Finnish. There are roughly 2000 forms for each noun, 6000 for each adjective and 12000 for each verb. The characteristics of these forms and their usage in Finnish has been extensively documented in an online Finnish grammar, "Iso suomen kielioppi" (Hakulinen et al., 2004). It is not possible to only chop off word endings, because changes also take place in the stem when inflectional morphemes are added, e.g., "nojatuoli" [armchair], "nojatuoleja" [armchairs], "nojatuoleissa" [in the armchairs], "nojatuoleissani" [in my armchairs], "nojatuoleissanikin" [also in my armchairs]. In practice, Finnish words can represent expressions that in English are rendered as a phrase, so Finnish needs a morphological analyzer to separate the base form from the endings. As a bonus to the morphological processing, many of the inflectional morphemes that are separated from the base form correspond to stop-words in English.

In addition to gluing inflectional morphemes onto the words, Finnish also has the orthographic convention of writing newly formed compound words without separating spaces, i.e., "nappanahkanojatuoli" [calf-skin armchair]. The English word *armchair* can be seen as a compound as well, but typically a modern *armchair* is not perceived only as a *chair* with *arm*rests, but as something slightly more comfortable, so the *armchair* has a lexicalized meaning of its own. This means that, for newly coined non-lexicalized compounds, it is essential that the morphological analysis separates the non-lexicalized parts in Finnish; otherwise the compositional meaning is lost. Long newly formed compounds also lack predictive power since they are rare by definition whereas the compound parts may give essential clues to the topic of the narrative. It should be noted that a Finnish writer could also choose to write "nappanahkainen nojatuoli" [calf-skin armchair], and with the increased influence of English, this convention is perceived as more readable.

The structure of named-entities, i.e., places, organizations, persons and other names, follows the conventions mentioned for regular words. In particular, place names tend to be written in one or two words at most because they are of older origin. Person names have a similar structure as in English with given name and surname. However, long organization names tend to be formulated as multi-word expressions following newer writing tendencies.

#### *2.3.2 Named-entity recognition in Finnish*

For named-entity recognition in many languages it is possible to do string matching directly on the surface forms in written text. In Finnish, we need more in-depth morphological processing to deal with the inflections and the compound words. For out-of-vocabulary words, we also need guessers. To cope with morphological ambiguity, we need a tagger before we can apply named-entity recognition.

Language technological applications for agglutinating languages such as Finnish, benefit greatly from high co-

verage morphological analyzers providing word forms with their morphological analyses, e.g.,

"nojatuole+i+ssa+ni+kin : nojatuoli *Noun Plural 'In' My' 'Also'* " [also in my armchairs].

However, morphological analysis makes applications dependent on the coverage of the morphological analyzer. Building a high coverage morphological analyzer (with an accuracy of over 95%) is a substantial task and, even with a high-coverage analyzer, domain-specific vocabulary presents a challenge. Therefore, accurate methods for dealing with out-of-vocabulary words are needed.

With the Helsinki Finite-State Transducer (HFST) tools (Lindén et al., 2011), it is possible to use an existing morphological analyzer for constructing a morphological guesser based on word suffixes. Suffix based guessing is sufficient for many agglutinating languages such as Finnish (Lindén and Pirinen, 2009), where most inflection and derivation is marked using suffixes. Even if a word is not recognized by the morphological analyzer, the analyzer is likely to recognize some words which inflect similarly as the unknown word. These can be used for guessing the inflection of the unknown word.

Guessing of an unknown word such as "twiitin" (the genitive form of "twiitti", tweet, in Finnish) is based on finding recognized word forms like "sviitin" (genitive form of "sviitti" hotel suite in Finnish), that have long suffixes such as "-iitin", which match the suffixes of the unrecognized word. The longer the common suffix, the likelier it is that the unrecognized word has the same inflection as the known word. The guesser will output morphological analyses for "twiitin" in order of likelyhood.

A morphological reading is not always unique without context, e.g., "alusta" can be an inflected form of "alku" [beginning], "alunen" [plate], "alustaa" [found] or "alus" [ship]. To choose between the readings in context it is possible to use, e.g., an hidden Markov model (HMM) which is essentially a weighted finite-state model. Finite-state transducers and automata can more generally be used for expressing linguistically relevant phenomena for tagging and parsing as regular string sets, demonstrated by parsing systems like Constraint Grammar (Karlsson, 1990) which utilizes finite-state constraints. Weighted machines offer the added benefit of expressing phenomena as fuzzy sets in a compact way.

Using tagged input, a named entity recognizer (NER) for Finnish marks names in a text, typically with information on the type of the name (Nadeau and Sekine, 2007). Major types of names include persons, locations, organizations and events. NER tools often also recognize temporal and numeric expressions. NER tools typically use gazetteers, lists of known names, to ensure that high-frequency names are recognized with the correct type. For Finnish, the gazetteer is included in the morphological analyzer because names inflect. In addition, names and their types can be recognized based on internal evidence, i.e., the structure of the name itself (e.g., ACME Inc., where *Inc.* indicates that ACME denotes a company), or based on external evidence, i.e., the context of the name (e.g., *works for* ACME; ACME *hired a new CEO*) (MacDonald, 1996).

## 2.4 Association analysis

### *2.4.1 Extracting word associations*

One of the goals of the Software Newsroom is to give an overview of popular topics discussed in the internet communities. This gives journalists an opportunity to react to these topics on a short notice. In the Software Newsroom, word association analysis is used for detecting novelty in the contents of a given set of documents. For instance, consider a web forum where people discuss about different topics, e.g., fashion, technology, politics, economics, computer games, etc. As an example, consider that a new smartphone *SoftSmart* has a feature which automatically disables GPS when you are indoors. It turns out that it has a bug, and in some very specific cases (e.g., for instance when you are on the top floor of a building) it starts to drain your battery because the signal strength is varying. It is reasonable to believe that many *SoftSmart* users will go to web forums and start discussing about the problems. Even more, it might turn out that there is an easy fix available and this is posted somewhere to the forum. The problem is, that there are thousands of similar problems being discussed all over the world, so it is not feasible for a technology journalist to monitor all the forums.

If we could automatically detect this as a trendy topic, then this information would be invaluable for a technology journalist, as she/he could then learn more about this and write a news story. From the language analysis point of view, the text written by people in web forums and other web communities introduce problems - the text contains slang, typing errors, words from different languages, etc. These aspects add another goal for the association analysis - our goal is to develop a method which is not fixed to any specific vocabulary. Our idea is to analyze the associations between words and to look for such associations which are novel with regards to other documents.

Considering the *SoftSmart* example, there are words which co-occur in sentences but the association between them is most probably very common, such as *SoftSmart - battery, battery - drain, SoftSmart - GPS,* etc. For the *SoftSmart* case, the words for which the association is rather specific could be *battery - floor, floor - drain, battery - top, Softsmart - floor* and so on. In association analysis, our goal is to automatically detect the latter ones. Note,

that the association itself might be surprising, though it is between very common words, like 'battery' and 'floor'.

Finding associations between concepts which can be represented as sets of items is a very much studied area which originates from the idea of finding correlations in market basket data (Agrawal et al., 1993). The bag-of-words model of representing documents as sets of unordered words is a common concept in information retrieval (Harris, 1954; Salton, 1993). Often, the bag-of-words model is used together with the tf-idf measure that measures word specificity with respect to the document corpus (Salton, 1993).

Analysing word associations in document is not a new idea. There are various word association measures available - the log-likelihood ratio test (Dunning, 1993), the chi-squared test, Latent Semantic Indexing (Dumais et al., 1988), pointwise mutual information (Church and Hanks, 1990), Latent Dirichlet Allocation (Blei et al., 2003), etc. There is also a method for pairs, which is inspired by tf-idf, called tpf-idf-tpu which is a combination of using term pair frequency, its inverse document frequency and the term pair uncorrelation for determining the specific pairs of a document (Hynönen et al., 2012).

In this paper, we present a method for analyzing and representing documents on the word association level. We use the log-likelihood ratio as the basis for our method. As mentioned before, finding associations between documents is a very common concept and the main goal for all the methods is to discover statistically strong associations between words. In some instances we are interested in such associations that are specific to a certain set of document. For instance, consider a set of documents about the singer Freddie Mercury. Imagine, that we create pairs of all the words which co-occur in the same sentence and the weight is determined by their co-occurrence statistics (e.g., weighted by the log-likelihood ratio test). Now, if we order the pairs decreasingly by association strength, we will most probably obtain pairs such as: 'freddie'-'singer', 'freddie'-'aids', 'freddie'-'bohemian', 'aids'-death', 'aids'-sick' etc. The point here is that some of the associations are important and relevant to the document set (e.g., the first two). On the other hand, the last two associations between words are very common. And this is defines our goal - we are looking for word associations which are specific to a certain set of documents *and* at the same time are uncommon with respect to other documents.

In the following, we introduce methods for extracting word associations that are specific to a set of documents. For this we define two concepts: *background associations*, which are the common associations between words and *foreground associations*, where the weight is higher for associations that are novel with respect to the background associations.

After we have given an overview of the core methods, we will present applications of these models in the Software Newsroom. First we will look at the possible representations of foreground associations and discuss the possible usefulness of explicit graph representations. Then we will provide an idea of diverging word clouds which illustrate word associations rather than frequencies. Finally, we propose a simple, yet intuitive way of generating summaries of a set of documents by using foreground associations.

### *2.4.2 Background associations*

*Background associations* represent common-sense associations between terms, where the weight depends on the strength of the association. For example, the connection between the words 'car' and 'tire' should be stronger than the connection between 'car' and 'propeller'. In our methodology, these associations are extracted from a corpus of documents, motivated by the observation that co-occurrence of terms tends to imply some semantic relation between them (slightly misleadingly often called semantic similarity). Background associations are calculated by identifying words which co-occur in the same sentence. The strength between the words is calculated using the log-likelihood method (Dunning, 1993; Toivonen et al., 2012). The latter paper describes how the word associations are calculated and also demonstrates the relationship between such associations and relations in WordNet (Miller, 1995).

### *2.4.3 Foreground associations*

In contrast to the common associations in the background, *foreground associations* represent novel associations of a (small) set $F$ of documents called the foreground documents.

However, the background associations do have a central role here: they tell us what is known, so that we can infer what is novel in any given document. The weighting scheme in the foreground also uses the log-likelihood ratio test. However, now we use the background to obtain the expected number of co-occurrences and to see how much the observed number of co-occurrences in the foreground documents F deviates from it. The result of this test gives higher weights to those term pairs that are more frequent in the foreground $F$ than they are in the background, i.e., especially those pairs which have a small likelihood of occurring together in the background.

In our implementation of this idea, the foreground weights are based on the log-likelihood ratio where the alternative model is based on the foreground documents $F$ and the null model on the background corpus $C$.

Let parameters $p_{ij}^{null}$ be the maximum likelihood parameters for the corpus C, i.e., (Equation 4):

$$
\begin{aligned}
p_{11}{}^{\text{null}} &= p(x \wedge y;\ C) \\
p_{21}{}^{\text{null}} &= p(x \wedge \neg y;\ C) \\
p_{12}{}^{\text{null}} &= p(\neg x \wedge y;\ C) \\
p_{22}{}^{\text{null}} &= p(\neg x \wedge \neg y;\ C)
\end{aligned}
\qquad [4]
$$

where *x* and *y* denote the events that "word x (respectively y) occurs in a randomly chosen sentence (of the given corpus)". For the background associations, these parameters are used as the alternative model, and here they are used as the null model. Set the alternative model parameters $p_{ij}$ in turn to be the maximum likelyhood parameters for the document set F (Equations 5),

$$
\begin{aligned}
p_{11} &= p(x \wedge y;\ F) \\
p_{21} &= p(x \wedge \neg y;\ F) \\
p_{12} &= p(\neg x \wedge y;\ F) \\
p_{22} &= p(\neg x \wedge \neg y;\ F)
\end{aligned}
\qquad [5]
$$

The log-likelihood ratio (LLR) for the foreground associations is then computed according to Equation 6.

$$\mathrm{LLR}(x,y) = -2\sum_{i=1}^{2}\sum_{j=1}^{2} k_{ij}\log(p_{ij}{}^{\text{null}}/p_{ij}) \quad [6]$$

The foreground association weights are assigned by this LLR function. Using this function, we give higher weight to such associations which are more likely to appear in the foreground and less likely in the back-ground. Note, that the log-likelihood ratio could be also negative. In this case the word association is weaker in the foreground than in the background. In our work we omit associations with negative weights.

*2.4.4 Applications*

In the following, we present applications in the Software Newsroom that employ the background/foreground associations method. In the first application we describe, the associations in the set of documents are represented as an explicit graph. In the remainder of the subsection we will demonstrate two different Software Newsroom applications - diverging word cloud generation and document summarization. For a single document experiment we will use the English Wikipedia as the background corpus and a story from BBC: "Google tests balloons to beam internet from near space" (Kelion, 2013) as the foreground document we are interested in.

The simplest way of representing the information is by showing the top-k (where k is an integer) word pairs of the news story. In order to show the differences between a standard co-occurrence calculation and our foreground method in we have, in Table 1, presented the top-5 pairs of the Google news story. For comparison, the left column lists the most strongly associated word pairs as measured using standard methods, while the right column lists the top-5 pairs obtained by the foreground method.

The pairs suggest that the foreground method is able to grasp the main associations of the news story better than the classical co-occurrence measures. By this we mean that the associations of the foreground contain more relevant associations, such as 'superpressure' and 'balloons' or 'google' and 'balloons'. Representing associations as a simple list makes them individually easy to understand, but does not give a picture of the network of connections. On the other hand, a graphical representation (Figure 1) of this network may be difficult, especially for novice users. On the other hand, when a user is familiar with such data representation it gives a quick and general view of the data. In our work, the explicit graph is not a favored method for illustrating or representing information. We put more emphasis on designing methods that employ the foreground graph.

*Table 1: The top-5 pairs for the BBC news story "Google tests balloons to beam internet from near space". The left column shows pairs calculated using the standard co-occurrence calculation method (log-likelihood ratio); the right column shows the top-5 pairs obtained using the foreground association method*

| Long-likelihood ratio | Foreground method |
|---|---|
| plastic - made | superpressure - balloons |
| months - airborne | launched - new_zealand |
| suggested - atmosphere | google - balloons |
| special - fitted | suggested - atmosphere |
| force - air | force - air |

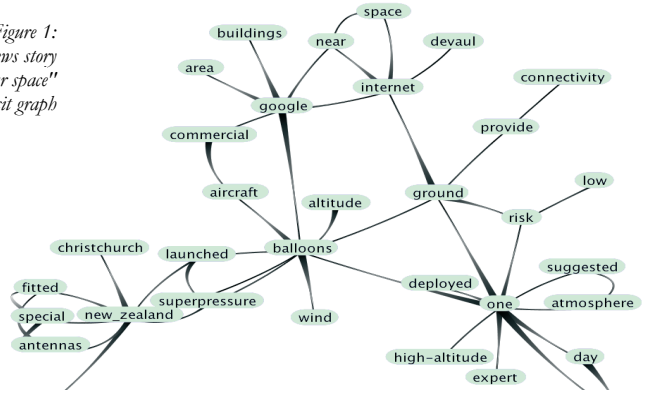
We now propose a new type of word clouds, *diverging (association) clouds,* that aim at helping users to explore the novel associative knowledge emerging from textual documents. Given a search term, the diverging cloud of a document highlights those words that have a special association with the search term. As a motivating application, consider word clouds as summaries of news stories. If the user has a special interest, say 'iPhone', we would first of all like the word clouds to be focused or conditioned on this search term, i.e., only show terms to which 'iPhone' is associated in the news story. Secondly, we would like to see only novel information about the iPhone, not the obvious ones such as 'Apple' and 'mobile'. The diverging clouds aim to do exactly this, directly based on the foreground associations of a news story as a representation of potentially new semantic associations. For a sample of diverging (association) clouds, see Figure 4 in the Results section.

In news, it is very common that the information on a certain event comes in over time. This is even more so for news published on the web or discussed in internet forums. For instance, considering an incident (e.g., the Boston Marathon bombing) which has a large impact and is related to many people, information and updates concerning the event are usually published frequently on news websites. For each news story update, most parts remain the same, some of the information changes, something is added, and something is removed. To

get an overall picture of the event, one should go through many articles and collect the new bits of information from each of them while at the same time most of the information is redundant.

*Automatic document summarization* is a method for overcoming this problem. The main goal of document summarization is to represent the information in a set of documents in a short and possibly non-redundant manner.

*Figure 1:*
*A subset of the foreground associations for the news story "Google tests balloons to beam internet from near space" represented as an explicit graph*




Two different approaches have been used in document summarization - generating text using the documents as reference (Hori et al., 2003; Knight et al., 2002) and selecting a representative set of sentences from the documents, e.g., by using Support Vector Machines (Yeh et al., 2005), Hidden Markov Models (Conroy and O'leary, 2001), and Conditional Random Fields (Sheh et al., 2007).

In this paper, we sketch a tentative method which uses sentence extraction for document summarization using foreground associations as reference. In the future, we plan to enhance this method and also combine it with text generation into a hybrid method. Sentence selection is based on the principle that our goal is to cover as much of the foreground associations as possible with the minimum number of sentences. The intuitive idea is that the foreground associations describe the most relevant aspects of the document set. Now, if these associations are covered by a certain set of documents, it is reasonable to assume that we have also captured the essentials of these documents.

A greedy algorithm for selecting the sentences is the following:

1. Select the strongest association *e* from the foreground associations.
2. Let *S(e)* be the set of all sentences which contain both words of *e*.
3. For each sentence *s* in *S(e)* calculate the score of the sentence as the sum of the foreground association weights for all the word pairs found in the sentence.
4. Output the highest scoring sentence *s**.
5. Remove *e* and all other pairs which appear in *s** from the foreground associations.
6. If the size of the summary is not sufficient and the foreground is not empty, go to (1).

The design of the algorithm follows the following two principles. First, the highest weighted pair should contain the most important information. This is the reason why we start looking for the sentences where the words of the highest association in the foreground appear. The second principle is that we do not want to include information which is already included. This is the reason for the step (5) above. If we are interested in penalizing sentences which contain pairs which are already covered, then in step (5) it is possible to change the values of the association weights into negative constants. For our experiments with summary generation on the topic of Google balloons, see the Results section 3.2.2.

