# Event representation across genre

Pivovarova, Lidia

The Association for Computational Linguistics
2013

Pivovarova , L , Huttunen , S & Yangarber , R 2013 , Event representation across genre . in Proceedings of the The 1st Workshop on EVENTS : Definition, Detection, Coreference, and Representation . The Association for Computational Linguistics , Atlanta , pp. 29-37 , NAACL Workshop on EVENTS: Definition, Detection, Coreference, and Representation , Atlanta, GA , United States , 01/01/1800 .

http://hdl.handle.net/10138/42892

publishedVersion

# Event representation across genre

**Lidia Pivovarova, Silja Huttunen and Roman Yangarber**

University of Helsinki
Finland

## Abstract

This paper describes an approach for investigating the representation of events and their distribution in a corpus. We collect and analyze statistics about subject-verb-object triplets and their content, which helps us compare corpora belonging to the same domain but to different genre/text type. We argue that event structure is strongly related to the genre of the corpus, and propose statistical properties that are able to capture these genre differences. The results obtained can be used for the improvement of Information Extraction.

## 1 Introduction

The focus of this paper is collecting data about certain characteristics of events found in text, in order to improve the performance of an Information Extraction (IE) system. IE is a technology used for locating and extracting specific pieces of information—or "facts"—from unstructured natural-language text, by transforming the facts into abstract, structured objects, called *events*.

In IE we assume that events represent real-world facts and the main objective is to extract them from plain text; the nature of the events themselves rarely receives in-depth attention in current research.

Events may have various relationships to real-world facts, and different sources may have contradictory views on the facts, (Saurí and Pustejovsky, 2012). Similarly to many other linguistic units, an event is a combination of meaning and form; the structure and content of an event is influenced by both the structure of the corresponding real-world fact and by the properties of the surrounding text.

We use the notion of *scenario* to denote a set of structured events of interest in a real-world domain: e.g., the MUC Management Succession scenario, (Grishman and Sundheim, 1996), within the broader Business domain.

The representation and the structure of events in text depends on the scenario. For example, Huttunen et al. (2002a; Huttunen et al. (2002b) points out that "classic" MUC scenarios, such as Management Succession or Terrorist Attacks, describe events that occur in a specific point in time, whereas other scenarios like Natural Disaster or Disease Outbreak describe processes that are spread out across time and space. As a consequence, events in the latter, "nature"-related scenarios are more complex, may have a hierarchical structure, and may overlap with each other in text. Linguistic cues that have been proposed in Huttunen et al. (2002a) to identify the overlapping or partial events include specific lexical items, locative and temporal expressions, and usage of ellipsis and anaphora.

Grishman (2012) has emphasized that for fully unsupervised event extraction, extensive linguistic analysis is essential; such analysis should be able to capture "modifiers on entities, including quantity and measure phrases and locatives; modifiers on predicates, including negation, aspect, quantity, and temporal information; and higher-order predicates, including sequence and causal relations and verbs of belief and reporting." It is clear that such sophisticated linguistic analysis increases the importance of text style and genre for Information Extraction.

The idea of statistical comparison between text types goes back at least as far as (Biber, 1991). It was subsequently used in a number of papers on automatic text categorization (Kessler et al., 1997; Stamatatos et al., 2000; Petrenz and Webber, 2011).

Szarvas et al. (2012) studied the linguistic cues indicating uncertainty of events in three genres: news, scientific papers and Wikipedia articles. They demonstrate significant differences in lexical usage across the genres; for example, such words as *fear* or *worry* may appear relatively often in news and Wikipedia, but almost never in scientific text. They also investigate differences in syntactic cues: for example, the relation between a proposition and a real-word fact is more likely to be expressed in the passive voice in scientific papers (*it is expected*), whereas in news the same words are more likely appear in the active.

Because events are not only representations of facts but also linguistic units, an investigation of events should take into account the particular language, genre, scenario and medium of the text—i.e., events should be studied in the context of a *particular corpus*. Hence, the next question is how corpus-driven study of events should be organized in practice, or, more concretely, what particular statistics are needed to capture the scenario-specific characteristics of event representation in a particular corpus, and what kind of markup is necessary to solve this task. We believe that answers to these questions will likely depend on the ultimate goals of event detection. We investigate IE in the business domain—thus, we believe that preliminary study of the corpus should use exactly the same depth of linguistic analysis as would be later utilized by the IE system.

