Natural Language Engineering (2021), **27**, pp. 575–607
doi:10.1017/S1351324920000236

CAMBRIDGE UNIVERSITY PRESS

ARTICLE

Computational generation of slogans

Khalid Alnajjar* and Hannu Toivonen

Department of Computer Science and HIIT, University of Helsinki, Helsinki 00014, Finland
*Corresponding author. E-mail: khalid.alnajjar@helsinki.fi

(Received 13 May 2019; revised 2 April 2020; accepted 3 April 2020; first published online 3 June 2020)

Abstract

In advertising, slogans are used to enhance the recall of the advertised product by consumers and to distinguish it from others in the market. Creating effective slogans is a resource-consuming task for humans. In this paper, we describe a novel method for automatically generating slogans, given a target concept (e.g., car) and an adjectival property to express (e.g., elegant) as input. Additionally, a key component in our approach is a novel method for generating nominal metaphors, using a metaphor interpretation model, to allow generating metaphorical slogans. The method for generating slogans extracts skeletons from existing slogans. It then fills a skeleton in with suitable words by utilizing multiple linguistic resources (such as a repository of grammatical relations, and semantic and language models) and genetic algorithms to optimize multiple objectives such as semantic relatedness, language correctness, and usage of rhetorical devices. We evaluate the metaphor and slogan generation methods by running crowdsourced surveys. On a five-point Likert scale, we ask online judges to evaluate whether the generated metaphors, along with three other metaphors generated using different methods, highlight the intended property. The slogan generation method is evaluated by asking crowdsourced judges to rate generated slogans from five perspectives: (1) how well is the slogan related to the topic, (2) how correct is the language of the slogan, (3) how metaphoric is the slogan, (4) how catchy, attractive, and memorable is it, and (5) how good is the slogan overall. Similarly, we evaluate existing expert-made slogans. Based on the evaluations, we analyze the method and provide insights regarding existing slogans. The empirical results indicate that our metaphor generation method is capable of producing apt metaphors. Regarding the slogan generator, the results suggest that the method has successfully produced at least one effective slogan for every evaluated input.

Keywords: Natural language generation; Slogan generation; Metaphor generation; Computational creativity

1. Introduction

Slogans are memorable short phrases that express an idea. They are frequently used in advertising and branding to enhance the recall of products by customers and to distinguish it from competitors. For example, the phrase "Connecting People" triggers many of us to think of *Nokia*. This effect of associating a phrase (i.e., slogan) with a concept (i.e., brand) highlights the significance of slogans in advertising. Coming up with successful slogans is challenging, both for humans and machines. This paper proposes a method to draft advertising slogans computationally.

Advertising professionals often resort to rhetorical devices to create memorable and catchy slogans. A common such rhetorical device is metaphor; *Nokia*'s "Connecting People" and *Red Bull*'s "Red Bull gives you wings" are both examples of metaphoric slogans. The subtle metaphor in *Nokia*'s slogan paints an image of mobile devices establishing intimate relations between people, in addition to providing a concrete means of communication between them. *Red Bull*'s slogan is more obviously metaphoric since a drink cannot give wings. An interpretation of the metaphor is

that the drink helps you exceed your physical limits. According to Reinsch (1971), metaphors and similes make messages more persuasive.

We propose a novel method for generation of metaphorical candidate slogans for a given target concept (e.g., *car*) and property (e.g., *elegant*). The intended use of the method is to assist advertising professionals during brainstorming sessions, not to substitute the professionals. Examples of slogans created by the method for the above input include "The Cars Of Vintage," "Travel Free," "The Cars Of Stage," and "Travel, Transport and Trip." Behind each such generated slogan is a computationally generated metaphor, intended to associate the property *elegant* to cars. For instance, the slogan "The Cars Of Stage" establishes an analog between cars and dancers, suggesting that cars are as elegant as dancers.

Our work contributes to the fields of Natural Language Processing (NLP) and Natural Language Generation (NLG) in two ways. On the one hand, this work computationally processes and generates short expressions. On the other hand, the focus is on figurative language, especially generation and evaluation of some figurative devices.

More specifically, the contributions are the following: (1) We start by providing a characterization and review of the field of slogan generation, also including some other creative expressions. (2) Our main contribution is a novel computational method for generation of slogans. (3) A key component of the slogan generation method is an algorithm for generating metaphors, a novel contribution in itself. (4) We evaluate the proposed method, including the metaphor generator, and provide extensive experimental results based on crowdsourcing. A partial description of a preliminary version of the method is given by Alnajjar, Hadaytullah, and Toivonen (2018).