## 3. Results

### 3.1 Topic analysis

We have applied topic mining to both the English and the Finnish languages. Our first experiment was to construct a model using 2 million articles from the English Wikipedia (at the time this comprised 13% of the entire English Wikipedia).

The vocabulary is created by extracting eight features from the raw text divided into subdocuments. These features are based on part-of-speech (POS) classification (extracting and lemmatizing nouns, verbs, adjectives and adverbs) and named entity recognition (NER) (tagging words and groups of words as persons, locations, organizations and miscellaneous). For POS tag-

ging we use FreeLing (Padró et al., 2010) and for NER tagging the Illinois Named Entity Tagger (Ratinov and Roth, 2009). The documents and features are forwarded to the model trainer MPCA as a bag-of-words presentation. The MPCA produces the K (in these examples K=50) strongest topics for the user to name. This name is not used for the projection of text against the model (Equation 3), but it associates a numbered topic to a semantically meaningful context, which is essential for humans who exploit the method. Tables 2 and 3 present one of the fifty topics generated. The topic has intuitively been given the name "Space missions". Documents/texts under analysis are projected against the created model in order to find which topics the text is most strongly related to. Feature extraction and bag-of-words presentation are applied to the single document (as is done for the entire corpus in the model creation) before applying Equation 3 to the projection.

*Table 2:*
*The fourteen strongest Named Entity tags for the topic "Space missions". The un-normalized weighting factor corresponds to the incidence of the Named Entity in the particular topic. LOC stands for location, MISC for miscellaneous, ORG for organization, and PER for person. The weight is given at the left side of each word*

| Weight | LOC | Weight | MISC | Weight | ORG | Weight | PER |
|---|---|---|---|---|---|---|---|
| 14.25 | Russia | 34.11 | Russian | 5.71 | NASA | 1.51 | Venus |
| 6.31 | Moscow | 11.34 | Soviet | 5.27 | Sun | 0.77 | Ivan |
| 5.01 | Earth | 6.98 | Ukrainian | 1.84 | Apollo | 0.59 | Pluto |
| 4.85 | Ukraine | 2.09 | Estonian | 1.13 | Mars | 0.53 | Mars |
| 4.28 | Soviet Union | 1.98 | Georgian | 1.05 | Moon | 0.43 | Galileo |
| 1.78 | Kiev | 1.87 | Russians | 0.96 | Saturn | 0.42 | Mercury |
| 1.70 | Estonia | 1.56 | Latvian | 0.75 | NGC | 0.38 | Moon |
| 1.66 | Mars | 1.05 | Soyuz | 0.64 | Nikon | 0.38 | Vladimir |
| 1.56 | Jupiter | 1.03 | Belarusian | 0.61 | ISS | 0.35 | Ptolemy |
| 1.52 | USSR | 0.74 | Titan | 0.57 | GPS | 0.34 | Kepler |
| 1.35 | Georgia | 0.62 | Martian | 0.53 | ESA | 0.31 | Lenin |
| 1.30 | Belarus | 0.62 | Gregorian | 0.52 | Gemini | 0.30 | Boris |
| 1.31 | Latvia | 0.54 | Chechen | 0.50 | Canon | 0.28 | Koenig |
| 1.17 | Saint Petersburg | 0.50 | Earth | 0.44 | AU | 0.28 | Star |

Table 4 shows an example based on the BBC article entitled "Storm Sandy: Eastern US gets back on its feet" (31 October 2012). Table 4 presents the five strongest topics given by the model for this news article. The numbers in front of the topics are normalized statistical weights of each topic. Table 5 presents the Named Entities given by the NER tagger for this news article.

*Table 3: The strongest Part of Speech (POS) tags for the topic "Space missions". JJ stands for adjective, NN for noun, RB for adverb and VB for verb. The weight is given at the left side of each word*

| Weight | JJ | Weight | NN | Weight | RB | Weight | VB |
|---|---|---|---|---|---|---|---|
| 2.63 | solar | 2.06 | star | 15.16 | man | 2.85 | see |
| 1.90 | light | 1.89 | space | 2.50 | approximately | 1.93 | take |
| 1.71 | lunar | 1.37 | system | 2.34 | away | 1.52 | discover |
| 1.41 | html | 1.34 | planet | 1.98 | z_times | 1.43 | show |
| 1.40 | russian | 1.11 | object | 1.66 | close | 1.36 | give |
| 1.08 | red | 1.06 | camera | 1.56 | actually | 1.32 | move |
| 1.06 | astronomical | 0.96 | light | 1.47 | relatively | 1.30 | name |
| 1.05 | bright | 0.93 | satellite | 1.46 | slightly | 1.28 | find |
| 1.04 | scientific | 0.87 | crater | 1.44 | probably | 1.26 | appear |
| 1.04 | black | 0.86 | day | 1.22 | roughly | 1.13 | observe |
| 1.03 | dark | 0.84 | mission | 1.20 | sometimes | 1.11 | launch |
| 0.99 | similar | 0.81 | orbit | 1.19 | currently | 1.01 | call |
| 0.97 | visible | 0.80 | distance | 1.16 | long | 0.86 | refer |
| 0.90 | optical | 0.75 | lens | 1.16 | directly | 0.77 | base |

*Table 4: The strongest topics of the BBC, 31 October 2012 article "Storm Sandy: Eastern US gets back on its feet". The normalization is such that the total sum of the weights of all words in the material is 1*

| | Weight | Topic Name |
|---|---|---|
| 1 | 0.0512 | US politics |
| 2 | 0.0511 | Sci-fi and technology |
| 3 | 0.0391 | US traffic and information networks |
| 4 | 0.0384 | Latin America |
| 5 | 0.0342 | Physics |

*Table 5: The Named Entities for the BBC news article "Storm Sandy: Eastern US gets back on its feet" (31 October 2012)*

| Freq. | Type | Entity | Freq. | Type | Entity |
|---|---|---|---|---|---|
| 1 | MISC | Democratic | 1 | PER | Andrew Cuomo |
| 1 | MISC | Earth A | 1 | PER | Barack Obama |
| 1 | MISC | Jersey Shore | 2 | PER | Chris Christie |
| 1 | MISC | Nasdaq | 2 | PER | Christie |
| 3 | MISC | Republican | 1 | PER | Donna |
| 1 | LOC | Atlantic City | 1 | PER | Joseph Lhota |
| 1 | LOC | Canada | 1 | PER | Michael Bloomberg |
| 1 | LOC | Caribbean | 1 | PER | Mitt Romney |
| 1 | LOC | Easton | 1 | PER | Mt Washington |
| 1 | LOC | Haiti | 2 | PER | Obama |
| 1 | LOC | Hudson River | 1 | PER | Paul Adams |
| 1 | LOC | JFK | 1 | PER | Romney |
| 4 | LOC | Manhattan | 6 | PER | Sandy |
| 2 | LOC | Maryland | 1 | ORG | AP |
| 1 | LOC | NY City | 1 | ORG | CNN |
| 1 | LOC | New Hampshire | 1 | ORG | Coriolis Effect |
| 5 | LOC | New Jersey | 1 | ORG | Little Ferry |
| 7 | LOC | New York | 1 | ORG | MTA |
| 2 | LOC | New York City | 1 | ORG | Metropolitan Transit Authority |
| 1 | LOC | New York Stock Exchange | 1 | ORG | Moonachie |
| 1 | LOC | New York University | 1 | ORG | National Weather Service |
| 1 | LOC | Ohio | 1 | ORG | New York Stock Exchange |
| 1 | LOC | Queens | 1 | ORG | Newark Liberty |
| 1 | LOC | Teterboro | 1 | ORG | Tisch Hospital |
| 4 | LOC | US | 1 | ORG | Trams |
| 1 | LOC | Washington DC | 1 | ORG | US Department of Energy |

Tables 6 and 7 explore the fifth strongest topic of this news article, "physics", showing a collection of the strongest individual Wikipedia articles on this topic, and strongest features of this topic.

The Finnish language Wikipedia turned out to be far less extensive than the English one. Instead, we used a collection of 73 000 news articles from the Finnish News Agency (STT). Generally, the text in this material is of good quality, but there are some limitations: sports news are dominating and there are very few information technology related news (no Apple, Google, Facebook, Twitter, etc.). The STT news used here date from the years 2002-2005 including also 5000 news from February 2013. For POS tagging the STT news we used a commercial morphological parser, FINTWOL by Ling-Soft Ltd., and for NER tagging we created lists of NER tagged words to which we compared single and groups of POS tagged and lemmatized words. As an example for Finnish, Tables 8 and 9 present results based on an article about the re-election of Giorgio Napolitano as the president of Italy (*Talouselämä*, 22 April 2013).

*Table 6: The strongest individual Wikipedia articles for the topic "Physics"*

| |
|---|
| Terahertz time-domain spectroscopy |
| List of materials analysis methods |
| Fiber laser |
| Cryogenic particle detectors |
| Varistor |
| Neutron generator |
| Laser ultrasonics |
| Optical amplifier |
| Thyristor |
| Electric current |
| Neutron source |
| Voltage-regulator tube |
| Switched-mode power supply |
| Gas-filled tube |
| Isotopes of plutonium |
| Superconducting magnet |

*Table 7: The strongest features for the topic "Physics"*

| | |
|---|---|
| NE-LOC | US, Europe, Chernobyl, Hiroshima, Earth. |
| NE-MISC | X-ray, Doppler, CO2, °C, CMOS, Fresnel. |
| NE-ORG | CERN, IPCC, IAEA}. |
| NE-PER | Maxwell, Edison, Gibbs, Watt, Richter, Einstein, Rutherford, Faraday, Bohr}. |
| POS-JJ | nuclear, electrical, magnetic, liquid, thermal, atomic, mechanical, solid}. |
| POS-NN: | energy, power, system, gas, material, temperature, pressure, air, effect, frequency, wave, field, heat, particle, unit, process, signal, mass, device, surface, circuit, light. |
| POS-RB: | relatively, extremely, slowly, fast. |
| POS-VB: | produce, require, cause, measure, reduce, increase, generate, allow, apply, create. |

*Table 8: The five strongest topics for a Talouselämä, 22 April 2013 article (English translation in parenthesis)*

| Number | Weight | Topic Name |
|---|---|---|
| 1 | 0.1014 | Vaalit (elections) |
| 2 | 0.0567 | Kansainvälinen konflikti (international conflict) |
| 3 | 0.0497 | Sää (weather) |
| 4 | 0.0438 | Aseellinen selkkaus (armed conflict) |
| 5 | 0.0434 | Tuloneuvottelut (income negotiations) |

*Table 9: The Named Entities for the Talouselämä, 22 April 2013 article*

| Freq. | Type | Value |
|---|---|---|
| 2 | MISC | presidentti (president) |
| 1 | MISC | radikaali (radical) |
| 2 | LOC | Italia (Italy) |
| 1 | LOC | maa (country) |
| 2 | PER | Napolitano |
| 1 | ORG | hallitus (government) |
| 2 | ORG | parlamentti (parliament) |

Figure 2 shows all the fifty topics obtained for the *Talouselämä*, 22 April 2013 article. The highest peak is the strongest topic "president". The number of Named Entities for the example in Finnish is much smaller than that in English. The state of the art NER taggers for Finnish are not as evolved as the taggers for English.

The overall results are, in fact, better for the BBC article; there are more NER tagged words and the strongest topics correspond better to the semantic contents of the article.

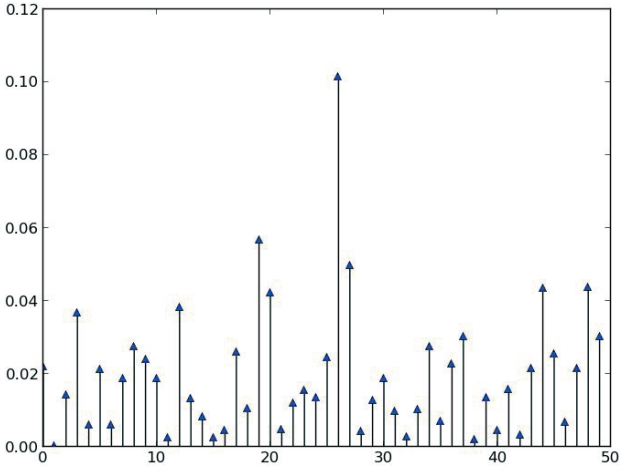



*Figure 2: The fifty strongest topics for one news article projected against model created with STT news data. The horizontal axis shows the number of the topic and the vertical axis shows normalized weight of the topic*

However, the results using the model created with STT data are far better than those created with the Finnish Wikipedia. This is demonstrated in Figure 3 where the strongest topics do not as strongly rise above the rest and, furthermore, the five strongest topics are mostly not significant: Finnish politics, philosophy and religion, natural sciences, computer games, and banks and monetary policies. This shows that the corpus and named entity data used to create the model is sufficiently extensive and of good quality.

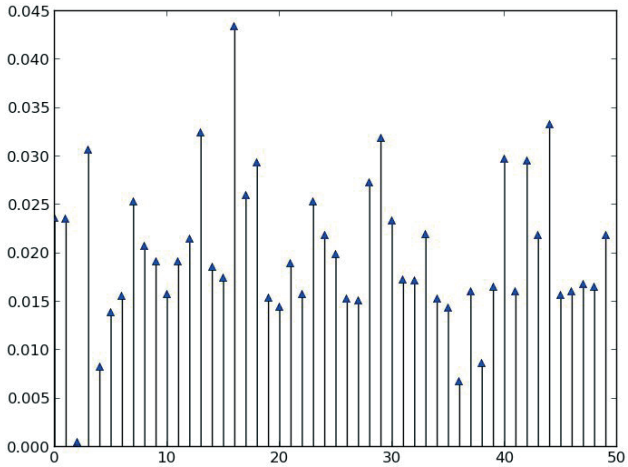



*Figure 3: The fifty strongest topics for one news article projected against a model created using the Finnish Wikipedia*

As another source for the corpus in Finnish we used the free newspaper sheet *Metro*. For POS tagging of *Metro* news we used the Open Source Morphology for Finnish, OmorFi (Lindén et al., 2011), and for NER tagging we used a combination of OMorFi and our own POS tagging version created for STT news.

## 3.2 Word association analysis

### *3.2.1 Diverging word clouds*

In this section we present some results of the Software Newsroom applications that use word association analysis as their basis. As before, for single document experiments we use the English Wikipedia as the background corpus and a story from BBC, "Google tests balloons to beam internet from near space" (Kelion, 2013), as the foreground document that we are interested in.

Given a document *d* and a word *w*, a specification for the corresponding diverging association cloud is directly obtained from the foreground associations of the document: take the top *n* words associated with *w* in the foreground and position them in the word cloud according to their weights. Figure 4 illustrates the idea using the document on Google balloons.


*Figure 4: The diverging word cloud created from the foreground associations of the news story "Google tests balloons to beam internet from near space". The search term for the left diverging cloud is 'google' and for the right diverging word cloud the search term is 'balloons'. We used an internet tool (Word Clouds for Kids, 2012) for generating the word clouds*

These association clouds give a good idea of what the document could be about. Such word clouds could also be implemented in an interactive manner: as the user clicks on a word in the cloud, the selected word becomes the next search term and, correspondingly, all divergent clouds are re-rendered for all documents using the new focus. A drawback of this method is that it takes time to get used to the fact that the word cloud is conditioned on the search term and thus interpreting the results could be non-intuitive for novice users. In order to alleviate the problem, it would be possible to also present the search term together with the words but currently we do not have a clear idea of how to present this in an intuitive manner.

### *3.2.2 Summary generation*

For our experiments with the summary generation algorithm presented earlier, we collected four news stories on the same topic from different news sources - BBC: "Up, up and away: Google to launch Wi-Fi balloon experiment" (Kelion, 2013), National Geographic: "Goog-

le's Loon Project Puts Balloon Technology in Spotlight" (Handwerk, 2013), ARS Technica: "Google's balloon-based wireless networks may not be a crazy idea" (Brodkin, 2013), CNN: "Google tests balloons to beam internet from near space" (Smith-Spark, 2013).

For the background associations we used Wikipedia news stories and the foreground associations were calculated using the respective news stories. We then applied the algorithm which we described on this set of data. The total number of words in all the four documents was 3696.

The first four sentences, containing a total of 120 words, returned by the algorithm were the following:

- Google is reportedly developing wireless networks for sub-Saharan Africa and Southeast Asia that would combine a technology well established for such purposes (TV White Spaces) with one that's a bit more exotic - balloons that transmit wireless signals. - *ARS Technica*
- Project Loon balloons are made of plastic just 3 mm (0.1in) thick, another Orlando-based firm, World Surveillance Group, sells similar equipment to the US Army and other government agencies. - *BBC*
- It has been working on improving connectivity in the US with Google Fiber and bringing the internet to underserved populations overseas through White Spaces networks. - *ARS Technica*
- A company called Space Data makes balloon-based repeater platforms for the US Air Force that "extend the range of standard-issue military two-way radios from 10 miles to over 400 miles." - *ARS Technica*

The application of the algorithm yields promising results. Our next goals are improving and evaluating the current method. It is important to note here that the way the extracted sentences are presented to the user is also a very important aspect. For instance, consider the third sentence which has a co-reference resolution problem (i.e., the sentence starts with "it" and we do not know what "it" is). In such cases it makes sense to present consecutive sentences together in the summary regardless how they are ordered by the algorithm. In some cases this could help to overcome the co-reference resolution problem. It is also possible to provide some context to the user, for instance, when the user's cursor hovers over an extracted sentence, the sentences which are before and after it in the news story can be shown.

## 4. Discussion

Application of MPCA seems to work well for news search by topic analysis. It is likely that also other variants of probabilistic modeling perform well for news identification. Our second approach, association analysis, also clearly enhances the effectiveness of the "news nose". A question then arises, whether other methods could be effective as well, or even better than the adopted approaches.