## 2 Problem Statement

### 2.1 Events in the Business domain

We investigate event structure in the context of PULS,[1] an IE System, that discovers, aggregates, verifies and visualizes events in various scenarios. This paper focuses on the Business domain, in which scenarios include investments, contracts, layoffs and other business-related events, which are collected in a database to be used for decision support. In the Business domain, PULS currently handles two types of documents: news reports and short summaries written by professional business analysts. Thus, events extracted from both corpora relate to approximately the same real-world facts.

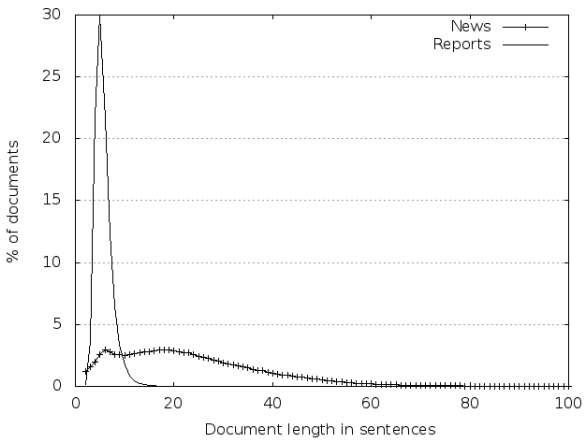


Figure 1: Distributions of document length in the news and business analysts' reports corpora

Both corpora are in English (though some of the analysts' reports are based on news articles written in other languages). We collected a corpus of reports containing 740 thousand documents over three years 2010-2012, and a news corpus containing 240 thousand documents over the same period.

The two corpora demonstrate significant linguistic differences. First, the documents have different length: the average length of an analyst reports is 5.5 sentences including the title, and 80% of the documents have length between 4 and 7 sentences, (see Figure 1). News articles are on average 19 sentences long—and much more varied in length.

The topical structure is also quite different for the two corpora. Each analyst report is most typically dedicated to a particular single real-world event. Also, the reports tend to have a standardized, formulaic structure. The analysts who generate these reports tend to follow a specific, strict style and structure over time.

By contrast, documents in the news corpus are much more heterogeneous. These texts can follow a wide variety of different styles—short messages, surveys, interviews, etc. News documents can focus not only strictly on business events but on related topics as well. For example, political events have complex interaction with and impact on business ac-

[1] More information is available at: http://puls.cs.helsinki.fi

tivity, and therefore political news frequently appear in business news feeds.

PULS aims to use the same processing chain for various types of input documents. One key goal of the current work is to investigate whether different IE processing approaches are needed for documents belonging to different text types, as exemplified by analyst reports vs. articles from news feeds.

To summarize, the goals of the present work are:

- investigate how text genre influences event representation;
- find formal markers able to capture and measure the differences in corpus style/genre;
- propose a methodology for adaptating an IE system to a different text genre.

### 2.2 System Description

In this section we describe how the IE system is used in a "pattern-mining mode," to address the aforementioned problems.

PULS is a pipeline of components, including: a shallow parser/chunker; domain ontologies and lexicons; low-level patterns for capturing domain-specific entities and other semantic units, such as dates and currency expressions; higher-level patterns for capturing domain-specific relations and events; inference rules, which combine fragments of an event that may be scattered in text—that a pattern may not have picked up in the immediate context (e.g., the date of the event); reference resolution for merging co-referring entities and events.

The ontology and the lexicon for the Business domain encode the taxonomic relations and support merging of synonyms: e.g., the ontology stores the information that *cellphone* and *mobile phone* are synonymous, and that a super-concept for both is PRODUCT.

Low-level patterns are used to extract entities from text, such as company names, dates, and locations. On a slightly higher level, there are patterns that match contexts such as *range (collection, line, etc.) of X* and assign them the type of *X*. For instance, the phrase *a collection of watches* would be assigned semantic type *watch*, etc. The top-level patterns in all IE scenarios are responsible for finding the target events in text.