This paper is structured as follows. We begin by covering the necessary conceptual background regarding slogans and metaphor in Section 2, and we review related work on slogan and metaphor generation in Section 3. In Section 4, we describe a novel method for generating metaphorical slogans. We report on empirical evaluation in Section 5 and discuss the results in Section 6. Section 7 contains our conclusions.

2. Background

Slogans. We define a slogan, from an advertising perspective, as a concise, advertisable, and autonomous phrase that expresses a concept (e.g., an idea, product, or entity); the phrase will be frequently repeated and associated with the concept. Elements of advertisability include creativity, catchiness (i.e., draws attention), memorability (i.e., easy to memorize and recall), clearness (i.e., does not cause confusion), informativeness (i.e., has a message), and distinctiveness (i.e., uniqueness) (Dahl 2011). Creating slogans that exhibit these elements manifests the difficulty of the task.

Slogans change over time and typically are not fixed for advertising campaigns (Kohli, Suri, and Thakor 2002). Brands may change their slogans, for instance, to target a certain audience, provide a new persuasive selling statement for a given product, or reflect changes in the company's values. Mathur and Mathur (1995) have found that firms that change their slogan seem to have positive effects on their market value. Continuous change of slogans can benefit from a slogan generator such as the one introduced in this paper.

Slogans, taglines, and mottoes are similar to the extent that they are considered synonyms. Slogans and taglines are often used interchangeably; however, slogans are made for an advertising campaign whereas taglines are employed as an identifiable phrase for the brand. In other words, a slogan is made for one or more advertising campaigns but a tagline is typically made once for the lifetime of the company. On the other hand, mottoes are sayings that represent a group's (e.g., corporate, political, and religious) vision such as *Google*'s previous motto "Don't be evil".[a] In this

[a] From https://en.wikipedia.org/wiki/Google.

paper, we use the term slogan to refer to all these collectively, given the similarities they have and the difficulties in distinguishing them.

Rhetorical devices. Language is a device for communication; slogans convey a message to the receiver, usually a persuasive one about a concept (i.e., product, service, or company). Like poems, slogans have a stylistic language concerned with *how* a message is expressed. Rhetorical devices such as figures of speech are examples of stylistic language. They exploit the listeners' knowledge of the language and persuade them by redirecting their thinking toward a path intended by the speaker.

Many slogans employ rhetorical devices. For instance, *Yellow Page*'s slogan "Let your fingers do the walking" uses personification expressing fingers as entities capable of walking. Previous research suggests that slogans employing rhetorical devices tend to be favored and remembered better by consumers (Reece, Van den Bergh, and Li 1994). Moreover, different rhetorical devices in slogans have various effects on consumers. For instance, Burgers *et al.* (2015) suggest that slogans containing conventional metaphors are liked and considered more creative than slogans containing irony.

Metaphor. Metaphor is a figurative expression where some properties get implicitly highlighted or attributed from one concept to another one. For instance, the metaphor "time is money" implies that time is valuable without saying it directly: by equating *time* and *money*, the property *valuable* of *money* is attributed to the concept *time*. Other interpretations are also possible, as is usual with metaphors.

A metaphor involves two concepts, a tenor and a vehicle (Richards 1936). In "time is money," *time* is the tenor and *money* is the vehicle. As another example, in *Oakmont Bakery*'s slogan "We create delicious memories,"[b] the tenor (pastry) is implicitly compared to memorable events (e.g., a wedding), implying that it is their cakes that make the event remembered for long. In this example, like in many slogans, the tenor and vehicle are not mentioned explicitly but rather must be inferred from the context. A nominal metaphor, on the other hand, is a metaphor in the simple form "tenor is [a\n] vehicle." "Time is money" is an example of a nominal metaphor.

Multiple theories exist in the literature about metaphors, providing us with guidance into what characteristics are exhibited by metaphors and what makes a metaphor apt. The *salience imbalance theory* (Ortony *et al*. 1985; Ortony 1993) states that metaphoricity occurs when the tenor and vehicle share attributes but some are highly salient for the vehicle only, and this imbalance causes these attributes to be highlighted by the metaphorical expression. Tourangeau and Sternberg (1981) argue that *similarities* within and between the domain of the vehicle and that of the tenor are aspects humans consider when comprehending metaphors. Katz (1989) points out that *concrete* vehicles that are *semantically moderately distant* from the tenor result in apt metaphors. An important property of metaphors is that they are *asymmetrical* in the sense that the metaphor "A is B" highlights different properties than "B is A."