In contrast to statistical methods such as PCA, *cluster analysis* can best be seen as a heuristic method for exploring the diversity in a data set by means of pattern generation (van Ooyen, 2001). Cluster analysis may be applied for finding similarities and trends in data (described using the common term *pattern recognition*). An example of cluster analysis is the *expectation maximization* (EM) algorithm, which has recently been applied to astronomical data for identifying stellar clusters from large collections of infrared survey data (Solin, Ukkonen and Haikala, 2012). Cluster analysis has also been used in, e.g., market research within a more general family of methodologies called *segmentation methods*. These can be used to identify groups with common attitudes, media habits, lifestyle, etc. Cluster analysis is probably less well suited for news search than probabilistic models like MPCA, since the semantic contents of articles that contain more than one topic are not resolved by cluster analysis (e.g., Newman at al., 2006), while probabilistic modeling clearly performs well in such cases provided that the corpus and named entity data used for the model creation are sufficiently extensive. This will result in only a small number of unrecognized words that cannot be tagged, and thus a high resolving power of topics and named entities.

*Supervised learning* methods divide objects such as text documents into predefined classes (Yang, 1999). Cluster analysis and PCA are data driven methods which can extract information from documents without *a priori* knowledge of what the documents may contain (Newman et al., 2006), and topic categorization (i.e., a topic model) is created by the algorithm without rules or restrictions on the contents of a topic, which is why such methods are called *unsupervised learning*. Obviously supervised learning is poorly suited for news search from arbitrary textual data, since the topics of potential news in the material cannot be predicted, and it is thus impossible to recognize new emerging topics.

A further, more advanced analysis of complex data may incorporate the use of *semantic networks*. Methods of this category are *Traditional* and *Improved Three-Phase Dependency Analysis* (TPDA, ITPDA). These algorithms have been applied to recognition of semantic information in visual content and they use Bayesian networks to automatically discover the relationship networks among the concepts. These methods can be applied, for example, to automatic video annotation. (Wang, Xu and Liu, 2009).

In this paper, we have mainly interpreted the associations on a single association level rather than as a network. But these associations, both background and foreground, can also be seen as a kind of semantic network where words are nodes and the edges represent the associations. Analyzing the background associations as a network might give interesting results in automatic word domain discovery or for finding interesting subnetworks that connect two words. The same applies for the foreground associations, which might provide interesting inference and application possibilities when interpreted as word networks and used as such. Thus, in the future, our models and methods could be improved in their accuracy. More efficient, scalable algorithms could be designed and, perhaps more interestingly, additional novel applications could be invented with help of the background and foreground models, especially in the broad areas of information browsing and retrieval.

Considering the topic model and data used for the training, our experiments indicate that the comprehensiveness, quality, and also the semantic similarity of the text corpus and named entity data with the data under analysis are critical to the effectiveness of the search algorithm. This is of course obvious, but poses a challenge for automated news search since language evolves and the language used in, e.g., social media that obeys no standard rules diffuses with an increasing speed to various media channels. Should we accept this and modify the models and additionally also adopt slang in the presentation of news, or try to force the users to educate themselves in order to write in decent standard language also in social media?

An aspect of crucial importance in (automated) news search is the quality of the data. The internet is full of hoaxes and distorted information, and finding assurance for the reliability of potential news may sometimes be challenging, and will require too much time.

This may lead to that the potential news becomes yesterday's news or that it is published by a competitor before sufficient background information is found. The Software Newsroom should therefore trace all possible metadata on the sources, time, places, people, and organizations associated with the creation of the information found by automated means. While this cannot rely merely on software, automation can be used to significantly improve the effectiveness and speed of the process.

## 5. Conclusions

We have developed and applied methods for automated identification of potential news from textual data for use in an automated news search system called Software Newsroom. The purpose of the tools is to analyze data collected from the internet and to identify information that has a high probability of containing news. The identified potential news information is summarized in order to help understanding the semantic contents of the data and also to help in the news editing process.

Two different methodological approaches have been applied to the news search. One method is based on topic analysis which uses MPCA for topic model creation and data profiling. The second method is based on association analysis that applies LLR. The two methods are used in parallel to enhance the news recognition capability of Software Newsroom.

For the topic mining we have created English and Finnish language corpora from Wikipedia and several Finnish language corpora from Finnish news archives, and we have used bag-of-words presentations of these corpora as training data for the topic model. We have made experiments of topic analysis using both the training data itself and arbitrary text parsed from internet sources. The selected algorithmic approach is found to be well suited for the task, but the effectiveness and success of news search depends strongly on the extensiveness and quality of the training data used for the creation of the topic model. Also, semantic similarity of the target text with the corpus used for the model creation generally improves the search effectiveness. The large difference between the language commonly used in user-created internet content and standard language poses a challenge for news search from social networks, since a significant part of the language is not recognized by the part-of-speech and name entity taggers. A simple solution for this would be a translator that would preprocess the unknown slang words, turning them into standard language. Another would be a slang-based corpus. The latter has the disadvantage that the resulting raw news material would be composed of slang and it would have to be translated into standard language before publishing. Thus, our plan is to collect a small dictionary of the most common words used in social media and use them for further experiments on social media.

In the association analysis we have used a methodology for detecting novel word associations from a set of documents. For detecting novel associations we first used the background corpus from which we extracted such word associations that are common. We then collected the statistics of word co-occurrences from the set of documents that we are interested in, looking for such associations which are more likely to appear in these documents than in the background.

We also demonstrated applications of Software Newsroom based on association analysis - association visualization as a graph, diverging (association) clouds which

are word clouds conditioned on a search term, and a simple algorithm for text summarization by sentence extraction. We believe that the background-foreground model has significant potential in news search. The simplicity of the model makes it easy to implement and use. At the same time, our experiments indicate great promise in employing the background-foreground word associations for different applications.

The combination of the two methods has not yet been implemented. This is in our plans for the near future. and the application of both methods on social media using a pre-translator of social media language is underway. Potential future work also includes experiments on automated news generation and application of our methods for other purposes, e.g., improvement of recommendations.

**Acknowledgements**

We are grateful to Jukka Mauno from MTV3 for raising the idea of Software Newsroom. We acknowledge Tekes, the Finnish Funding Agency for Technology and Innovation, and Tivit Oy for financing this research. Especially we wish to thank the Focus Area Director of the Next Media research programme, Eskoensio Pipatti, for his support. This work has also been supported by the Algorithmic Data Analysis (Algodan) Centre of Excellence of the Academy of Finland and Hecse, Helsinki Doctoral Programme in Computer Science - Advanced Computing and Intelligent Systems. Finally, we thank the referees for valuable comments.

**References**

Agrawal, R., Imieliński, T. and Swami, A., 1993. Mining association rules between sets of items in large databases. *ACM SIGMOD Record*. 22(2), pp. 207-216

Allan, J., 2002. *Topic Detection and Tracking: Event-based Information Organization*. Dordrecht: Kluwer Academic Publishers

Berry, M.W., Dumais, S.T. and O'Brien, G.W., 1994. Using Linear Algebra for Intelligent Information Retrieval. *SIAM Review,* 37(4), pp. 573-595

Blei, D.M., Ng, A.Y. and Jordan, M.I., 2003. Latent Dirichlet Allocation. *The Journal of Machine Learning Research*, 3, pp. 993-1022

Brodkin J. (2013). Google's balloon-based wireless networks may not be a crazy idea. *ARS Technica*, June 2, 2013. [Online] Available at: <http://arstechnica.com/information-technology/2013/06/googles-balloon-based-wireless-networks-may-not-be-a-crazy-idea/>. [Accessed 26 June 2013]

Buntine, W. and Jakulin, A., 2004. Applying discrete PCA in data analysis. *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence* (UAI2004), pp. 59-66

Buntine, W. and Jakulin, A., 2006. Discrete component analysis. *Subspace, Latent Structure and Feature Selection, Lecture Notes in Computer Science*. Vol. 3940, pp. 1-33

Church, K.W. and Hanks, P., 1990. Word association norms, mutual information, and lexicography. *Computational linguistics*, 16(1), pp. 22-29

Comintelli, 2013. [Online] Available at: <http://www.comintelli.com/Company/Press-Releases/Esmerk-launches-new-current-awareness-platform-Esm> [Accessed 31 October 2013]

Conroy, J.M. and O'leary, D.P. 2001. Text summarization via hidden markov models. *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval,* pp. 406-407

Dumais, S. T., Furnas, G.W., Landauer, T.K., Deerwester, S. and Harshman, R., 1988. Using latent semantic analysis to improve access to textual information. *Proceedings of the SIGCHI conference on Human factors in computing systems,* pp. 281-285

Dunning, T., 1993. Accurate methods for the statistics of surprise and coincidence. *Computational linguistics*, 19(1), pp. 61-74

Hakulinen, A., Vilkuna, M., Korhonen, R., Koivisto, V., Heinonen, T. R., and Alho, I., 2004. *Iso suomen kielioppi*. Helsinki: Suomalaisen Kirjallisuuden Seura (Available online at http://scripta.kotus.fi/visk/etusivu.php)

Handwerk, B. (2013). Google's Loon Project Puts Balloon Technology in Spotlight. *National Geographic*, June 18, 2013. [Online] Available at: <http://news.nationalgeographic.com/news/2013/06/130618-google-balloon-wireless-communication-internet-hap-satellite-stratosphere-loon-project/>. [Accessed 26 June 2013]

Harris, Z., 1954. Distributional Structure. *Word*, 10(23), pp 146-162

Hofmann, T., 1999. Probabilistic Latent Semantic Indexing, *Proc. 22nd Annual International SGIR Conference on Research and Development in Information Retrieval*, pp. 50-57

Hori, C. and Furui, S., 2003. A new approach to automatic speech summarization. *IEEE Transactions on Multimedia*, 5(3), pp. 368-378

Huang, S., Peng, X., Niu, Z. and Wang, K., 2011. News topic detection based on hierarchical clustering and named entity. *7th International Conference on Natural Language Processing and Knowledge Engineering.* pp. 280-284

Hynönen, T., Mahler, S. and Toivonen, H., 2012. Discovery of novel term associations in a document collection. *Bisociative Knowledge Discovery, Lecture Notes in Computer Science,* Vol. 7250, pp. 91-103

Karlsson, F., 1990. Constraint grammar as a framework for parsing running text. *Proceedings of the 13th Conference on Computational Linguistics*, Vol. 3., pp. 168-173

Kelion L. (2013). Google tests balloons to beam internet from near space. *BBC*, June 15, 2013. [Online] Available: <http://www.bbc.co.uk/news/technology-22905199>. [Accessed 26 June 2013]

Kimura, M., Saito, K. and Uera, N., 2005. Multinomial PCA for extracting major latent topics from document streams, *Proc. 2005 IEEE International Joint Conference on Neural Networks*, Vol. 1, pp. 238-243

Knight, K. and Marcu, D., 2002. Summarization beyond sentence extraction: A probabilistic approach to sentence compression. *Artificial Intelligence*, 139(1), pp. 91-107

Lindén, K., Axelson, E., Hardwick, S., Pirinen, T.A. and Silfverberg, M., 2011. HFST-Framework for Compiling and Applying Morphologies. In: Mahlow, C. and Piotrowski, M., eds.(2011). *Systems and Frameworks for Computational Morphology. Communications in Computer and Information Science*, Vol. 100. Berlin-Heidelberg: Springer. pp. 67-85

Lindén, K. and Pirinen, T., 2009. Weighted finite-state morphological analysis of Finnish compounds. In: Jokinen, K. and Bick, E., eds.(2009). *Proc. Nordic Conference of Computational Linguistics*. Odense: NEALT

McDonald, D. D., 1996. Internal and external evidence in the identification and semantic categorization of proper names. In: Boguraev, B. and Pustejovsky, J., eds. (1996). *Corpus Processing for Lexical Acquisition.* Cambridge, MA: MIT Press. pp. 21-39

Meltwater, 2013. [Online] Available at: <http://www.meltwater.com/products/meltwater-buzz-social-media-marketing-software/> [Accessed 31 October 2013]

Miller G.A., 1995. WordNet: A Lexical Database for English, *Communications of the ACM,* 38(11), pp. 39-41

Nadeau, D. and Sekine, S., 2007. A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30(1), pp. 3-26

Newman, D., Chemudugunta, C., Smyth, P. and Steyvers, M., 2006. Analyzing Entities and Topics in News Articles using Statistical Topic Models. *Lecture Notes in Computer Science*, Volume 3975, pp. 93-104

van Ooyen, A., 2001. Theoretical aspects of pattern analysis. In: L. Dijkshoorn, K. J. Tower, and M. Struelens, eds. *New Approaches for the Generation and Analysis of Microbial Fingerprints.* Amsterdam: Elsevier, pp. 31-45

Padró, L., Reese, S., Agirre, E. and Soroa, A., 2010. Semantic Services in FreeLing 2.1: WordNet and UKB. *Proceedings of the Global Wordnet Conference 2010*

Pearson, K., 1901. On Lines and Planes of Closest Fit to Systems of Points in Space. *Philosophical Magazine*, 2(11), pp. 559-572

Ratinov, L. and Roth, D., 2009. Design Challenges and Misconceptions in Named Entity Recognition. *Proceedings of the Thirteenth Conference on Computational Natural Language Learning*, pp. 147-155

Salton, G., 1991. Developments in Automatic Text Retrieval. *Science*, Vol 253, pp. 974-979

Shen, D., Sun, J.T., Li, H., Yang, Q. and Chen, Z., 2007. Document summarization using conditional random fields. *Proceedings of the 20th international joint conference on Artifical intelligence,* Vol. 7, pp. 2862-2867

Smith-Spark L. (2013). Up, up and away: Google to launch Wi-Fi balloon experiment. *CNN*, June 15, 2013. [Online] Available at: <http://www.bbc.co.uk/news/technology-22905199>. [Accessed 26 June 2013]

Solin, O., Ukkonen, E., and Haikala, L., 2012. Mining the UKIDSS Galactic Plane Survey: star formation and embedded clusters, *Astronomy & Astrophysics*, Volume 542, A3, 23 p

Toivonen H., Gross, O., Toivanen J.M. and Valitutti A., 2012. Lexical Creativity from Word Associations. *Synergies of Soft Computing and Statistics for Intelligent Data Analysis. Advances in Intelligent Systems and Computing.* Vol. 190*,* pp. 17-24

Wang, F., Xu, D. and Liu, J., 2009. Constructing semantic network based on Bayesian Network. *1st IEEE Symposium on Web Society*, pp. 51-54

Word Clouds for Kids, 2013. [Online] Available at: <http://www.abcya.com/word_clouds.htm>. [Accessed 26 June 2013]

Yang, Y., 1999. An Evaluation of Statistical Approaches to Text Categorization. *Information Retrieval*, Vol 1, pp. 67-88

Yeh, J.Y., Ke, H.R., Yang, W.P. and Meng, I., 2005. Text summarization using a trainable summarizer and latent semantic analysis. *Information Processing & Management*, 41(1), pp. 75-95

# Paper III

Oskar Gross, Antoine Doucet, and Hannu Toivonen

**Document Summarization Based on Word Associations**

In *Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval*, 1023-1026, ACM, 2014.



III

# Document Summarization Based on Word Associations

Oskar Gross
Department of Computer Science and HIIT
University of Helsinki, Finland
oskar.gross@cs.helsinki.fi

Antoine Doucet
GREYC, CNRS UMR 6072
University of Normandy, Unicaen, France
antoine.doucet@unicaen.fr

Hannu Toivonen
Department of Computer Science and HIIT
University of Helsinki, Finland
hannu.toivonen@cs.helsinki.fi

## ABSTRACT

In the age of big data, automatic methods for creating summaries of documents become increasingly important. In this paper we propose a novel, unsupervised method for (multi-)document summarization. In an unsupervised and language-independent fashion, this approach relies on the strength of word associations in the set of documents to be summarized. The summaries are generated by picking sentences which cover the most specific word associations of the document(s). We measure the performance on the DUC 2007 dataset. Our experiments indicate that the proposed method is the best-performing unsupervised summarization method in the state-of-the-art that makes no use of human-curated knowledge bases.

## Categories and Subject Descriptors

I.2.7 [**Natural Language Processing**]: Language models—*abstracting methods, summarization*

## General Terms

Algorithms, Experimentation, Languages

## Keywords

Multi-Document Summarization, Word Associations

## 1. INTRODUCTION

We propose a novel method for document summarization, Association Mixture Text Summarization, aimed to abstract a news story into a shorter text. Like most other methods, Association Mixture Text Summarization works in a sentence-based manner, selecting a set of sentences from the document to be summarized to constitute its summary. The sentences are chosen so that they collectively cover as much of the relevant information in the original document as possible. The main difficulties are to define what is relevant and to measure how well sets of sentences cover relevant information. Our method has three central characteristics:


*SIGIR '14,* July 06 - 11 2014, Gold Coast , QLD, Australia

http://dx.doi.org/10.1145/2600428.2609500

(1) *Relevance is based on the relative associations between words, helping to grasp the most salient information in a news story.* Much of the core content of news stories is in the links they establish, e.g., between people, acts, events, and places. We argue that associations at subtler levels can also be important, even ones between adjectives or adverbs and noun or verbs used in the news. Recognition of associations is based on statistical analysis of word co-occurrences within sentences. We believe that such associations reflect the key ideas of news and are useful for selecting sentences.

(2) *Novel associations in a document are recognized by contrasting them against a background corpus.* News stories are supposed to tell something new and a key problem in summarization is to identify what is new in a given document. We treat this as a novelty detection task by contrasting the document to a background corpus to see which associations are emphasized more in the document.

(3) *Natural language processing is trivial, making the method language-independent.* All processed documents are split to sentences and tokens (words) based on punctuation and whitespaces; numbers are removed, and the remaining tokens are used as they are, without any further processing.

In this paper we focus on the sentence selection subtask of document summarization. We do not address the issue of arranging or processing the sentences for improved readability. We evaluate the method in English using public benchmarks, and leave experiments with other languages for future work. In the experiments, our proposed method outperforms all unsupervised summarization methods that do not use semantic resources such as Wordnet.

This paper is organised as follows. We next briefly review related work. We then present the Association Mixture Text Summarization method in Section 3. The performance of the method is evaluated in Section 4, while Section 5 concludes this article with a discussion.