In the pattern-mining mode we use the general pattern SUBJECT–VERB–OBJECT, where the components may have any semantic type and are constrained only by their *deep* syntactic function—the system attempts to normalize many syntactic variants of the basic, active form: including passive clauses, relative clauses, etc.[2]

The idea of using very simple, local patterns for obtaining information from large corpora in the context of event extraction is similar to work reported previously, e.g., the bootstrapping approaches in (Thelen and Riloff, 2002; Yangarber et al., 2000; Riloff and Shepherd, 1997). Here, we do not use iterative learning, and focus instead on collecting and analyzing interesting statistics from a large number of S-V-O patterns. We collected all such "generalized" S-V-O triplets from the corpus and stored them in a database. In addition to the noun groups, we save the head nouns and their semantic classes. This makes it easy to use simple SQL queries to count instances of a particular pattern, e.g., all objects of a particular verb, or all actions that can be applied to an object of semantic class "PRODUCT." For each triplet the database stores a pointer the original sentence, making it possible to analyze specific examples in their context.

In the next two sections we present the statistics that we collected using the pattern-mining mode. This information reflects significant differences among the corpora genres and can be used to *measure* variety of genre. We believe that in the future such data analysis will support the adaptation of PULS to new text genres.

## 3 Statistical Properties of the Corpora

### 3.1 Personal pronouns

Pronouns play a key role in anaphoric relations; the more pronouns are present in the corpus, the more crucial anaphora resolution becomes. Analysis of relationships between frequencies of personal pronouns in text and the genre of the text is not new; it has been observed and studied extensively, going

[2]By normalization of syntactic variants we mean, for instance, that clauses like *"Nokia releases a new cellphone"* (active), *"a new cellphone is released by Nokia"* (passive), *"a new cellphone, released by Nokia,..."* (relative), etc., are all reduced to the same S-V-O form.

| Pronoun | Reports | | News | |
|---|---|---|---|---|
| | *Object* | *Subject* | *Object* | *Subject* |
| I/me | 0.003 | 0.007 | 0.2 | 1.0 |
| we/us | 0.001 | 0.001 | 0.4 | 1.7 |
| you | 0.002 | 0.003 | 0.3 | 0.8 |
| he/him | 0.05 | 0.4 | 0.6 | 2.2 |
| she/her | 0.007 | 0.05 | 0.1 | 0.5 |
| they/them | 0.3 | 0.6 | 0.8 | 1.3 |
| it | 1.1 | 2.6 | 1.5 | 2.3 |
| **Total** | **1.5** | **3.6** | **4.0** | **9.8** |

Table 1: Personal pronouns appearing in the subject or object position in the corpora. The numerical values are proportions of the total number of verbs.

back as far as, e.g., (Karlgren and Cutting, 1994). The analysis of pronoun distribution in our corpora is presented in Table 1, which shows the proportions of personal pronouns, as they appear in subject or object position with verbs in the collected triples. The numbers are relative to the count of all verb tokens in the corpus, i.e., the total number of the S–V–O triplets extracted from the corpus. The total number of triplets is approximately 5.7M in the report corpus and 11M in the news corpus.

It can be seen from Table 1 that personal pronouns are much more rare in the report corpus than in the news corpus. Only 1.5% of verbs in the reports corpus have a pronoun as an object, and 3.6% as a subject. By contrast, in the news corpus 4% of verbs have a personal pronoun as an object, and 9.8% as a subject. This corresponds to the observation in (Szarvas et al., 2012), that "impersonal constructions are hardly used in news media."

It is interesting to note the distribution of the particular pronouns in the two corpora. Table 1 shows that *it* is the most frequent pronoun, *they* and *he* are less frequent; the remaining pronouns are much less frequent in the report corpus, whereas in the news the remaining personal pronouns have a much more even distribution. This clearly reflects a more relaxed style of the news that may use rhetorical devices more freely, including citing direct speech and use a direct addressing the reader (*you*). It is also interesting to note that in the third-person singular, the feminine pronoun is starkly more rare in both corpora than the masculine, but roughly twice more rare among the analyst reports.

| | Reports | | News | |
|---|---|---|---|---|
| | *Subject* | *Object* | *Subject* | *Object* |
| *All* | 21.8 | 6.6 | 14.6 | 6.5 |
| *Business* | 27.1 | 8.1 | 20.1 | 9.5 |

Table 2: Distribution of proper names as subjects and objects, as a proportion the total number of all verbs (top row) vs. *business-related* verbs (bottom row).

### 3.2 Proper Names

Proper names play a crucial role in co-reference resolution, by designating anaphoric relations in text, similarly to pronouns. In the Business domain, e.g., a common noun phrase (NP) may co-refer with a proper name, as "the company" may refer to the name of a particular firm. A correctly extracted event can be much less useful for the end-user if it does not contain the specific name of the company involved in the event.

A verbs is often the key element of a pattern that indicates to the IE system the presence of an event of interest in the text. When the subject or object of the verb is a common NP, the corresponding proper name must be found in the surrounding context, using reference resolution or domain-specific inference rules. Since reference resolution is itself a phase that contributes some amount of error to the overall IE process, it is natural to expect that if proper-name subjects and objects are more frequent in the corpus, then the analysis can be more precise, since all necessary information can be extracted by pattern without the need for additional extra inference. Huttunen et al. (2012) suggests that the compactness of the event representation may be used as one of the discourse cues that determine the event relevance.

Table 2 shows the percentage of proper name objects and subjects for the two corpora. Proper-name objects have comparable frequency in both corpora, though proper-name subjects appear much more frequently in analyst reports than in news. Furthermore, for the *business verbs*, introduced below in section 4.1—i.e., the specific set of verbs that are used in event patterns in the Business scenarios—as seen in the second row of the table—proper-name objects and subjects are more frequent still. This suggests that business events *tend* to mention proper names.

| | Percentage of business verbs | | |
|---|---|---|---|
| *Corpus* | *Total* | *Title* | *1st sentence* |
| **Reports** | 49.5 | 7.6 | 13.8 |
| **News** | 31.8 | 0.6 | 1.1 |

Table 3: Business verbs in analyst reports and news corpora, as a proportion of the total number of verbs.

## 4 Business Verbs

### 4.1 Distribution of Business verbs

The set of *business-related verbs* is an important part of the system's domain-specific lexicon for the Business domain. These verbs are quite diverse: some are strongly associated with the Business domain, e.g., *invest*; some are more general, e.g., *pay, make*; many are ambiguous, e.g., *launch, fire*. Inside analyst reports these verbs always function as markers of certain business events or relations. The verbs are the key elements of the top-level patterns and it is especially crucial to investigate their usage in the corpora to understand how the pattern base should be fine-tune for the task.

Since the majority of these verbs fall in the ambiguous category, none of these verbs can by themselves serve a sufficient indicator of the document's topic. Even the more clear-cut business verbs, such as *invest*, can be used metaphorically in the non-business context. However, their *distribution* in the particular document and in the corpus as a whole can reflect the genre specificity of the corpus.

Table 3 shows the overall proportion of the business verbs, and their proportion in titles and in the first sentence of a documents. It suggests that almost 50% of the verbs in the report corpus are "business" verbs, and almost half of them are concentrated in the beginning of a document. By contrast, the fraction of business verbs in the news corpus is less than one third and they are more scattered through the text. This fact is illustrated by the plot in Figure 2.