## 2. RELATED WORK

Document summarization is a well-studied area. There are two types of summarizations methods: methods which select existing sentences and methods which generate sentences. Both of these types of methods can be either supervised or unsupervised, i.e., either learning from examples of existing summaries or not. We focus on the unsupervised domain, of which we give a very brief overview. Nenkova and McKeown [10] provide an exhaustive review of the topic.

Some methods use Latent Semantic Analysis (LSA) [2] as their basis (e.g. [4]). The state-of-the art in purely unsupervised summarization is represented by the DSDR method of

He et al. [5]. This approach generates a summary by using sentences that best "reconstruct" the original document. This work has been extended by Zhang et al. [11] who combined document reconstruction and topic decomposition.

A number of unsupervised methods take advantage of additional linguistic resources. In particular, the Two-Tiered Topic model by Celikyilmaz [1] uses Wordnet [9] and the DUC-provided user query for selecting the summary sentences. The Document Understanding Conference[1] (DUC) provides most evaluation procedures and collections in the summarization field. We provide further details in Section 4.

# 3. METHOD

The *Association Mixture Text Summarization* method proposed below takes as its input a *document* $D$ to be summarized and a *background corpus* $\mathcal{B}$ consisting of a set of documents representing the norm or the current state of information.

As a special case, the background corpus can be empty. Additionally, by extension, instead of a single document a set of documents can be summarized by simply giving their concatenation as the input document $D$, as will be done in the experimental section.

The method has two parts: (1) computation of document-specific word associations, and (2) selection of sentences with strong word associations. These two steps are described in the following subsections.

## 3.1 Finding Document-Specific Associations

We consider two relevance criteria for associations in the given document $D$.

First, an association between two words is more relevant if they are statistically mutually dependent, i.e., if they co-occur in $D$ more frequently than they would by chance. This, of course, is a classic idea.

Second, and more interestingly, the association is characteristic for document $D$ if the two words co-occur in $D$ more frequently than in the background corpus $\mathcal{B}$.

The second criterion is in principle more useful since it uses additional data to assess the association, but it is of little value if the background corpus is small or if the words or the word pair does not occur in the corpus. Our method therefore uses a mixture model of the two criteria above.

**Notation.** We first define the notation for various counts of words and word pairs in document $D$ and in background $\mathcal{B}$. Let $t_i$ and $t_j$ be words. We use $n_{ij}$ to denote the number of sentences in document $D$ that contain both words $t_i$ and $t_j$, $n_{i-j}$ the number of sentences containing word $t_i$ but not $t_j$, $n_{-ij}$ the number of sentences containing $t_j$ but not $t_i$, and $n_{-i-j}$ the number of sentences containing neither $t_i$ nor $t_j$. We use $n_{i\cdot} = n_{ij} + n_{i-j}$ to denote the total number of sentences containg word $t_i$, and respectively for $n_{\cdot j}$. Let $n = |D|$ denote the total number of sentences in document $D$. Finally, let $m_{ij}$, $m_{i-j}$, $m_{-ij}$, $m_{-i-j}$, $m_{i\cdot}$, $m_{\cdot j}$ and $m$ be the respective counts in the background corpus $\mathcal{B}$.

**Statistical Model.** Consider the association between words $t_i$ and $t_j$. We use multinomial distributions to model the probabilities of observing different combinations of existence/non-existence of words $t_i$ and $t_j$ in a sentence. The four respective model parameters are $p_{ij}$, $p_{i-j}$, $p_{-ij}$ and $p_{-i-j}$, affecting the likelihood of the observed counts $n_{ij}$, $n_{i-j}$, $n_{-ij}$ and $n_{-i-j}$. Three such models are given next, and the fit of the data to these models is later used to assign a weight to the association between $t_i$ and $t_j$. The third model is the Association Mixture model, while the first two are simpler models that will be used as the components of the mixture.

For convenience, we below define the models using parameters $p_{i\cdot}$ (the probability of observing word $t_i$), $p_{\cdot j}$ (the probability of observing word $t_j$), and $p_{ij}$ (the probability of observing both $t_i$ and $t_j$). These give more natural definitions for the models. The multinomial model parameters can then easily be obtained as $p_{i-j} = p_{i\cdot} - p_{ij}$; $p_{-ij} = p_{\cdot j} - p_{ij}$; $p_{-i-j} = 1 - p_{ij} - p_{i-j} - p_{-ij}$.

The independence model (component) $p^{\text{D-ind}}$ considers observed frequencies of words $t_1$ and $t_2$ only in document $D$ and assumes that they are statistically independent:

$$p^{\text{D-ind}}_{i\cdot} = n_{i\cdot}/n; \quad p^{\text{D-ind}}_{\cdot j} = n_{\cdot j}/n; \quad p^{\text{D-ind}}_{ij} = n_{i\cdot} \cdot n_{\cdot j}/n^2.$$

If the data fits this model badly, i.e., essentially if $n_{ij}$ deviates a lot from $n_{i\cdot} \cdot n_{\cdot j}/n$, then the words are likely to be statistically dependent.

The background model (component) $p^{\text{B}}$ estimates all three parameters from the respective relative frequencies in the background corpus $\mathcal{B}$:

$$p^{\text{B}}_{i\cdot} = m_{i\cdot}/m; \quad p^{\text{B}}_{\cdot j} = m_{\cdot j}/m; \quad p^{\text{B}}_{ij} = m_{ij}/m.$$

If the data fits this model badly then the word pair occurs in the document differently from the background. This signals that the association is novel.

The association mixture model $p^{\text{B+D-ind}}$ averages the two components above, weighted by their sample sizes $n$ and $m$: $p^{\text{B+D-ind}} = (n \cdot p^{\text{D-ind}} + m \cdot p^{\text{B}})/(n+m)$. This gives

$$\begin{aligned} p^{\text{B+D-ind}}_{i\cdot} &= (n_{i\cdot} + m_{i\cdot})/(n+m), \\ p^{\text{B+D-ind}}_{\cdot j} &= (n_{\cdot j} + m_{\cdot j})/(n+m), \\ p^{\text{B+D-ind}}_{ij} &= (n_{i\cdot} \cdot n_{\cdot j}/n + m_{ij})/(n+m). \end{aligned}$$

In other words, the mixture model combines information from document $D$ itself and from the background $\mathcal{B}$. Their relative weights adapt to their relative sizes, giving more emphasis to the statistically more reliable source of information.

**Association Weights.** The weight of the association between two words is based on a log-likelihood ratio test [3]. The test compares two models for each word pair: (1) a null model, in our case the mixture model, and (2) a maximum likelihood alternative model. If the likelihood of the alternative model is much higher, then the null model is less likely to be true. In other words, the mixture model is an expression of expectations, and we are actually interested in finding exceptions to them.

The maximum likelihood model $p^{\text{D}}$ is obtained by simply assigning the model parameters directly from the observed relative frequencies: $p^{\text{D}}_{i\cdot} = n_{i\cdot}/n$; $p^{\text{D}}_{\cdot j} = n_{\cdot j}/n$; $p^{\text{D}}_{ij} = n_{ij}/n$.

Let $L(p^{\text{D}})$ be the likelihood of the maximum likelihood model given the counts $n_{ij}$, $n_{i-j}$, $n_{-ij}$, $n_{-i-j}$ in document $D$, and let $L(p^{\text{B+D-ind}})$ be the likelihood of the mixture model given the same counts. We define the weight $w(t_i, t_j)$ of the association between $t_i$ and $t_j$ as the value of the respective log-likelihood ratio test:

$$w(t_i, t_j) = -2 \log \frac{L(p^{\text{B+D-ind}})}{L(p^{\text{D}})}.$$

[1] http://duc.nist.gov/

Multinomial coefficients in the likelihoods cancel out, and after simplification we have

$$w(t_i, t_j) = 2 \sum_{\substack{a \in \{\text{“}ij\text{”}, \text{“}i-j\text{”}, \\ \text{“}-ij\text{”}, \text{“}-i-j\text{”}\}}} n_a (\log p_a^{\text{D}} - \log p_a^{\text{B+D-ind}}).$$

The log-likelihood ratio test gives lower weights for word pairs that better match the mixture model and higher weights for those associations that are unexpected with respect to the mixture model. In text summarization, we are interested in word pairs that have a higher relative frequency in the document $D$ than in the background $\mathcal{B}$, and that have a high log-likelihood ratio.

## 3.2 Sentence Selection

The other subtask is to select from document $D$ sentences that contain strong word associations. In the sentence selection phase, our goal is to preserve as many of the stronger associations and thereby as much as possible of the core contents of the original document $D$.

Given a fixed target size of the summary (e.g. 250 words) and the association weights, we aim to pick sentences such that the sum of the log-likelihood ratios of word pairs in the summary is maximized. To avoid selecting sentences with too similar content, each pair is taken into account once.

Formally, let document $D$ be a set of sentences and let each sentence be a set of words. We call any subset $S = \{s'_1, \ldots, s'_s\} \subset D$ of sentences a *summary* of $D$. We define the total weight of associations in summary $S$ as

$$w(S) = \sum_{\substack{\{t_i, t_j\} \text{ s.t. } t_i \neq t_j \wedge \\ \exists s \in S: \{t_i, t_j\} \subset s}} w(t_i, t_j),$$

i.e., as a sum over the set of word pairs in any sentence of the summary. Every pair is only counted once.

In the sentence selection step we aim to find a summary $S^* \subset D$ with a maximal total weight, i.e.,

$$S^* = \arg \max_{\substack{S \subset D \\ ||S|| \leq L}} w(S),$$

where $||S||$ is the number of words in summary $S$. In our experiments below, the upper limit is set to $L = 250$ words.

This problem is similar to the weighted set cover problem [6]: use sentences of the document to cover as much of the associations as possible. Due to the limited length of the summary, a natural “cost” of a sentence is the number of words in it. Given the computational complexity of the task, we resort to a greedy algorithm [6] to find a summary $S$ that approximates the optimum $S^*$.

For the sake of simplicity, in the experiments below we add sentences to the summary $S$ until the maximum size is reached ($||S|| \geq L$) and then simply truncate the summary to $L$ words.

# 4. EXPERIMENTS

In this section, we describe experiments carried out to evaluate the proposed Association Mixture Text Summarization method. We aim to address the following questions: (1) How does the method perform in comparison to state-of-the-art unsupervised summarization methods? (2) What are the contributions of the components $p^{\text{B}}$ and $p^{\text{D-ind}}$ to the method? (3) What is the effect of the size of the background corpus $\mathcal{B}$ on the quality of the summaries?

## 4.1 Experimental Setup

For experiments and comparisons we use the DUC 2007 dataset consisting of 45 topics. Each topic of 25 documents from the AQUAINT corpus of English news is to be summarized into a collective abstract of at most 250 words.

The evaluation measure is the well-known ROUGE (Recall-Oriented Understudy for Gisting Evaluation) [8]. We use the model summaries of the DUC datasets and their associated tools to compute the ROUGE measures. According to Lin and Hovy [7] the ROUGE-1 score has the best correspondence with human judgements. It is therefore the main focus of our evaluation. We experimented with several background corpora: the Brown corpus, the Gütenberg corpus, the Reuters RCV-1 corpus, as well as combinations.

**Data Preprocessing.** We remove all markup tags from the documents and leave only the headline and textual content of the news story. We then split the content to sentences with the DUC 2003 sentence segmentation tool and keep all words of length at least two.

**Comparative Evaluation.** We compare the Association Mixture Text Summarization method against results given in literature for state-of-the-art unsupervised summarization methods: Document Summarization Based on Data Reconstruction, linear and non-linear (DSDR-lin, DSDR-non) [5], Topic DSDR (TDSRD) [11], Two-Tiered Topic Model (TTM) and Enriched TTM (ETTM) [1]. The last two use Wordnet and topic description as additional resources. We also include two baseline methods provided with the DUC: NIST BL and CLASSY04. The latter is actually a supervised method.

## 4.2 Results

**Association Mixture Model and Its Two Components:** In terms of F-measure for ROUGE-1, Figure 1 illustrates the performance of the overall model and the independence and background corpus components as functions of the size of the background corpus $\mathcal{B}$.

The performance improves from 0.380 to 0.422 as the size of the background $\mathcal{B}$ grows from 10 to 10,000 sentences. This illustrates how a larger background corpus is a simple but effective way to provide auxiliary information to the summarization process. In our experiments, 1,000–3,000 sentences were already sufficient as a background corpus. The improvement after this was very limited.

Next, consider the performance of the two components of the model individually. The independence component does obviously not depend on the background corpus $\mathcal{B}$ and is hence represented by a horizontal line on the figure.

The background component, in turn, shows a longer period of improvement than the Association Mixture model and converges later than the 1,000–3,000 sentences range.

Overall, the Association Mixture Text Summarization method seems to successfully combine the two components into a model that clearly dominates both of them. Contrary to our expectations, there is a clear margin over the background component for large background corpus sizes, even though the relative weight of the independence component is very small there.

**Comparison to Other Methods.** A comparison to state-of-the-art in unsupervised summarization methods shows that the Association Mixture model is very competitive (Table 1). ROUGE-1 results are additionally shown as thin, unlabeled horizontal lines in Figure 1.

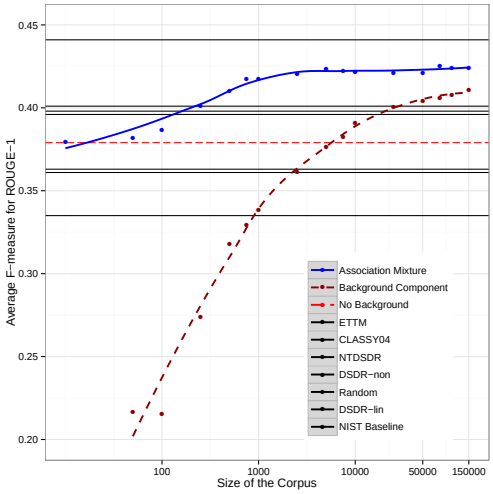



**Figure 1: Performance of the methods in terms of average ROUGE-1 F-measure, as the function of the size of the background corpus $\mathcal{B}$ (smooth curves obtained by LOESS regression).**

| Method | Rouge-1 | Rouge-2 | Rouge-3 | Rouge-L |
|---|---|---|---|---|
| NIST BL | 0.335 | 0.065 | 0.019 | 0.311 |
| DSDR-lin [5] | 0.361 | 0.072 | 0.021 | 0.324 |
| Random | 0.363 | 0.064 | 0.018 | 0.335 |
| DSDR-non [5] | 0.396 | 0.074 | 0.020 | 0.353 |
| NTDSDR [11] | 0.398 | 0.082 | - | 0.362 |
| CLASSY04 | 0.401 | 0.093 | 0.031 | 0.363 |
| Assoc. Mix.$^+$ | 0.424$^+$ | 0.104$^+$ | 0.036$^+$ | 0.384$^+$ |
| ETTM [1]$^*$ | 0.441$^*$ | 0.104$^*$ | - | - |
| TTM [1]$^*$ | 0.447$^*$ | 0.107$^*$ | - | - |

**Table 1: Average F measures for the DUC 2007 dataset.** $^*$Uses Wordnet and topic descriptions as additional resources. $^+$Uses background corpus as an additional resource. Paired Wilcoxon Test p-values are below 0.0004 between CLASSY04 and Assoc. Mix for all metrics.

The Association Mixture Text Summarization method outperformed all unsupervised approaches that do not rely on additional resources, and did this already with a background corpus of 300 sentences.

Among the tested methods, the Association Mixture Text Summarization method was only outperformed by the Two-Tiered Topic Models TTM and ETTM [1]. These methods use Wordnet and a topic description as additional resources, while we use a raw unprepared background corpus (with similar performance improvement with different genres and types of background corpora). It seems natural that methods using such manually crafted resources as Wordnet do better than methods using simple corpora.

## 5. CONCLUSIONS

In this paper we have proposed the Association Mixture Text Summarization method for creating (multi-)document summaries based on word associations. This approach has a number of characteristics: (i) it looks for relevant associations rather than words, (ii) it generalizes to multiple documents, (iii) it is unsupervised and uses simple resources, and thus it is (iv) largely language-independent.

In our experiments, the Association Mixture Text Summarization method outperformed resource-free unsupervised summarization methods and its performance was comparable to systems which use hand-crafted linguistic resources. Its performance converged when the size of the background reached approximately 1,000–3,000 sentences.

The only language-specific resource required by the method is a background corpus of some thousands of sentences, and the only required linguistic processing is the ability to split a text into sentences and its sentences into words. The simplicity of the method and its very modest requirements should make it universally applicable.

*Acknowledgements.*

This work has been supported by the European Commission (FET grant 611733, ConCreTe) and the Academy of Finland (Algodan Centre of Excellence).

## 6. REFERENCES

[1] A. Celikyilmaz and D. Hakkani-Tür. Discovery of topically coherent sentences for extractive summarization. In *ACL*, pages 491–499, 2011.

[2] S. C. Deerwester, S. T. Dumais, T. K. Landauer, G. W. Furnas, and R. A. Harshman. Indexing by latent semantic analysis. *JASIS*, 41(6):391–407, 1990.

[3] T. Dunning. Accurate methods for the statistics of surprise and coincidence. *Computational linguistics*, 19(1):61–74, 1993.

[4] Y. Gong and X. Liu. Generic text summarization using relevance measure and latent semantic analysis. In *24th international ACM SIGIR conference on Research and development in information retrieval*, pages 19–25. ACM, 2001.

[5] Z. He, C. Chen, J. Bu, C. Wang, L. Zhang, D. Cai, and X. He. Document Summarization Based on Data Reconstruction. In *Twenty-Sixth AAAI Conference on Artificial Intelligence*, pages 620–626, 2012.

[6] D. S. Johnson. Approximation algorithms for combinatorial problems. *Journal of Computer and System Sciences*, 9(3):256–278, Dec. 1974.

[7] C.-Y. Lin and E. Hovy. Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Proceedings of the 2003 Conference of the North American Chapter of the ACL*, NAACL, pages 71–78. Association for Computational Linguistics, 2003.

[8] C.-Y. Lin and F. Och. Looking for a few good metrics: Rouge and its evaluation. In *NTCIR Workshop*, 2004.