The first sentence is often the most informative part of text, since it introduces the topic of the document to the reader and the writer must do his/her best to attract the reader's attention. It was shown in (Huttunen et al., 2012) that 65% of highly-relevant events in the domain of medical epidemics appear in the title or in the first two sentences of a news article; Lin and Hovy (1997) demonstrated that

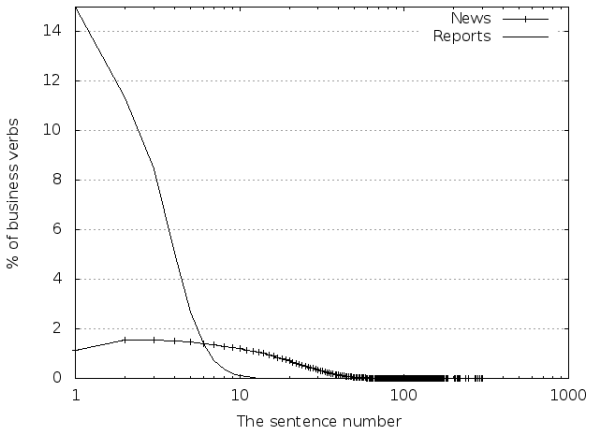


Figure 2: Percentage of business verbs in the text; sentence 0 refers to the title of the document. The fraction of verbs is presented as a percent of all verb instances in the corpus. Logarithmic scale is used for the x axis.

about 50% of topical keywords are concentrated in the titles. We have noticed that some documents in the news corpus have relevance to the business scenario, although relevant events still can be extracted from the second or third paragraphs of the text, mentioned incidentally. By contrast, each analyst report is devoted to a specific business event, and these events are frequently mentioned as early as in the title.

### 4.2 Case study: is "launch" a business verb?

A set of verbs such as *launch, introduce, release, present*,[3] etc., are used in the Business scenarios to extract events about bringing new products to market. In the domain ontology they are grouped under a concept called *LAUNCH-PRODUCT*. An example of a pattern that uses this concept is following:

```
np(COMPANY) vg(LAUNCH-PRODUCT)
    np(ANYTHING)
```

This pattern matches when a NP designating a company is followed by a verb from the ontology, followed by any other NP. This pattern matches, for example, such sentence as: *The Real Juice Company has launched Pomegranate Blueberry flavour to its line of 100% juices.* However, this pattern also overgenerates by matching sentences such as, e.g.: *Cen-*

[3]Note, the S-V-O triplet extraction also handles phrasal verbs, such as *roll out*, correctly, i.e., identifies them as a single linguistic unit, and treats them the same as single-word verbs.

*tral bank unveils effort to manage household debt.* Even among analyst reports, approximately 14% of the NEW-PRODUCT events found by the system are false positives. It is not feasible to collect a list of all possible products to restrict the semantic type of the object of the verb, since new, unpredictable types of products can appear on the market every day. It seemed more feasible to try to discover all *non-products* that can appear in the object slot, due to the ambiguity of the verbs in patterns—a kind of a black-list. We introduce an ontology concept NON-PRODUCT that groups nouns that can be matched by the LAUNCH verbs but are in fact not products, e.g., *budget, effort, plan, report, study*. The ontology supports multiple inheritance, so any of these words can be attached to other parents as well, if necessary.

If the <PRODUCT> slot in of event is filled by one of the black-listed concepts, the event is also black-listed, and not visible to the end-user. They are used as discourse features by learning algorithms that predict the relevance of other events from the same documents (Huttunen et al., 2012).

The NON-PRODUCT class is populated in an ad-hoc manner over time. The content of such a list depends on the particular corpus; the more diverse the topical and stylistic structure of the corpus, the more time-consuming and the less tractable such development becomes. Thus, an important task is to adjust the patterns and the class of NON-PRODUCT nouns to work for the news corpus, and to develop a feasible methodology to address the false-positive problem. We next show how we can use the pattern-mining mode to address these problems.

We extract all instances of the LAUNCH-PRODUCT verbs appearing in the corpora from the S–V–O database. In total 27.5% of all verb instances in reports corpus are verbs from this semantic class, in comparison to 0.7% in the news corpus. The number of distinct objects are approximately the same in both corpora: 3520 for reports and 3062 for news, see Table 4. In total 247 different objects from the report corpus attached to the semantic class PRODUCT in PULS ontology, and 158 objects have this semantic class in the news corpus.