[9] G. Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995.

[10] A. Nenkova and K. McKeown. A survey of text summarization techniques. In *Mining Text Data*, pages 43–76. Springer, 2012.

[11] Z. Zhang, H. Li, et al. TopicDSDR: combining topic decomposition and data reconstruction for summarization. In *Web-Age Information Management*, pages 338–350. Springer, 2013.

# Paper IV

Oskar Gross, Antoine Doucet, and Hannu Toivonen

**Language-Independent Text Summarization with Document Specific Word Associations**

In *Proceedings of 31st ACM Symposium on Applied Computing, Accepted for Publication*, ACM, 2016



IV

# Language-Independent Multi-Document Text Summarization with Document-Specific Word Associations

Oskar Gross
Department of Computer Science and HIIT
University of Helsinki, Finland
oskar.gross@cs.helsinki.fi

Antoine Doucet
Laboratoire Informatique, Image et Interaction
University of La Rochelle, France
antoine.doucet@univ-lr.fr

Hannu Toivonen
Department of Computer Science and HIIT
University of Helsinki, Finland
hannu.toivonen@cs.helsinki.fi

## ABSTRACT

The goal of automatic text summarization is to generate an abstract of a document or a set of documents. In this paper we propose a word association based method for generating summaries in a variety of languages. We show that a robust statistical method for finding associations which are specific to the given document(s) is applicable to many languages. We introduce strategies that utilize the discovered associations to effectively select sentences from the document(s) to constitute the summary. Empirical results indicate that the method works reliably in a relatively large set of languages and outperforms methods reported in MultiLing 2013.

## CCS Concepts

•**Information systems** → **Summarization;** *Data mining;* •**Computing methodologies** → *Semantic networks;* Information extraction; •**Applied computing** → Digital libraries and archives;

## Keywords

Natural language processing; text summarization; text mining; co-occurrence analysis

## 1. INTRODUCTION

The amount of information on the Internet is growing so rapidly that methods which are able to make it consumable for users, e.g., by summarization, are becoming more important every day. The problem is emphasized with news stories, where several news providers report on same events using similar facts. Automatic text summarization is one way to solve this problem by creating a comprehensive summary of a given set of documents. Effective summarization potentially makes it much easier for the readers to obtain the information efficiently.



*SAC 2016, April 04 - 08, 2016, Pisa, Italy*



ACM ISBN 978-1-4503-3739-7/16/04…$15.00

DOI: http://dx.doi.org/10.1145/2851613.2851647

In text summarization, one or more documents on some topic are abstracted into a shorter text. Summarization methods are needed essentially for all written languages but developing them separately is a huge effort. Motivated by this need, we introduce a summarization method that makes only some minimal assumptions about the language: that the text can be split to sentences (based on punctuation) and sentences further to words (based on white space). In the experiments of this paper, we applied it successfully on nine different languages without any language-specific resources, tools, or tuning.

The method we propose analyses co-occurrences of words in the given document and uses this information to pick suitable sentences from the document to produce a summary for it. It has been shown before that discovery of document-specific associations works well for summarization of the English language [12]. In this paper we extend the method and show that this method is more universal; in particular, we apply it to many languages and to multi-document summarization.

A central task in text summarization is to detect what is important in the given documents. The crux of the method proposed here is to statistically identify pairwise word associations which are characteristic and specific to these documents. To a degree this is similar to finding relevant words, e.g., using tf-idf. Obviously, associations (i.e., pairs of words) are more informative than individual words.

For instance, *accident* is a frequent word in news stories, and so is *Obama* at the time of writing of this paper. A hypothetical document talking about an accident to President Obama is characterized by the *combination* of these two common words, and our goal is to be able to recognize such unexpected combinations. In contrast, a purely keyword-based method fails to discover the connection, and may actually miss both words if they are sufficiently common in news in general.

On the other hand, the combination of *Obama* and *president* in a news story is not interesting since it is not unexpected. The method we use therefore down-weighs word pairs that are frequent in general.

The main contributions of the paper are the following.

- Using document-specific word associations as a model of the document, we propose two novel measures of how well a summary represents that model. Both outperform the previous measure based on word associations [12]. We also consider two alternative optimiza-

tion techniques to find a set of sentences to be used as the summary; one technique is a greedy one, the other one uses a genetic algorithm.

- The method is based on a simple model: a document is a set of sentences, and each sentence is a set of words. This makes the method easily applicable to different languages; we have used it on nine languages without any language-specific pre-processing at all. The model also makes multi-document summarization trivial: a set of documents is simply a larger set of sentences.
- The method outperforms existing methods when tested in multi-document summarization tasks in nine different languages; we evaluated the method experimentally on the tasks of MultiLing 2013 [9], an event for multilingual multi-document summarization. In six languages it gives the best results, in the remaining three it is among the best ones.

The strong empirical results in document summarization suggest that document-specific associations do capture essential aspects of the documents across several languages. There probably are other applications for such automatically extracted information besides summarization.

In the rest of the paper we will first give an overview of related work in language-independent text summarization. In Section 3 we describe the problem formally. We continue by introducing the method in Section 4. The performance of the method is assessed empirically in Section 5. The paper is concluded in Section 6.

## 2. RELATED WORK

Text summarization is the task of automatically building short summaries of longer documents. It is a well-studied area, addressed with two main approaches. The first approach is to select existing sentences (or phrases or words) to form the summaries, in what is termed "extraction-based" summarization. In contrast, "abstraction-based" methods use natural language generation methods to represent the original document in a condensed form. Hybrids exist where sentences are altered, using techniques such as sentence compression around the key parts of the text. In addition, all of these approaches can either be supervised or unsupervised.

In this paper, we focus on unsupervised approaches, in which there is no human intervention in the summarization process whatsoever. An exhaustive review of such techniques is provided by Nenkova and McKeown [20]. Further, we focus on extraction-based approaches.

To perform unsupervised summarization, several techniques rely on Latent Semantic Analysis (LSA) [5] as their basis (e.g. [11]). An example of purely unsupervised summarization is the DSDR method of He et al. [13]. This approach generates a summary by using sentences that best "reconstruct" the original document, in its diversity. This work has been extended by combining document reconstruction and topic decomposition [23].

An approach more closely related to ours is that of Baralis et al. [1, 2] who treat sentences as sets of items (i.e., words) and choose the sentences by using an approach based on frequent weighted itemsets. The difference to our method is that we neither use frequent itemsets nor association rules but exploit all pairs of words co-occurring in the same sentence. Perhaps more importantly, our method calculates its measure of relevance of associations by incorporating information from a background corpus, in order to contrast the document against general expectations about word associations.

A number of unsupervised methods take advantage of additional linguistic resources. In particular, the Two-Tiered Topic model by Celikyilmaz [4] uses Wordnet [19] and the DUC-provided user query for selecting the summary sentences. The Document Understanding Conference(DUC) provides most evaluation procedures and collections in the summarization field.

In this paper, our applications are in the specific task of multi-document summarization, in which a single summary needs to be constructed for a set of documents written about the same topic. This task has been shown to be more complex than single-document summarization as a larger set of documents inevitably induces a wider thematic diversity [17, 18].

Few techniques are language-independent, unsupervised and effective also in multi-document summarization. The most successful approach of the multilingual multi-document summarization workshop (MultiLing 2013) was UWB [22], a method based on singular value decomposition (SVD). UWB performed best in almost all the languages tested in MultiLing 2013.

## 3. PROBLEM

We will next formulate the problem of text summarization. Since the evaluation of summaries is an integral part of the problem, we also discuss methods to evaluate the generated summaries.

### *Formulation.*

Let $U$ be the universe of all possible sentences. We denote by $D$ the given set of documents to be summarized. We ignore sentence order in documents, so each document $d \in D$ is simply a subset of all possible sentences, $d \subset U$.

Given a set of documents $D$, consisting in total of $c$ words, the task is to summarize it into a document $\hat{d}$ consisting of at most $k$ words, where $k \ll c$. Conceptually the goal is to create a document $\hat{d}$ such that the information contained in $\hat{d}$ is in some sense as similar to the document set $D$ as possible:

$$\hat{d} = \max_{\substack{d' \subset U: \\ ||d'|| \leq k}} sim(d', D),$$

where $d'$ can be any set of sentences consisting of at most $k$ words.

In extraction-based summarization, the universe $U$ is restricted to the sentences found in $D$, i.e., $U = \bigcup_{d \in D} d$.

### *Evaluation.*

Evaluation of summaries is difficult since the similarity function $sim()$ above is difficult if not impossible to define objectively. In practical evaluations of summaries, it usually is based on human assessment, or on some rough computational similarity measure between a computer-generated summary and human-written summaries.

A classical method for automatic evaluation of summaries is ROUGE [15], also used in this paper. ROUGE uses n-gram analysis to calculate a similarity between human-written model summaries and automatically generated summaries. According to Lin and Hovy [16], ROUGE-1 score

corresponds best to human judgement. Giannakopoulos and Karkaletsis have also proposed graph based measures AutoSummENG and MeMog for evaluating summaries [10]. They show that these measures correlate well with the ROUGE-2 measure.

*Complexity.*

Extraction-based summarization is a restricted version of the general summarization problem. Under some reasonable assumptions the problem then reduces to a (weighted) set cover problem [14]: we have to choose a set $\hat{d} \subset U$ of sentences such that $\hat{d}$ maximally covers the information in $D$.

The set cover problem is known to be NP-hard, so even if $sim()$ could be defined optimally and even if it could be computed efficiently, the problem would still remain computationally hard.

# 4. METHOD

The key problem in extraction-based summarization is how to measure the importance of a sentence. We use a method that estimates the importance of word pairs in the given document, and then weighs a sentence by the word pairs it contains.

## 4.1 Defining Document-Specific Word Associations

Let us start with some notation and simplifying assumptions we make.

Let $T$ denote the set of all words. A sentence $s$ is, in our model, simply a set $s = \{t_1, \cdots, t_k\}$ of words $t_i \in T$, i.e., we ignore the order of words.

In the case of multi-document summarization, as in the experiments of this paper, we simply consider the documents to be summarized as one long document $d_s = \bigcup_{d \in D} d$. Multi-document summarization is thus trivially reduced to single-document summarization.

*Mixture Model for Word Co-occurrence.*

The goal is to identify word associations that are more common in the document than expected. We next describe the statistical Mixture model for what is considered "expected", following Gross et al. [12]. Co-occurrences of words are here considered on sentence level. The Mixture model considers and combines two aspects of what is expected.

First, if $t_i$ = *Obama* and $t_j$ = *Putin* are both frequent within a document, then it is likely that they also co-occur several times in the same sentence within the document. To estimate this probability, the method needs frequencies of $t_i$ and $t_j$ in the document $d_s$ to be summarized. These are denoted by $n_i$ and $n_j$, respectively, while $n_{ij}$ denotes the frequency of their co-occurrence and $n$ the total number of sentences in $d_s$. Assuming that $t_i$ and $t_j$ are statistically independent, their expected frequency of co-occurrence is then $E_d(n_{ij}) = n_i \cdot n_j / n$.

Second, if the pair $t_i$ = *Barack* and $t_j$ = *USA* co-occurs frequently in news stories in general, then their co-occurrence is not unexpected in a given news document $d_s$. In order to estimate how often word pairs are likely to co-occur in general, the method also computes word and word pair frequencies in a background corpus $\mathcal{B}$. These frequencies are denoted by $m_i$, $m_j$, $m_{ij}$, and $m$, similarly to the counts obtained for the document $d_s$. The expected frequency of co-occurrence in $d_s$ then is $E_{\mathcal{B}}(n_{ij}) = n \cdot m_{ij}/m$.

The Mixture model combines these two models and estimates the probabilities of words $t_i$ and $t_j$ and of their co-occurrence, denoted by $p_i$, $p_j$ and $p_{ij}$, respectively, as

$$\begin{aligned} p_i &= (n_i + m_i)/(n+m), \\ p_j &= (n_j + m_j)/(n+m), \\ p_{ij} &= (n_i \cdot n_j/n + m_{ij})/(n+m). \end{aligned}$$

Probabilities $p_i$ and $p_j$ are obtained in a straightforward manner from the frequencies of $t_i$ and $t_j$, respectively, in the union of $\mathcal{B}$ and $d_s$. This definition of $p_i$ equals the average of probabilities $n_i/n$ and $m_i/m$ weighted by their sample sizes $n$ and $m$, respectively (and similarly for $p_j$).

The probability $p_{ij}$ of co-occurrence is conceptually also estimated from the union of $\mathcal{B}$ and $d_s$, but *not* using the observed frequency of co-occurrence $n_{ij}$ in $d_s$—since we want to estimate if it is unexpected or not—but instead under the assumption that words $t_i$ and $t_j$ are statistically independent in $d_s$. This definition equals the average of probabilities $E_d(n_{ij})/n$ and $E_{\mathcal{B}}(n_{ij})/n$, weighted again by the sample sizes $n$ and $m$, respectively.

As can be seen from above, the method combines two models into one mixture model: one based on the document itself, another one based on the background; hence the name *Mixture*. The motivation for using the Mixture model is that we cannot always assume that the distributions between the background and the document are similar, thus we draw from both models.

*Weighting Word Associations.*

We use log-likelihood ratio (llr) to measure the unexpectedness of word associations in $d_s$ [6]. The measure compares the fit of two multinomial models to the data, one is a null model and the other is an alternative model. The null model is the Mixture model described above, defining what is expectable under the assumptions of the model. The alternative model is the maximum likelihood model obtained from $d_s$, where probabilities are estimated directly from the document itself: $q_i = n_i/n$; $q_j = n_j/n$; $q_{ij} = n_{ij}/n$.

The multinomial models actually have as their parameters the probabilities of the mutually exclusive cases of co-occurrence of $t_i$ and $t_j$ ($p_{ij}$; already known from above), of occurrence of $t_i$ without $t_j$ (denoted by $p_{i-j}$), of occurrence of $t_j$ without $t_i$ (denoted by $p_{-ij}$), and of absence of both (denoted by $p_{-i-j}$). We can obtain these parameters easily from the previously defined probabilities: $p_{i-j} = p_i - p_{ij}$; $p_{-ij} = p_j - p_{ij}$; $p_{-i-j} = 1 - p_{ij} - p_{i-j} - p_{-ij}$.

The log-likelihood ratio is then computed as [6, 12]

$$LLR(t_i, t_j) = 2 \sum_{\substack{a \in \{\text{“}ij\text{”}, \text{“}i-j\text{”}, \\ \text{“}-ij\text{”}, \text{“}-i-j\text{”}\}}} n_a (\log p_a - \log q_a).$$

*Document-specific associations* are now obtained by selecting those word pairs for which the log-likelihood ratio is greater than zero, $LLR(t_i, t_j) > 0$, and which co-occur at least twice in the document $d_s$. The latter condition reduces noise caused by rare words and co-occurrences.

## 4.2 Sentence Selection

Document-specific associations presumably carry essential information about the document, and earlier results indicate that this is indeed the case, at least in English [12]. The

next task is to take advantage of the discovered document-specific associations and pick sentences from the document to generate a summary of it. In this paper, we will use three strategies: a) pick sentences that cover as many of the associations as possible [12]; b) pick sentences that cover the most central nodes in the term-association graph; c) combine the two strategies above. We will next define sentence-scoring functions for these three strategies, and then will consider two optimization methods for picking the best possible sentences.

As some of the components also incorporate graph algorithms, we also consider a graph $G = (V, E, W)$, where $V = \bigcup_{s \in d_s} s$ is the set of nodes (all words in the document $d_s$),

$$E = \{\{t_i, t_j\} \mid t_i \neq t_j, \exists s \in d_s \text{ s.t. } \{t_i, t_j\} \subset s\}$$

is the set of edges (associations between words), and $W : V \times V \rightarrow \mathcal{R}$ maps an edge $e$ to a positive real number (i.e. edge weight). The log-likelihood ratio $LLR$ is used as the edge weight, i.e., $W(t_i, t_j) = LLR(t_i, t_j)$.

### *Covering Associations.*

The assumption given above is that stronger associations cover the most important relations between words in the given document. Hence, having as many of the most important associations also in the generated summary is a natural goal [12]. However, rather than aiming to actually replicate the document-specific associations in the summary, the aim is to have many of them as word co-occurrences. In other words, the goal is to pick sentences so that the words of each important association co-occur in at least one sentence of the summary. This choice is motivated by the need to produce short summaries; statistics based on the number of co-occurrences would indeed have large variance and would not likely be reliable.

This task now reduces to the weighted set cover problem. The best summary consists of the set of sentences that covers as many of the heaviest associations (edges) as possible. The score of a summary $S$ (a set of sentences) is

$$cover(S) = \sum_{\substack{e \in E: \\ \exists s \in S \text{ s.t. } e \subset s}} W(e),$$

and the summarization task now reduces to finding the set $S$ of sentences that matheximizes the score, when the size of the summary $S$ is constrained to at most $k$ words.

### *Covering Central Words.*

We propose word-centrality as an alternative measure to the graph coverage above. We still use word-associations, but instead of covering associations (edges in the word association graph), we aim to cover important words (nodes in the word association graph). The rationale here is that the most central words in the graph induced by pairwise associations are central concepts of the document.

To measure the importance of words, given word associations, we use the document graph $G$ and calculate the closeness centrality [8] for each of the nodes in the graph. For a node $v \in V$, the centrality is

$$C(v) = \frac{|V|}{\sum_{u \in V} d(u, v)},$$

where $d(u, v)$ is the length of the shortest path between nodes $u$ and $v$; the length of a path is computed as the sum of inverse weights $1/W(e)$ of its edges.

Similarly to the association cover, we now obtain a centrality score for summary $S$ as follows:

$$centrality(S) = \sum_{\substack{v \in V: \\ \exists s \in S \text{ s.t. } v \in s}} C(v).$$

### *Covering Associations and Central Words.*

While both measures above are based on document-specific word associations, it possible that they capture different nuances of the document. In case these differences are complementary, some combination of the measures potentially outperforms either one.