For 21% of *launch* verbs in the report corpus, and 34% in the news corpus, the system is not able to extract the objects, which may be a consequence of the more diverse and varied language of news. Recall,

| Corpus | LAUNCH-PRODUCT | distinct objects | PRODUCT objects |
|---|---|---|---|
| *Reports* | 204193 | 3520 | 247 |
| *News* | 77463 | 3062 | 158 |

Table 4: Distributions of LAUNCH-PRODUCT verbs in the corpora

that the system extracts a *deep-structure* verbal arguments, i.e., for a sentence like *"A new cell-phone has been launched by company XYZ"* it identifies *cell-phone* as the (deep) object, and the agent *company XYZ* as the (deep) subject.

It is interesting to examine the particular sets of words that can appear in the object position. We collected the 50 most frequent objects of the LAUNCH-PRODUCT verbs for each corpus; they are shown in Table 5 ranked by frequency (we show the top 30 objects to save space). The table shows the semantic class according to our ontology.

Of the 50 most frequent objects, 24 belong to the semantic class PRODUCT in the report corpus, while only 8 objects do in the general news corpus. By contrast, 20 objects belong to the NON-PRODUCT class in the news corpus and only 9 objects in reports. Moreover, 8 objects in the news corpus are not found in the ontology at all, in comparison to only one such case from the report corpus.

Some object classes may mean that the event is still relevant for the business domain, though it does not belong to the NEW-PRODUCT scenario. For example, when object is an advertising campaign the event is likely to belong to the MARKETING scenario, when the object is a facility (*factory, outlet, etc.*) it is likely INVESTMENT. Inference rules may detect such dependencies and adjust the scenario of these events in the Business domain.

The inference rules are supported by the same domain ontology, but can test domain- and scenario-specific conditions explicitly, and thus can be more accurate than the generic reference resolution mechanism. However, this also means that inference rules are more sensitive to the corpus genre and may not easily transfer from one corpus to another.

In some cases an object type cannot be interpreted as belonging to any reasonable event type, e.g., if it is an ORGANIZATION or PERSON. Such cases may arise due to unusual syntax in the sentence that

| Rank | Reports | | | News | | |
|---|---|---|---|---|---|---|
| | **Object** | **Freq** | **Class** | **Object** | **Freq** | **Class** |
| 1 | *Proper Name unspecified* | 19987 | | *Proper Name unspecified* | 5971 | |
| 2 | product | 7331 | PROD | report | 1078 | NON |
| 3 | service | 6510 | PROD | result | 851 | NON |
| 4 | campaign | 3537 | CAMP | plan | 805 | NON |
| 5 | project | 2870 | PROD | product | 792 | PROD |
| 6 | range | 2536 | COLL | service | 648 | PROD |
| 7 | plan | 2524 | NON | it | 618 | PRON |
| 8 | organization | 2450 | ORG | data | 552 | |
| 9 | system | 2166 | FAC | campaign | 510 | CAMP |
| 10 | line | 1938 | COLL | organization | 495 | ORG |
| 11 | model | 1920 | PROD | statement | 467 | NON |
| 12 | application | 1345 | PROD | *Proper Name person* | 449 | PER |
| 13 | website | 1321 | PROD | program | 439 | |
| 14 | flight | 1315 | PROD | *Proper Name company* | 432 | ORG |
| 15 | *Proper Name company* | 1232 | ORG | information | 411 | NON |
| 16 | brand | 1200 | COLL | detail | 398 | NON |
| 17 | offer | 1187 | NON | investigation | 380 | NON |
| 18 | production | 1112 | NON | website | 373 | PROD |
| 19 | programme | 998 | NON | measure | 368 | NON |
| 20 | store | 993 | PROD | they | 363 | PRON |
| 21 | *currency* | 958 | CUR | he | 358 | PRON |
| 22 | route | 954 | PROD | device | 352 | PROD |
| 23 | drink | 891 | PROD | system | 340 | FAC |
| 24 | solution | 883 | NON | smartphone | 337 | PROD |
| 25 | smartphone | 852 | PROD | attack | 335 | |
| 26 | fragrance | 824 | PROD | figure | 318 | NON |
| 27 | card | 802 | PROD | opportunity | 295 | INV |
| 28 | fund | 801 | PROD | fund | 290 | NON |
| 29 | scheme | 773 | NON | *currency* | 287 | CUR |
| 30 | facility | 756 | FAC | model | 286 | COLL |

Table 5: The most frequent objects of LAUNCH verbs. Class labels: PROD: product, NON: non-product (blacklisted), CAMP: advertising campaign, INV: investment. Domain independent labels: COLL: collective; PRON: pronoun, FAC: facility, ORG: organization, PER: person, CUR: currency,

confuses the shallow parser.