We propose to define such a combination simply as their sum. However, to give both components roughly equal weight, we first normalize both scores to be between 0 and 1:

$$combined(S) = \frac{cover(S)}{\sum_{\{t_i, t_j\} \in E} W(t_i, t_j)} + \frac{centrality(S)}{\sum_{v \in V} C(v)}.$$

### *Greedy Optimization Strategy.*

As was already noted above, the problem of selecting an optimal set of sentences to form summary $S$ is NP-hard. We will use two alternative heuristics this: a greedy strategy is described in this subsection, and a method based a genetic algorithm in the next one.

The standard greedy algorithm first takes the sentence which covers as many word associations as possible, and then chooses the next sentence ignoring the already covered pairs [12]. We can directly apply the same greedy strategy also to cover central words, or the combined measure.

For the sake of simplicity, consider the graph $G$ induced from the document $d_s$ and an initially empty set $S = \emptyset$, to which sentences will be added to constitute the final summary. The score, according to which individual sentences $s$ are selected by the greedy approach, normalizes the additional coverage given by a sentence $s$ with its length $|s|$: $cover(s) = \sum_{e \in E:\ e \subset s} W(e)/|s|$.

Similarly, we obtain a sentence scoring function based on word centrality: $centrality(s) = \sum_{t \in s} C(t)/|s|$.

Algorithm 1 describes the greedy process for the original graph cover. It can be easily adapted for *centrality*() and *combined*(). The algorithm first selects the best-scoring sentence $\hat{s}$ and adds this to the summary, $S = \hat{s} \cup S$. The graph is next updated by removing from $G$ all the edges between nodes which co-occur in $\hat{s}$. This step downweighs sentences that contain already covered pairs, and effectively also prevents selection of duplicates. The sentence selection and graph update process is then repeated until no more sentences can be added to the summary within the limit of $k$ words. When applied to *centrality*() or *combined*(), a record must be kept of words not yet covered. However, node-centrality scores should *not* be updated in the process.

### *Genetic Algorithm.*

The second approach for finding sentences which best cover the document-specific associations or words is an evolutionary algorithm. We chose the evolutionary algorithm

**Algorithm 1** Greedy Selection Algorithm

**procedure** GREEDYSELECT
  **Input:** $d_s$, a set of sentences to be summarized
  **Output:** $S \subset d_s$, a summary of $d_s$
  $S \leftarrow \emptyset$ ▷ An initially empty summary
  $ls \leftarrow 0$ ▷ Current summary length
  **while** $ls < k$ **do**
    $\hat{s} \leftarrow null$
    $\hat{s} \leftarrow \underset{\substack{s \in d_s: \\ |s|+ls \leq k}}{\operatorname{argmax}} cover(s)$
    **if** $\hat{s} = null$ **then**
      *break*
    **end if**
    $S \leftarrow S \cup \hat{s}$
    **for** $(t_i, t_j) \subset \hat{s}$ **do**
      $W(t_i, t_j) \leftarrow 0$
    **end for**
    $ls \leftarrow ls + |s|$
  **end while**
  **return** S
**end procedure**

| Parameter | Value |
|---|---|
| $\lambda$, population size | 300 |
| $\mu$, number of best individuals selected for producing offspring | 100 |
| crossover rate (probability of crossover) | 0.3 |
| mutation rate (probability of mutation) | 0.7 |
| number of iterations | 150 |

**Table 1: The parameters for the genetic algorithm.**

as an alternative optimization method as it makes few assumptions about the underlying fitness landscape and it is easy to apply to different kinds of problems.

For the evolutionary algorithm we need to define the genome, mutation, crossover and scoring function. We defined the *genome* to be a set of sentences (technically, sentence identifiers). The *crossover* function is defined between two individuals $a$ and $b$ (sets of sentences) and produces two individuals $a'$ and $b'$ to the offspring. For each sentence in $a$ and $b$ we will uniformly randomly assign the sentence to $a'$ or $b'$. The *mutation* function takes an individual $a$ as input and generates a new modified individual $a'$, by randomly adding or removing a random sentence in $a'$. The *scoring* function is either $cover(S)$, $centrality(S)$ or $combined(S)$. However, if the individual contains more than $k$ words then the score is 0.

We used the $(\mu + \lambda)$ elitist strategy [21] for the optimization and the DEAP [7] package for its implementation.

In order to use the strategy there is a number of parameters to set. The parameter values were obtained by rough experimental analysis of the convergence speeds on the English language (Table 1).

# 5. EVALUATION

We next carry out an empirical evaluation of the proposed method. The general aims are to obtain a view to the overall performance of the method, and to the effects of its various components (scoring methods, optimization techniques). Specifically, we will look for answers to the following questions. (1) Which graph based scoring method captures the information of the documents best (i.e. *cover*(), *centrality*(), *combined*())? (2) What is the effect of the optimization strategy on the results (greedy vs. genetic algorithm)? (3) How reliably and consistently does the method work for different languages? (4) How does the method perform in comparison to other systems?

## 5.1 Experimental Setup

### *Evaluation Method.*

We will use the ROUGE [15] evaluation method for evaluating summaries. The ROUGE method uses the overlap of n-grams between model summaries, written by humans, and generated summaries to measure the similarity. For instance, ROUGE-1 score just looks at unigrams, ROUGE-2 score looks at 2-grams and ROUGE-L looks for the longest common sequence between two texts. The ROUGE score breaks down into two components, precision and recall. For evaluation we will use the combined score, F-measure, computed as the harmonic mean between precision and recall. We originally attempted to use the evaluation method MeMoG [10], also used in MultiLing 2013, but were not able to reproduce the evaluation results published in MultiLing so we resorted to ROUGE standard instead.

### *Dataset.*

We use the MultiLing-2013 [9] dataset to evaluate our method. The dataset contains documents in 10 different languages – English, French, Chinese, Romanian, Spanish, Hindi, Arabic, Hebrew, Greek and Czech. Our method assumes that the text has been (or can trivially be) broken to words. Since this assumption does not hold for Chinese, we omitted it from our experiments.

MultiLing contains 15 topics for each language except for French and Hindi, for which the number of topics is 10. On average each topic consists of 10 documents which need to be collectively summarized into a text of 250 words. In our case, this multi-document summarization task is trivially reduced to single-document summarization by concatenating the documents into one set of sentences. The background corpus consists of the documents in the same language except the documents being summarized.

Additionally, MultiLing 2013 has made available the summaries generated by systems that participated in the event. In our comparisons below with other systems, we have computed ROUGE scores etc. for the other systems from the original summaries they have provided, i.e., we did not re-implement nor re-run any of the systems.

### *Notation.*

We have above proposed several alternative configurations for word-association based summarization, and we use the following notation to denote these configurations. First, the options for sentence-scoring are the graph cover measure (denoted by $G$), word centrality measure ($C$), and their combination ($G+C$). Second, optimization strategies for sentence selection are the greedy method ($GR$) and the genetic algorithm ($GA$). We refer to a combination of a scoring measure and an optimization method by their concatenation, e.g., $G\_GA$ refers to the graph cover scoring, optimized using a genetic algorithm.

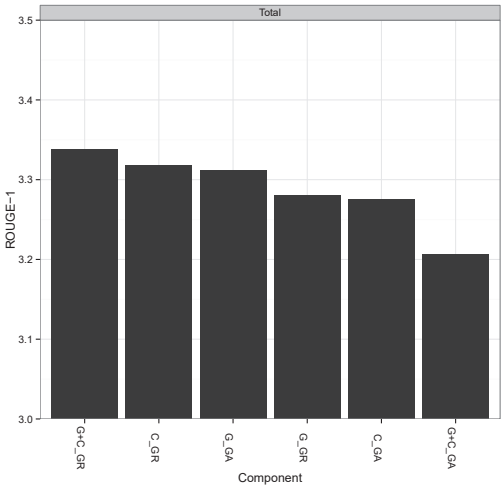



**Figure 1: The performance of the different scoring methods with different optimization strategies. Note that the y-axis is limited to the range $3 - 3.5$.**

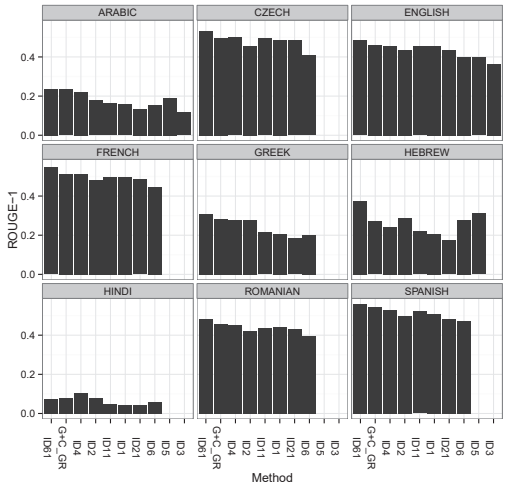



**Figure 2: A comparison between all systems for all languages. Note that ID61 is not a real summarization method (see text).**

## 5.2 Sentence Scoring Methods

First, we will take a look at how the different sentence scoring measures perform with different optimization methods. Instead of looking at individual languages here, we will compare the total scores obtained over all the languages with ROUGE-1. The scores for different combinations of scores and optimization strategies can be seen in Figure 1.

Best results are obtained with the combined measure $G + C$, followed by the word centrality-based measure $C$ and then the graph cover based measure $G$. The differences between these measures are relatively small, however. For the question (1) we conclude that most likely both the word centrality measure and the graph association measure cover important parts of the document. As the $G + C$ measure performed in our experiment a bit better than either of the individual measures alone, it suggests that the measures do capture different nuances of the documents.

Between the two optimization methods (question 2), the greedy algorithm tends to perform better than the genetic algorithm (Figure 1). This is a slight surprise since the genetic algorithm should be able to explore a much wider space of possible summaries. The relatively poor performance of the genetic algorithm here is probably due to the simplistic setup; genetic algorithms designed specifically for the weighted set cover problem are known to produce better results than standard solutions [3]. The greedy method is known to be suboptimal, but a positive interpretation of the results here is that the greedy method actually performs well and cannot be easily outperformed.

## 5.3 Language-Wise Performance

Next we will take a look at summarization performances for individual languages, for the best variant $G + C\_GR$ of our method, as well as the participants of MultiLing 2013. Our main aim in this subsection is to compare the stability of the performance or our method in different languages; a systematic comparison to the other methods is provided in the next subsection.

Before going to the results, let us introduce the baseline methods for MultiLing 2013: a global baseline (ID6) and a global topline (ID61). The *global baseline* system ID6 is a simple vector space model based approach. It finds the centroid $C$ in the vector space and tries to generate text which is most similar to the centroid, according to the cosine measure. The *global topline* method ID61 is not a real summarization method, it is an approximation of the upper limit of performance in extraction-based summarization. It works similarly to ID6, but "cheats" by using human-written summaries to generate the vector space, and then chooses sentences from the original documents to create text which is most similar to the centroid. Among the summarization methods of MultiLing 2013, ID4 denotes the best performing method, UWB [22]. For other methods we refer to the MultiLing 2013 overview paper [9].

Results over all methods and languages can be seen in Figure 2. There are two main observations to be made.

First, the proposed method is highly competitive against the other systems. It actually performs best among the automatic systems for six out of the nine languages (recall that ID61 is not an actual summarization system but an approximation of the upper limit). The method proposed here is outperformed only on Hebrew, Hindi and Czech.

Second, the results indicate that the proposed method is robust with regard to different languages, in the sense that it consistently ranks among the best ones and never loses much to the best one. On the other hand, some languages seem much more difficult for all methods, especially Hindi and Arabic, but also Greek and Hebrew, so robustness here

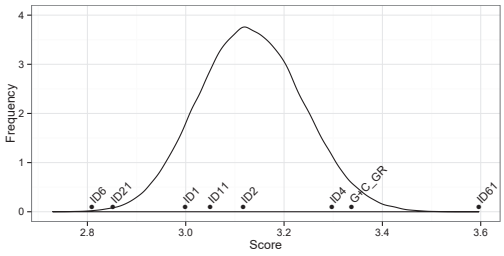



**Figure 3: The permutation test for MultiLing 2013 systems comparison to G+C_GR method. Note that ID61 is not a real summarization method.**

does not mean equally good absolute performance over all languages.

The answer to question (3) thus is that the proposed method seems to be generally applicable to many languages, with varying absolute performance but consistent relative performance in comparison to other methods applicable over a set of languages.

## 5.4 Statistical Comparison to Other Methods

Figure 2 already indicated strong relative performance of the method in comparison to other methods. We will now compare the performances of different methods statistically. We compare the total scores of our method, over all languages, to the scores of those methods that have results for all the languages in MultiLing 2013 (ID3 and ID5 were omitted since they only have results for some languages).

To avoid parametric assumptions about the distribution of scores, we carried out a permutation test as follows. The null hypothesis is that the proposed method is not statistically different from the other methods. In particular, for any given language, the proposed method could have received any of the scores that any method obtained for that language. Sampling a single random total score from this null hypothesis is easy: pick a random score for each language (among the ones obtained by the other systems) and sum up the scores.

By repeating this process 100 000 times we obtain an approximation of the distribution of total scores under the null hypothesis; this is shown as the curve in Figure 3. The total score of 3.337 obtained by our method can now be contrasted against the null distribution. The tail of the distribution starting from score 3.337 contains only 2.7% of the randomizations, i.e., the one-tailed empirical p-value is 0.027. Obviously, the same procedure can be used to obtain p-values for any of the methods.

Figure 3 also shows the total scores obtained by different methods. We can see that the global topline ID61 performs much better than any of the automatic systems. Among the real systems, the proposed method $G + C\_GR$ performs best, and is statistically significantly different from the other systems at level $< 0.05$ (empirical p-value 0.027). The significance level of ID4 is $< 0.1$ (empirical p-value 0.060).

A pairwise comparison between ID4 (UWB [22]) and $G + C\_GR$ using paired Wilcoxon rank sum test indicates that the methods are not statistically significantly different (p-value 0.20). Among the different configurations of our method we tested (Figure 1), ID4 would rank in the middle. On the other hand, even the worst of the configurations, the poorly optimized version $G + C\_GA$ of the same combined model, clearly outperforms the next best method, ID2.

| Method | ROUGE-1 | ROUGE-2 | ROUGE-L |
|---|---|---|---|
| ID61 | 3.60 | 1.51 | 3.09 |
| G+C_GR | **3.34** | 1.28 | **2.89** |
| ID4 | 3.30 | **1.36** | 2.87 |
| ID2 | 3.12 | 1.06 | 2.70 |
| ID11 | 3.05 | 1.13 | 2.61 |
| ID1 | 3.00 | 1.10 | 2.58 |
| ID21 | 2.85 | 1.01 | 2.44 |
| ID6 | 2.81 | 0.86 | 2.25 |

**Table 2: The average ROUGE scores for all the MultiLing 2013 methods. Note that ID61 is not a real summarization method (see text).**

Finally, Table 2 shows results also for ROUGE-2 and ROUGE-L. With ROUGE-2, ID4 (UWB) performs best, followed by the Mixture model. With ROUGE-L, the Mixture model wins again, with a small margin over ID4.

The answer to question (4) is that the performance of the proposed method is statistically significantly better than the performance of the other methods in general. It is not statistically significantly better than the UWB system [22] but the proposed method is more easily applicable to different languages: while UWB uses language-specific stop-word lists and various tunable parameters, the Mixture model has no parameters and uses no language specific resources except for a background corpus.

# 6. CONCLUSIONS

We have introduced a new method for automatically creating summaries for documents. The method is statistical in nature, and is based on analysis of the document itself, as well as comparing it to other documents. Word associations that are characteristic and specific to the given document are recognized first, and then a summary is constructed by picking those sentences from the document that best cover information in the strongest associations. We proposed new measures for the coverage that outperformed the previous measure [12].

The method is essentially language-independent: it only uses punctuation and white space to identify sentences and words. In our experiments, we did *not* use stemming or lemmatization, stopword lists, or any other language-specific tools or resources. These could probably be used to produce better results, but our goal here was to develop techniques that are readily applicable to a wide range of languages.

We evaluated the proposed method empirically using multi-document summarization tasks in nine different languages from MultiLing 2013. Overall, the method outperformed all methods that participated MultiLing: it ranked first in six languages out of nine, and was among the best ones in the remaining three. A statistical analysis shows that it is significantly better than the other methods in general (but not significantly better in a pairwise test than UWB [22], the best method of MultiLing 2013).

The superior performance of the method is striking given its extreme simplifications. Sentences are treated simply

as sets of words, and documents as sets of sentences. The multi-document summarization problem is trivially reduced to single-document summarization by taking the union of all documents. The method was successfully applied to nine different languages without any changes between languages. The results indicate strongly that document-specific word associations do capture central information of documents across several languages.

While the results are relatively speaking good, the summarization problem is all but solved. The coherence and fluency of generated summaries is an issue especially for methods based on sentence selection, such as ours. Further work is needed in making summaries better in these respects. Furthermore, interesting results could be obtained with hybrid approaches combining together language generation techniques and sentence selection techniques based on document-specific associations.

*Acknowledgements.*

This work has been supported by the European Commission (FET grant 611733, ConCreTe) and the Academy of Finland (decision 276897, CLiC).

## 7. REFERENCES

[1] E. Baralis and L. Cagliero. Learning from summaries: supporting e-learning activities by means of document summarization. *Emerging Topics in Computing, IEEE Transactions on*, (99):1–12, 2015.

[2] E. Baralis, L. Cagliero, A. Fiori, and P. Garza. Mwi-sum: A multilingual summarizer based on frequent weighted itemsets. *ACM Transactions on Information Systems*, 34(1):5:1–5:35, 2015.

[3] J. Beasley and P. Chu. A genetic algorithm for the set covering problem. *European Journal of Operational Research*, 94(2):392–404, 1996.

[4] A. Celikyilmaz and D. Hakkani-Tür. Discovery of topically coherent sentences for extractive summarization. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pages 491–499, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.

[5] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman. Indexing by latent semantic analysis. *Journal of the American Society for Information Science (JAIS)*, 41(6):391–407, 1990.

[6] T. Dunning. Accurate methods for the statistics of surprise and coincidence. *Computational linguistics*, 19(1):61–74, 1993.

[7] F.-A. Fortin, D. Rainville, M.-A. G. Gardner, M. Parizeau, C. Gagné, et al. DEAP: Evolutionary algorithms made easy. *The Journal of Machine Learning Research*, 13(1):2171–2175, 2012.

[8] L. C. Freeman. Centrality in social networks conceptual clarification. *Social networks*, 1(3):215–239, 1979.

[9] G. Giannakopoulos. Multi-document multilingual summarization and evaluation tracks in acl 2013 multiling workshop. In *Proceedings of the MultiLing 2013 Workshop on Multilingual Multi-document Summarization*, pages 20–28, Sofia, Bulgaria, August 2013. Association for Computational Linguistics.

[10] G. Giannakopoulos and V. Karkaletsis. AutoSummENG and MeMoG in evaluating guided summaries. In *Proceedings of the Text Analysis Conference (TAC 2011)*, Gaithersburg, Maryland, USA, November.

[11] Y. Gong and X. Liu. Generic text summarization using relevance measure and latent semantic analysis. In *24th international ACM SIGIR conference on Research and development in information retrieval*, pages 19–25, New Orleans, LA, USA, September 2001.

[12] O. Gross, A. Doucet, and H. Toivonen. Document summarization based on word associations. In *Proceedings of the 37th international ACM SIGIR conference on research & development in information retrieval*, pages 1023–1026, Gold Cost, Australia, 2014. ACM.

[13] Z. He, C. Chen, J. Bu, C. Wang, L. Zhang, D. Cai, and X. He. Document Summarization Based on Data Reconstruction. In *Twenty-Sixth AAAI Conference on Artificial Intelligence*, pages 620–626, Toronto, Ontario, Canada, July 2012.

[14] D. S. Johnson. Approximation algorithms for combinatorial problems. *Journal of Computer and System Sciences*, 9(3):256–278, Dec. 1974.

[15] C.-Y. Lin. ROUGE: A package for automatic evaluation of summaries. In *Text summarization branches out: Proceedings of the ACL-04 workshop*, pages 25–26, Barcelona, Spain, July 2004.

[16] C.-Y. Lin and E. Hovy. Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Proceedings of the 2003 Conference of the North American Chapter of the ACL*, NAACL, Edmonton, Canada, May–June 2003.

[17] H. Liu, H. Yu, and Z. Deng. Multi-document summarization based on two-level sparse representation model. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence,*, pages 196–202, Austin, Texas, USA, January 2015.

[18] K. McKeown and D. R. Radev. Generating summaries of multiple news articles. In *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '95, pages 74–82, New York, NY, USA, 1995. ACM.

[19] G. Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995.

[20] A. Nenkova and K. McKeown. A survey of text summarization techniques. In *Mining Text Data*, pages 43–76. Springer, 2012.

[21] H.-P. Schwefel. *Numerical optimization of computer models*. John Wiley & Sons, Inc., 1981.

[22] J. Steinberger. The UWB summariser at multiling-2013. In *Proceedings of the MultiLing 2013 Workshop on Multilingual Multi-document Summarization*, pages 50–54, Sofia, Bulgaria, August 2013. Association for Computational Linguistics.

[23] Z. Zhang, H. Li, et al. TopicDSDR: combining topic decomposition and data reconstruction for summarization. In *Web-Age Information Management*, pages 338–350. Springer, 2013.

# Paper V

Jukka M. Toivanen, Oskar Gross, Hannu Toivonen

**The Officer Is Taller Than You, Who Race Yourself! Using Document Specific Word Associations in Poetry Generation**

In *The Fifth International Conference on Computational Creativity*, 355-362, Josef Stefan Institute, Ljubljana, Slovenia, 2014.

# The Officer Is Taller Than You, Who Race Yourself! Using Document Specific Word Associations in Poetry Generation

**Jukka M. Toivanen, Oskar Gross, Hannu Toivonen**
Department of Computer Science and Helsinki Institute for Information Technology HIIT
University of Helsinki, Finland
jukka.toivanen@cs.helsinki.fi, oskar.gross@cs.helsinki.fi, hannu.toivonen@cs.helsinki.fi

**Abstract**

We propose a method for automatic poetry composition with a given document as inspiration. The poems generated are not limited to the topic of the document. They expand the topic or even put it in a new light. This capability is enabled by first detecting significant word associations that are unique to the document and then using them as the key lexicon for poetry composition.

## Introduction

This paper presents an approach for generating poetry with a specific document serving as a source of inspiration. The work is based on the corpus-based poetry composition method proposed by Toivanen et al. (2012) which uses text mining and word replacement in existing texts to produce new poems. We extend that approach by using a specific news story to provide replacement words to the automatic poetry composition system. New contributions of this work are in constructing a model of document-specific word associations and using these associations to generate poetry in such a way that a single generated poem is always based on a single document, such as a news story.

The method for finding document-specific word associations is based on contrasting them to general word associations. In a given document, some of the document's word associations are long-established and hence well-known links which are part of people's commonsense knowledge, whereas some are new links, brought in by the document. Especially in the case of news stories, these links are exactly the new information the document focuses on, and they can be used in a poetry generation system to produce poems that loosely reflect the topic and content of the specific document. However, the story or message of the document is not directly conveyed by the produced poem as the process of poetry composition is based on the use of word associations. Thus, the generated poetry is roughly about the same topic as the document but it does not contain the actual content of the document. Poetry composed with these word associations may evoke fresh mental images and viewpoints that are related to the document but not exactly contained in it.

The general goal of this work on poetry generation is to develop maximally unsupervised methods to produce poetry out of given documents. Thus, we want to keep manually crafted linguistic and poetry domain knowledge at minimum in order to increase the flexibility and language independence of the approach.

The next sections present briefly related work on poetry generation, introduce the method of constructing document-specific associations called here foreground associations and outline the procedure of using these associations in a poetry generation system. We also present some examples produced by the method and outline directions for future work.

## Related Work

**Poetry generation** Several different approaches have been proposed for the task of automated poetry composition (Manurung, Ritchie, and Thompson 2000; Gervás 2001; Manurung 2003; Diaz-Agudo, Gervás, and González-Calero 2002; Wong and Chun 2008; Netzer et al. 2009; Colton, Goodwin, and Veale 2012; Toivanen et al. 2012; Toivanen, Järvisalo, and Toivonen 2013). A thorough review of the proposed methods and systems is not in the scope of this paper but, for instance, Colton et al. (2012) provide a good overview.

The approach of this paper is based on the work by Toivanen et al. (2012). They have proposed a method where a template is extracted randomly from a given corpus and words in the template are substituted by words related to a given topic. In this approach the semantic coherence of new poems is achieved by using semantically connected words in the substitution. In contrast to that work, we use document-specific word associations as substitute words to make the new poems around specific stories. Toivanen et al. (2013) have also extended their previous work by using constraint-programming methods in order to handle rhyming, alliteration, and other poetic devices.

Creating poetry from news stories was also proposed by Colton et al. (Colton, Goodwin, and Veale 2012). Their method generates poetry by filling in user-designed templates with text extracted from news stories.

**Word association analysis** There is a vast number of different methods for co-occurrence analysis. In our work we have been careful not to fall into developing hand-tailored

methods, but to use more general approaches (i.e. statistics), which could be applied to all languages in which different words are detectable in text. Most prominent statistical methods for word co-occurrence analysis are log-likelihood ratio (Dunning 1993), Latent Semantic Analysis (Deerwester et al. 1990), Latent Dirichlet Allocation (Blei, Ng, and Jordan 2003) and Pointwise Mutual Information (Church and Hanks 1990; Bouma 2009).

In this work we build on the background association calculation method proposed by Gross et al. (2012) and its recent extension to document specific associations (Gross, Doucet, and Toivonen 2014). We will describe these models in some detail in the next section.

## What is Important in a News Story?

To produce a poem from a given news story, we first identify the essential features of its contents. News stories are normally summarized by their headlines, leads, topics, or keywords. For producing a poem, we are less interested in readily written descriptions such as the title and the lead, but more in text fragments such as keywords that we can use in poetry production. This also makes the approach more generic and not limited to just news stories.

Instead of keywords or topics, we propose to search for *pairs of associated words* in the document, as in Gross et al. (2014). The rationale is that often the core of the news content can be better summarized by the links the story establishes e.g. between persons, events, acts etc.

For illustration we use a BBC newspaper article on Justin Bieber drinking and driving on the streets of Miami, published on January 24, 2014[1]. As an example, consider the sentence *"Pop star Justin Bieber has appeared before a Miami court accused of driving under the influence of alcohol, marijuana and prescription drugs."* The associations which are rather common in this sentence are, e.g. "pop" and "star", "justin" and "bieber", "miami" and "court" – words which we know are related and which we would think of as common knowledge. The interesting associations in this sentence could be "bieber" and "alcohol", "bieber" and "prescription", "justin" and "alcohol" and so on.

We model the problem of discovering interesting associations in a document as novelty detection, trying to answer the questions "Which word pairs are novel in this document?" In order to judge novelty, we need a reference of commonness. We do this by contrasting the given *foreground* document to a set of documents in some *background corpus*. The idea is that any associations discovered in the document that also hold in the background corpus are not novel and are thus ignored. We next present a statistical method for extracting document-specific word associations.

We use the log-likelihood ratio (LLR) to measure document-specific word associations. LLR is a standard method for finding general associations between words (Dunning 1993). In our previous work, we have used it to build a weak semantic network of words for use in computational creativity tasks (Gross et al. 2012; Toivonen et al. 2013; Huovelin et al. 2013). In contrast to that work, here we look for deviations from the normal associations. This approach, outlined below, seems to be powerful in catching document specific information since it has been used as a central component in a successful document summarization method (Gross, Doucet, and Toivonen 2014).

We count co-occurrences of words which appear together in the same sentence. We do this both for the background corpus and the foreground document. Using LLR, we measure the difference in the relative co-occurrence frequencies. More specifically, the test compares two likelihoods for the observed frequencies: one (the null model) assumes that the probability of co-occurrence is the same as in the background corpus, the other (the alternative model) is the maximum likelihood model, i.e., it assumes that the probabilities are the same as the observed relative frequencies. We will next describe the way to calculate document specific association strengths in more detail.

### Counting Co-Occurrences

Consider two words $w_1$ and $w_2$ which appear in the document. We denote the number of times $w_1$ and $w_2$ appear together in a same sentence by $k_{11}$. The number of sentences in which $w_1$ appears without $w_2$ is denoted by $k_{12}$, and for $w_2$ without $w_1$ by $k_{21}$. The number of sentences in which neither of them occurs is denoted by $k_{22}$. In a similar way, we denote the counts of co-occurrences of words $w_1$ and $w_2$ in the background corpus by $k'_{ij}$ (cf. Table 1).

**Foreground Counts**

| | $w_1$ | $\neg w_1$ |
|---|---|---|
| $w_2$ | $k_{11}$ | $k_{12}$ |
| $\neg w_2$ | $k_{21}$ | $k_{22}$ |

**Background Counts**

| | $w_1$ | $\neg w_1$ |
|---|---|---|
| $w_2$ | $k'_{11}$ | $k'_{12}$ |
| $\neg w_2$ | $k'_{21}$ | $k'_{22}$ |

Table 1: The foreground and background contingency tables for words $w_1$ and $w_2$.

### Probabilities

We use a multinomial model for co-occurrences of words $w_1$ and $w_2$. In the model, each of the four possible combinations ($w_1$ and $w_2$ vs. $w_1$ alone vs. $w_2$ alone vs. neither one) has its own probability. In effect, we will normalize the values in the contingency tables of Table 1 into probabilities. These probabilities are denoted by $p_{ij}$ such that $p_{11} + p_{12} + p_{21} + p_{22} = 1$.

Let $m = k_{11}+k_{12}+k_{21}+k_{22}$ be the number of sentences in the foreground document. The values of the parameters can then be estimated directly from the document as $p_{ij} = \frac{k_{ij}}{m}$. The respective parameters can also be estimated from the background corpus. Let $m'$ be the number of sentences in the background, and let $q_{ij}$ be the parameters (instead of $p_{ij}$) of the multinomial model; then $q_{ij} = \frac{k'_{ij}}{m'}$.

Next we will use these probabilities in likelihood calculations.

[1] http://www.bbc.co.uk/news/world-us-canada-25863200

### Log-Likelihood Ratio

To contrast the foreground document to the background corpus, we will compare the likelihoods of the counts $k_{ij}$ in the foreground and background models. The foreground model is the maximum likelihood model for those counts, so the background model can never be better. The question is if there is a big difference between the models.

Let $P = \{p_{ij}\}$ and $Q = \{q_{ij}\}$ be the parameters of the two multinomial probability models, and let $K = \{k_{ij}\}$ be the observed counts in the document. Then, let $L(P, K)$ denote the likelihood of the counts under the foreground model, and let $L(Q, K)$ be their likelihood under the background model:

$$L(P,K) = \binom{k_{11}+k_{12}+k_{21}+k_{22}}{k_{11},k_{12},k_{21},k_{22}} p_{11}^{k_{11}} p_{12}^{k_{12}} p_{21}^{k_{21}} p_{22}^{k_{22}}$$

$$L(Q,K) = \binom{k_{11}+k_{12}+k_{21}+k_{22}}{k_{11},k_{12},k_{21},k_{22}} q_{11}^{k_{11}} q_{12}^{k_{12}} q_{21}^{k_{21}} q_{22}^{k_{22}}.$$

For contrasting the foreground to the background we compute the ratio between the likelihoods under the two models:

$$\lambda = \frac{L(Q,K)}{L(P,K)}. \quad (1)$$

The log-likelihood ratio test $D$ is then defined as

$$D = -2\log\lambda. \quad (2)$$

Given our multinomial models, the multinomial coefficients cancel out so the log-likelihood ratio becomes

$$D = -2\log\left(\frac{q_{11}^{k_{11}} q_{12}^{k_{12}} q_{21}^{k_{21}} q_{22}^{k_{22}}}{p_{11}^{k_{11}} p_{12}^{k_{12}} p_{21}^{k_{21}} p_{22}^{k_{22}}}\right), \quad (3)$$

which after further simplification equals

$$D = 2\sum_{i=1}^{2}\sum_{j=1}^{2} k_{ij}(\log(p_{ij}) - \log(q_{ij})).$$

The likelihood ratio test now gives higher values for word pairs whose co-occurrence distribution in the document deviates more from the background corpus.

For improved statistical robustness, we include the respective document in the background model, and in the case that the pair only co-exists in the document we estimate their joint co-occurrence probability under the assumption that the words are mutually independent. For more details, see Gross et al. (2014) who refer to these models as a Mixture model and an Independence model.

Given a document, we can now compute the above likelihood ratios for all pairs of words in the document. For poetry composition, we then pick from each document word pairs with the highest likelihood ratios and with $p_{11} > q_{11}$ to find the most exceptionally frequent pairs.

## Poetry Composition

We compose poetry using a word substitution method as described by Toivanen et al. (2012). Instead of explicitly representing a generative grammar of the output language or manually designing templates, the method copies a concrete instance from an existing text (of poetry) and substitute most of its contents by new words. One word of the original text is replaced at a time with a new, compatible word. In this method, compatibility is determined by syntactic similarity of the original and substitute word. Depending on the language, this requires varying degrees of syntactical and morphological analysis and adaptation. For more details on this part, see Toivanen et al. (2012).

In the current method, in contrast to the previous work outlined above, the topics and semantic coherence of the generated poetry are controlled by using the foreground associations. The document-specific foreground associations are used to provide semantically interconnected words for the content of a single poem. These words reflect the document in question but do not convey the actual content of the document. The idea is to produce poetry that evokes fresh mental images and thoughts which are loosely connected to the original document. Thus, the aimed style of the poetry is closely related to the imagist movement in the early 20th-century poetry which emphasised mental imagery as an essence of poetry. In the reported experiments, the corpus from which templates were taken contained mostly Imagist poetry from the Project Gutenberg.[2]

## Examples

Following is an excerpt of the previously introduced BBC news story which we used for generating poems.

> **Justin Bieber on Miami drink-drive charge after 'road racing'**
>
> Pop star Justin Bieber has appeared before a Miami court accused of driving under the influence of alcohol, marijuana and prescription drugs. Police said the Canadian was arrested early on Thursday after racing his sports car on a Miami Beach street. They said he did not co-operate when pulled over and also charged him with resisting arrest without violence and having an expired driving licence. (...)

The article then goes on to discuss the issue in more detail and to give an account of the behaviour of Justin Bieber.

We use Wikipedia as the background corpus, as it is large, represents many areas of life, and is freely available. Contrasting the Justin Bieber story to the contents of Wikipedia, using the model described in the previous section, we obtain a list of word pairs ranked by how specific they are to the news story (Table 2). Pairs with lower scores tend to be quite common associations (e.g. los angeles, sports car, street car, etc). Pairs with top scores seem to capture the essence of the news story well. Clearly the associations suggest that the news story has something to do with Bieber, police, Miami and alcohol (and "saying" something, which is not typical in Wikipedia, our background corpus, but is typical in news stories like this one).

Using words in the top associations, the following sample poem was generated:

*Race at the miami-dade justins in the marijuana!*

[2] http://www.gutenberg.org

| Top pairs | Bottom pairs |
|---|---|
| say, bieber | los, angeles |
| say, police | later, jail |
| miami, bieber | sport, car |
| miami, say | car, early |
| bieber, police | thursday, early |
| beach, bieber | marijuana, alcohol |
| beach, police | prescription, alcohol |
| car, say | sport, thursday |
| bieber, alcohol | car, street |
| bieber, los | prescription, marijuana |

Table 2: The top and the bottom foreground associations for the Justin Bieber's news story.

*The officer is taller than you, who race yourself*
*So miami-dade and miami-dade: race how its entourages are said*
*Co-operate and later in the singer, like a angeles of alcohols*
*Racing with jails and singers and co-operate race.*

This poem was one of the many we generated and, in a humorous way, it covers many different aspects of the news story. (Currently, our implementation does not fix capitalization and articles in the results, nor does it recognize compound words. These are left for future work; here we present results in the exact form produced by the implementation without editing them manually in any way.)

In order to illustrate the effect of using document specific associations, we next fix the template used for word substitution and two types of poems: 1) using words related to Justin Bieber in general, using Wikipedia as the background corpus (Toivanen et al. 2012), and 2) and using document specific words from the news story given above.