In summary, the results obtained from the S-V-O pattern-mining can be used to improve the performance of IE. First, the most frequent subjects and objects for the business verbs can be added to the ontology; second, inference rules and patterns are adjusted to handle the new concepts and words.

It is very interesting to investigate—and we plan to pursue this in the future—how this can be done fully automatically; the problem is challenging since the semantic classes for these news concepts depend on the domain and task; for example, some objects are of type PRODUCT (e.g., "video"), and others are of type NON-PRODUCT (e.g., "attack", "report", etc.). Certain words can be ambiguous even within a limited domain: e.g., *player* may designate a COMPANY ("a major player on the market"), a PRODUCT (DVD-player, CD-player, etc.), or a person (tennis player, football player, etc.); the last meaning is relevant for the Business domain since sports personalities participate in promotion campaigns, and can launch their own brands. Automating the construction of the knowledge bases is a challenging task.

In practice, we found that the semi-automated approach and the pattern-mining tool can be helpful for analyzing genre-specific event patterns; it provides the advantages of a corpus-based study.

## 5 Conclusion

We have described an approach for collecting useful statistics about event representation and distribution of event arguments in corpora. The approach was easily implemented using pattern-based extraction of S-V-O triplets with PULS; it can be equally efficiently implemented on top of a syntactic parser, or a shallow parser of reasonable quality. An ontology and lexicons are necessary to perform domain-specific analysis. We have discussed how the results of such analysis can be exploited for fine-tuning of a practical IE scenario.

The pattern-mining process collects *deep-structure* S–V–O triplets from the corpus—which are "potential" events. The triplets are stored in a database, to facilitate searching and grouping by words or by semantic class appearing as the arguments of the triplets. This helps us quickly find all realizations of a particular pattern—for example, all semantic classes that appear in the corpus as objects of verbs that have semantic class LAUNCH-PRODUCT. The subsequent analysis of the frequency lists can help improve the performance of the IE system by suggesting refinements to the ontology and the lexicon, as well as patterns and inference rules appropriate for the particular genre of the corpus.

Our current work includes the adaptation of the IE system developed for the analyst reports to the general news corpus devoted to the same topics. We also plan to develop a hybrid methodology, to combine the presented corpus-driven analysis with open-domain techniques for pattern acquisition, (Chambers and Jurafsky, 2011; Huang and Riloff, 2012).

The approach outlined here for analyzing the distributions of features in documents is useful for studying events within the context of a corpus. It demonstrates that event structure depends on the text genre, and that genre differences can be easily captured and measured. By analyzing document statistics and the output of the pattern-mining, we can demonstrate significant differences between the genres of analyst reports and general news, such as: sentence length, distribution of the domain vocabulary in the text, selectional preference in domain-specific verbs, word co-occurrences, usage of pronouns and proper names.

The pattern mining collects other statistical features, beyond those that have been discussed in detail above. For example, it showed that active voice is used in 95% of the cases in the news corpus in comparison to 88% in the analyst report corpus. It is also possible to count and compare the usage of other grammatical cues, such as verb tense, modality, etc. Thus, we should investigate not only lexical and semantic cues, but also broader syntactic preferences and selectional constrains in the corpora.

In further research we plan to study how the formal representation of the genre differences can be used in practice, that is, for obtaining directly measurable improvements in the quality of event extraction. Taking into account the particular genre of the corpora from which documents are drawn will also have implications for the work on performance improvements via cross-document merging and inference, (Ji and Grishman, 2008; Yangarber, 2006).

The frequency-based analysis described in Section 4.2 seems to be effective. Sharpening the results of the analysis as well as putting it to use in practical IE applications will be the subject of further study.