These poems are generated using words related Justin Bieber:

Is it the youtube, the justin,
the release of second times,
and the second celebrities of our says?
These are but brauns.

Is it the atlanta, the mallette,
the music of first uniteds,
and the song yorks of our defs?
These are but news.

Is it the chart, the braun,
the def of first ushers,
and the musical stratfords of our nevers?
These are but youtubes.

The following three poems have been produced using document specific associations:

Is it the miami, the street,
the jail of co-operate officers,
and the co-operate singers of our prescriptions?
These are but alcohols.

Is it the car, the sport,
the angeles of co-operate justins,
and the early lamborghinis of our entourages?
These are but singers.

Is it the entourage, the sport,
the singer of later lamborghinis,
and the early thursdays of our singers?
These are but justins.

Finally, instead of evaluating the methods with test subjects, we let the readers of this paper decide for themselves by providing a collection of 18 poems at the end of this paper. To make this reader evaluation as fair as possible, *we did not select or edit the poems in any way*. We selected three news stories, of different topics and of sufficiently general interest, based on their original contents but not on the poems produced. Then, without any testing of the suitability of those stories for association extraction and poetry generation, we ran the poetry machinery and added the first poems produced for each of the news stories in the collection at the end of this paper.

The three news stories are the following:

- The aforementioned news story about Justin Bieber.
- A news story *Ukrainian Prime Minister Resigns as Parliament Repeals Restrictive Laws*[3] published by NY Times on January 28.
- A news story *The return of the firing squad? US states reconsider execution methods*[4] published by The Guardian on January 28.

To get some understanding how different background corpora affect the results, we used two different background corpora: the English Wikipedia and the Project Gutenberg corpus. We used each background to generate three poems from each news story: in each collection of six poems, poems 1–3 are generated by using Wikipedia as background, and poems 4–6 using Project Gutenberg as background.

## Conclusions and Future Work

In this paper we have proposed a novel approach for using document-specific word associations to provide content words in a poetry generation task. As a novel part of the methodology, we use a recent model that extracts word pairs that are specific to a given document in a statistical sense.

Instead of an objective evaluation with some fixed criteria, we invite the readers of this paper to read the poems generated by the system — called P.O. Eticus — in the next pages and form their own opinions on the methods and results.

Automated methods for poetry generation from given documents could have practical application areas. For instance, the methodology has already been used in an art project exhibited in Estonia and Finland (Gross et al. 2014). Similarly the poems could be used for entertainment or as

[3] http://nyti.ms/1k0kj9r
[4] http://gu.com/p/3m8p5

automatically generated thought-provoking mechanisms in news websites or internet forums.

An interesting direction for further developments would be combining together documents on the same topic and then producing poems which give an overview of the *diverse* aspects of the topic. For instance each verse could cover some specific documents, or a step further we could use document clustering for identifying key subtopics and creating verses from these.

## Acknowledgments

This work has been supported by the Future and Emerging Technologies (FET) programme within the Seventh Framework Programme for Research of the European Commission, under FET grant number 611733 (ConCreTe), the Algorithmic Data Analysis (Algodan) Centre of Excellence of the Academy of Finland, and the Helsinki Doctoral Program in Computer Science (HECSE).

## References

Blei, D. M.; Ng, A. Y.; and Jordan, M. I. 2003. Latent Dirichlet allocation. *Journal of Machine Learning Research* 3:993–1022.

Bouma, G. 2009. Normalized (pointwise) mutual information in collocation extraction. In *Proceedings of the Biennial GSCL Conference*, 31–40.

Church, K. W., and Hanks, P. 1990. Word association norms, mutual information, and lexicography. *Computational Linguistics* 16(1):22–29.

Colton, S.; Goodwin, J.; and Veale, T. 2012. Full FACE poetry generation. In *Proceedings of the Third International Conference on Computational Creativity*, 95–102.

Deerwester, S.; Dumais, S. T.; Furnas, G. W.; Landauer, T. K.; and Harshman, R. 1990. Indexing by latent semantic analysis. *Journal of the American Society for Information Science* 41(6):391–407.

Diaz-Agudo, B.; Gervás, P.; and González-Calero, P. A. 2002. Poetry generation in COLIBRI. In *ECCBR 2002, Advances in Case Based Reasoning*, 73–102.

Dunning, T. 1993. Accurate methods for the statistics of surprise and coincidence. *Computational linguistics* 19(1):61–74.

Gervás, P. 2001. An expert system for the composition of formal spanish poetry. *Journal of Knowledge-Based Systems* 14(3–4):181–188.

Gross, O.; Toivonen, H.; Toivanen, J. M.; and Valitutti, A. 2012. Lexical creativity from word associations. In *Knowledge, Information and Creativity Support Systems (KICSS), 2012 Seventh International Conference on*, 35 –42.

Gross, O.; Toivanen, J.; Lääne, S.; and Toivonen, H. 2014. Arts, news, and poetry - the art of framing. Submitted for review.

Gross, O.; Doucet, A.; and Toivonen, H. 2014. Document summarization based on word associations. In *Proceedings of the 37th international ACM SIGIR conference on Research and Development in Information Retrieval*. ACM.

Huovelin, J.; Gross, O.; Solin, O.; Lindn, K.; Maisala, S.; Oittinen, T.; Toivonen, H.; Niemi, J.; and Silfverberg, M. 2013. Software newsroom - an approach to automation of news search and editing. *Journal of Print and Media Technology Research* 3(2013):3:141–156.

Manurung, H. M.; Ritchie, G.; and Thompson, H. 2000. Towards a computational model of poetry generation. In *Proceedings of AISB Symposium on Creative and Cultural Aspects and Applications of AI and Cognitive Science*, 79–86.

Manurung, H. 2003. *An evolutionary algorithm approach to poetry generation*. Ph.D. Dissertation, University of Edinburgh, Edinburgh, United Kingdom.

Netzer, Y.; Gabay, D.; Goldberg, Y.; and Elhadad, M. 2009. Gaiku : Generating haiku with word associations norms. In *Proceedings of NAACL Workshop on Computational Approaches to Linguistic Creativity*, 32–39.

Toivanen, J. M.; Toivonen, H.; Valitutti, A.; and Gross, O. 2012. Corpus-based generation of content and form in poetry. In *International Conference on Computational Creativity*, 175–179.

Toivanen, J. M.; Järvisalo, M.; and Toivonen, H. 2013. Harnessing constraint programming for poetry composition. In *International Conference on Computational Creativity*, 160–167.

Toivonen, H.; Gross, O.; Toivanen, J.; and Valitutti, A. 2013. On creative uses of word associations. In *Synergies of Soft Computing and Statistics for Intelligent Data Analysis, Part 1*, number 190 in Advances in Intelligent Systems and Computing. Springer. 17–24.

Wong, M. T., and Chun, A. H. W. 2008. Automatic haiku generation using VSM. In *Proceedings of ACACOS*, 318–323.

## Justin Bieber on Miami drink-drive charge after 'road racing'

Poems by P.O.Eticus

1. *It races at the singer, the later, racing singer, and he is race*
*within its officer and prescription. Inside is his thursday, his street, his sport, his*
*lamborghini, and his entourages. He is racing, and the entourages are said with singers*
*of miami, racing through miami-dade miami-dade. A miami says itself up at*
*the early entourage, and through the miami-dade miami in the car he can say*
*miami lamborghini, lazily racing among co-operate singers. A lamborghini in a*
*early cars and angeleses, and members race into his car, raced,*
*thursday, saying up like angeleses of member, higher and higher. Justin! The*
*members say on their later says. The thursday races up in early*
*later miamis of co-operate marijuana and says into the court. Car! And*
*there is only the car, the car, the beach, and the racing thursday.*

2. *Fruit can not race*
*through this co-operate beach:*
*car can not race into sport*
*that angeleses up and races*
*the angeleses of sports*
*and biebers the singers.*

3. *There is a miami-dade here*
*within my miami,*
*but miami-dade and sport....*

4. *I say;*
*perhaps I have steped;*
*this is a driving;*
*this is a incident;*
*and there is home....*

5. *Oh, he was bieber*
*Which then was he among the ferrari?*
*The co-operate, the slow, the medication?*
*I have transfered a first raymond of thursdays in one*
*But not this, this sport*
*Car!*

6. *Make, You! and canadian my driver;*
*That my ferraris race me no longer,*
*But thursday in your home.*

## Ukrainian Prime Minister Resigns as Parliament Repeals Restrictive Laws

Poems by P.O.Eticus

1. *Water approved and restrictive by repealing building*
   *Which laws and governments it into sundayukraine police weeks*
   *Said with provincial opposition vote.*
   *The repealing of the leader upon the statement*
   *Is like a leader of week oppositions*
   *In a concrete statement new resignation.*

2. *The statement approves into the party, and the party says him in a leader of leader. But it is said with parliament and restrictive with sundayukraine streets. The week parliaments. Repealing, repealing, saying, repeal, resigning, resign the leaders. Over riots, and televisions, and votes, and streets. Approving its region on the vote the government legislations, blocks itself through the leaders, and ministers and repeals along the riots.*

3. *The svobodas*
   *police from the resigns,*
   *the televisions at their statements*
   *resign lower through the ukraines.*

4. *And always concrete! Oh, if I could ride*
   *With my week resigned concrete against the repeal*
   *Do you resign I'd have a parliament like you at my television*
   *With your azarov and your week that you resign me? O ukrainian week,*
   *How I resign you for your parliamentary legislation!*

5. *Concrete one,*
   *new and restrictive,*
   *provincial repeal,*
   *region,*
   *concrete and leader you are vote*
   *in our weeks.*

6. *Resigned amid jan*
   *We will avoid all azarov;*
   *And in the government*
   *Resigning forth, we will resign restrictive votes*
   *Over the repealed administration of azarov.*

## The return of the firing squad? US states reconsider execution methods

Poems by P.O.Eticus

1. *Many one,*
   *many and lethal,*
   *recent injection,*
   *republican,*
   *recent and drug you are gas*
   *in our electrocutions.*

2. *You are not he.*
   *Who are you, choosing in his justice on the question*
   *And lethal and lethal to me?*
   *His doubt, though he rebuilt or found*
   *Was always lethal and recent*
   *And many to me.*

3. *I die;*
   *perhaps I have began;*
   *this is a doubt;*
   *this is a prisoner;*
   *and there is state....*

4. *You amid the public's pentobarbital longer,*
   *You trying in the josephs of the methods above,*
   *Me, your hanging on the michael, unusual franklins,*
   *Me unusual michael in the states, ending you use*
   *You, your court like a death, proposed, pentobarbital,*
   *You, with your death all last, like the wyoming on a ended!*

5. *Lawmaker and quiet:*
   *a brattin overdoses in the year courts*
   *behind the process with the many new injection*
   *across the brattin.*

6. *The longer rebuilds into the day, and the gas ends him in a supply of schaefer. But it is divulged with west and powerful with republican penalties. The process options. Coming, rebuilding, divulging, charles, looming, propose the news. Over officials, and spectacles, and senators, and burns. Begining its florida on the dodd the supply spectacles, franklins itself through the propofols, and burns and proposes along the gases.*

TIETOJENKÄSITTELYTIETEEN LAITOS
PL 68 (Gustaf Hällströmin katu 2 b)
00014 Helsingin yliopisto

DEPARTMENT OF COMPUTER SCIENCE
P.O. Box 68 (Gustaf Hällströmin katu 2 b)
FI-00014 University of Helsinki, FINLAND

JULKAISUSARJA **A**

SERIES OF PUBLICATIONS **A**

Reports are available on the e-thesis site of the University of Helsinki.

A-2010-1 M. Lukk: Construction of a global map of human gene expression - the process, tools and analysis. 120 pp. (Ph.D. Thesis)

A-2010-2 W. Hämäläinen: Efficient search for statistically significant dependency rules in binary data. 163 pp. (Ph.D. Thesis)

A-2010-3 J. Kollin: Computational Methods for Detecting Large-Scale Chromosome Rearrangements in SNP Data. 197 pp. (Ph.D. Thesis)

A-2010-4 E. Pitkänen: Computational Methods for Reconstruction and Analysis of Genome-Scale Metabolic Networks. 115+88 pp. (Ph.D. Thesis)

A-2010-5 A. Lukyanenko: Multi-User Resource-Sharing Problem for the Internet. 168 pp. (Ph.D. Thesis)

A-2010-6 L. Daniel: Cross-layer Assisted TCP Algorithms for Vertical Handoff. 84+72 pp. (Ph.D. Thesis)

A-2011-1 A. Tripathi: Data Fusion and Matching by Maximizing Statistical Dependencies. 89+109 pp. (Ph.D. Thesis)

A-2011-2 E. Junttila: Patterns in Permuted Binary Matrices. 155 pp. (Ph.D. Thesis)

A-2011-3 P. Hintsanen: Simulation and Graph Mining Tools for Improving Gene Mapping Efficiency. 136 pp. (Ph.D. Thesis)

A-2011-4 M. Ikonen: Lean Thinking in Software Development: Impacts of Kanban on Projects. 104+90 pp. (Ph.D. Thesis)

A-2012-1 P. Parviainen: Algorithms for Exact Structure Discovery in Bayesian Networks. 132 pp. (Ph.D. Thesis)

A-2012-2 J. Wessman: Mixture Model Clustering in the Analysis of Complex Diseases. 118 pp. (Ph.D. Thesis)

A-2012-3 P. Pöyhönen: Access Selection Methods in Cooperative Multi-operator Environments to Improve End-user and Operator Satisfaction. 211 pp. (Ph.D. Thesis)

A-2012-4 S. Ruohomaa: The Effect of Reputation on Trust Decisions in Inter-enterprise Collaborations. 214+44 pp. (Ph.D. Thesis)

A-2012-5 J. Sirén: Compressed Full-Text Indexes for Highly Repetitive Collections. 97+63 pp. (Ph.D. Thesis)

A-2012-6 F. Zhou: Methods for Network Abstraction. 48+71 pp. (Ph.D. Thesis)

A-2012-7 N. Välimäki: Applications of Compressed Data Structures on Sequences and Structured Data. 73+94 pp. (Ph.D. Thesis)

A-2012-8 S. Varjonen: Secure Connectivity With Persistent Identities. 139 pp. (Ph.D. Thesis)

A-2012-9 M. Heinonen: Computational Methods for Small Molecules. 110+68 pp. (Ph.D. Thesis)

A-2013-1 M. Timonen: Term Weighting in Short Documents for Document Categorization, Keyword Extraction and Query Expansion. 53+62 pp. (Ph.D. Thesis)

A-2013-2 H. Wettig: Probabilistic, Information-Theoretic Models for Etymological Alignment. 130+62 pp. (Ph.D. Thesis)

A-2013-3 T. Ruokolainen: A Model-Driven Approach to Service Ecosystem Engineering. 232 pp. (Ph.D. Thesis)

A-2013-4 A. Hyttinen: Discovering Causal Relations in the Presence of Latent Confounders. 107+138 pp. (Ph.D. Thesis)

A-2013-5 S. Eloranta: Dynamic Aspects of Knowledge Bases. 123 pp. (Ph.D. Thesis)

A-2013-6 M. Apiola: Creativity-Supporting Learning Environments: Two Case Studies on Teaching Programming. 62+83 pp. (Ph.D. Thesis)

A-2013-7 T. Polishchuk: Enabling Multipath and Multicast Data Transmission in Legacy and Future Interenet. 72+51 pp. (Ph.D. Thesis)

A-2013-8 P. Luosto: Normalized Maximum Likelihood Methods for Clustering and Density Estimation. 67+67 pp. (Ph.D. Thesis)

A-2013-9 L. Eronen: Computational Methods for Augmenting Association-based Gene Mapping. 84+93 pp. (Ph.D. Thesis)

A-2013-10 D. Entner: Causal Structure Learning and Effect Identification in Linear Non-Gaussian Models and Beyond. 79+113 pp. (Ph.D. Thesis)

A-2013-11 E. Galbrun: Methods for Redescription Mining. 72+77 pp. (Ph.D. Thesis)

A-2013-12 M. Pervilä: Data Center Energy Retrofits. 52+46 pp. (Ph.D. Thesis)

A-2013-13 P. Pohjalainen: Self-Organizing Software Architectures. 114+71 pp. (Ph.D. Thesis)

A-2014-1 J. Korhonen: Graph and Hypergraph Decompositions for Exact Algorithms. 62+66 pp. (Ph.D. Thesis)

A-2014-2 J. Paalasmaa: Monitoring Sleep with Force Sensor Measurement. 59+47 pp. (Ph.D. Thesis)

A-2014-3 L. Langohr: Methods for Finding Interesting Nodes in Weighted Graphs. 70+54 pp. (Ph.D. Thesis)

A-2014-4 S. Bhattacharya: Continuous Context Inference on Mobile Platforms. 94+67 pp. (Ph.D. Thesis)

A-2014-5 E. Lagerspetz: Collaborative Mobile Energy Awareness. 60+46 pp. (Ph.D. Thesis)

A-2015-1 L. Wang: Content, Topology and Cooperation in In-network Caching. 190 pp. (Ph.D. Thesis)

A-2015-2 T. Niinimäki: Approximation Strategies for Structure Learning in Bayesian Networks. 64+93 pp. (Ph.D. Thesis)

A-2015-3 D. Kempa: Efficient Construction of Fundamental Data Structures in Large-Scale Text Indexing. 68+88 pp. (Ph.D. Thesis)

A-2015-4 K. Zhao: Understanding Urban Human Mobility for Network Applications. 62+46 pp. (Ph.D. Thesis)

A-2015-5 A. Laaksonen: Algorithms for Melody Search and Transcription. 36+54 pp. (Ph.D. Thesis)

A-2015-6 Y. Ding: Collaborative Traffic Offloading for Mobile Systems. 223 pp. (Ph.D. Thesis)

A-2015-7 F. Fagerholm: Software Developer Experience: Case Studies in Lean-Agile and Open Source Environments. 118+68 pp. (Ph.D. Thesis)

A-2016-1 T. Ahonen: Cover Song Identification using Compression-based Distance Measures. 122+25 pp. (Ph.D. Thesis)