## Acknowledgements

We wish to thank Matthew Pierce and Peter von Etter for their help in implementation of the pattern mining more described in this paper. The work was supported in part by the ALGODAN: Algorithmic Data Analysis Centre of Excellence of the Academy of Finland.

## References

Douglas Biber. 1991. *Variation across speech and writing*. Cambridge University Press.

Nathanael Chambers and Dan Jurafsky. 2011. Template-based information extraction without the templates. In *Proceedings of ACL-HLT*, pages 976–986.

Ralph Grishman and Beth Sundheim. 1996. Message understanding conference-6: A brief history. In *Proceedings of COLING*, volume 96, pages 466–471.

Ralph Grishman. 2012. Structural linguistics and unsupervised information extraction. *Automatic Knowledge Base Construction and Web-scale Knowledge Extraction (AKBC-WEKEX 2012)*, pages 57–61.

Ruihong Huang and Ellen Riloff. 2012. Bootstrapped training of event extraction classifiers. *EACL 2012*, pages 286–295.

Silja Huttunen, Roman Yangarber, and Ralph Grishman. 2002a. Complexity of event structure in IE scenarios. In *Proceedings of the 19th International Conference on Computational Linguistics (COLING 2002)*, Taipei, August.

Silja Huttunen, Roman Yangarber, and Ralph Grishman. 2002b. Diversity of scenarios in information extraction. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC 2002)*, Las Palmas de Gran Canaria, Spain, May.

Silja Huttunen, Arto Vihavainen, Mian Du, and Roman Yangarber. 2012. Predicting relevance of event extraction for the end user. In T. Poibeau et al., editor, *Multi-source, Multilingual Information Extraction and Summarization*, pages 163–177. Springer-Verlag, Berlin.

Heng Ji and Ralph Grishman. 2008. Refining event extraction through cross-document inference. In *Proceedings of ACL-2008: HLT*, pages 254–262, June.

Jussi Karlgren and Douglas Cutting. 1994. Recognizing text genres with simple metrics using discriminant analysis. In *Proceedings of the 15th Conference on Computational Linguistics*, pages 1071–1075, Kyoto, Japan, August.

Brett Kessler, Geoffrey Numberg, and Hinrich Schütze. 1997. Automatic detection of text genre. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics*, pages 32–38. Association for Computational Linguistics.

Chin-Yew Lin and Eduard Hovy. 1997. Identifying topics by position. In *Proceedings of the fifth conference on Applied natural language processing*, pages 283–290. Association for Computational Linguistics.

Philipp Petrenz and Bonnie Webber. 2011. Stable classification of text genres. *Computational Linguistics*, 37(2):385–393.

Ellen Riloff and Jessica Shepherd. 1997. A corpus-based approach for building semantic lexicons. In *Proceedings of the Second Conference on Empirical Methods in Natural Language Processing*, pages 117–124. Association for Computational Linguistics, Somerset, New Jersey.

Roser Saurí and James Pustejovsky. 2012. Are you sure that this happened? Assessing the factuality degree of events in text. *Computational Linguistics*, 38(2):261–299.

Efstathios Stamatatos, Nikos Fakotakis, and George Kokkinakis. 2000. Text genre detection using common word frequencies. In *Proceedings of the 18th conference on Computational linguistics - Volume 2*, COLING '00, pages 808–814, Stroudsburg, PA, USA. Association for Computational Linguistics.

György Szarvas, Veronika Vincze, Richárd Farkas, György Mófra, and Iryna Gurevych. 2012. Cross-genre and cross-domain detection of semantic uncertainty. *Computational Linguistics*, 38(2):335–367.

Mark Thelen and Ellen Riloff. 2002. A bootstrapping method for learning semantic lexicons using extraction pattern contexts. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002)*.

Roman Yangarber, Ralph Grishman, Pasi Tapanainen, and Silja Huttunen. 2000. Automatic acquisition of domain knowledge for information extraction. In *Proceedings of the 18th International Conference on Computational Linguistics (COLING 2000)*, Saarbrücken, Germany, August.

Roman Yangarber. 2006. Verification of facts across document boundaries. In *Proceedings of the International Workshop on Intelligent Information Access (IIIA-2006)*, Helsinki, Finland, August.