Analysis of slogans. Reece, Van den Bergh, and Li (1994) have analyzed linguistic characteristics of slogans, in addition to other characteristics such as their themes, to find out how they affect receivers in recalling the brand. Their study indicates that utilizing linguistic devices has indeed affected the recall of the brand. The top eight slogans with high recall contained the following linguistic devices: (1) self-reference (i.e., having the brand name in the slogan), (2) alliteration, (3) parallel construction (i.e., repeating rhythm or words from the first phrase in the second phrase), (4) metaphor, and (5) use of a well-known phrase. The authors have also noticed that the slogan with the highest number of correct brand identifications made use of rhymes. As a result,

[b] Slogan examples in this paper are from http://www.textart.ru/, unless otherwise specified.

these linguistic devices seem to have a significant influence on recalling the brand, albeit, some of the frequently found linguistic devices in slogans did not have such outstanding influence, for example, puns.

Inspired by the analysis and taxonomy of linguistic devices used by Reece, Van den Bergh, and Li (1994), Miller and Toman (2016) manually analyzed slogans from various linguistic perspectives, focusing on rhetorical figures and covering other linguistic devices. Their research shows that linguistic devices existed in 92% of 239 slogans, out of which 80% and 42% were schematic and tropic rhetorical devices, respectively. Additionally, the two most common rhetorical devices which were found in figurative slogans are phonetic and semantic devices, covering 87% and 37% of them, respectively. Some phonetic devices appeared more than others, for example, both consonance and assonance occurred in 59% of figurative slogans whereas 32% and 4% of them had alliteration and rhyming, respectively. The semantic device with the highest frequency is metaphor, existing in 24% of rhetorical slogans. Other linguistic devices analyzed by the authors are syntactic, orthographic, and morphological devices which appeared in less than 30 slogans.

A similar manual analysis was conducted by Dubovičienė and Skorupa (2014). Their results also demonstrate that slogans use rhetorical devices frequently, especially figurative language and prosody. However, the percentages of individual rhetorical devices do not match the one by Miller and Toman (2016), which could be due to the difference in the analysis method and the sources of slogans used during the analysis.

Tom and Eves (1999) have found that advertisements containing rhetorical figures are more persuasive and have higher recall in comparison to slogans that do not utilize rhetorical figures. A research conducted by Reece, Van den Bergh, and Li (1994) suggests that recalling a slogan relies largely on the slogan itself, not on the advertising budget, years in use or themes. Furthermore, advertising slogans tend to contain positive words (Dowling and Kabanoff 1996) which would give the receiver a positive feeling about the brand.

Problem definition. We define the task of slogan generation from a computational perspective as follows. Given an input concept/tenor T (e.g., car) and an adjectival property P (e.g., elegant), produce slogans that associate concept T with property P. As a reminder from the beginning of this section, a slogan is a concise, advertisable, and autonomous phrase, where advertisable often implies creativity, catchiness, memorability, or related properties. "Car is elegant," an obvious output for the example task, clearly is not a good slogan.

As the above background on slogans indicates, slogans tend to include rhetorical devices. Among the schematic and tropic rhetorical devices, prosody and metaphor were found to be the most frequent devices (Miller and Toman 2016). Motivated by this, as well as by their effectiveness in enhancing the recall of the brand (Reece, Van den Bergh, and Li 1994), we focus on these two types of rhetorical devices. Besides the usage of rhetorical devices, slogans have positive sentiment and, as a rule, should neither carry negative words nor communicate negative meanings.

The specific slogan generation task that we consider in this paper is the following. Given an input concept/tenor T and an adjectival property P, produce positive slogans that are related to T, that metaphorically associate concept T with property P, and/or that are prosodic. An interesting subtask in its own right is to find a metaphorical vehicle v that attributes property P to concept/tenor T when the concept/tenor is equated with the vehicle.

3. Related work

Research on computational generation of creative expressions is relatively scarce. In this section, we briefly review related work on generating nominal metaphors and on generation of slogans and other creative expressions.

3.1. Computational generation of metaphors

Metaphor Magnet,[c] a web service built by Veale and Li (2012), generates and interprets metaphors by observing the overlap of stereotypical properties between concepts. The metaphor generation process accepts a tenor as input. It uses knowledge regarding properties strongly associated with the tenor to find other concepts, potential vehicles, that share those properties. The aptness of the potential metaphors is measured in the process. The interpretation model, in turn, looks at strongly associated properties shared by the two input concepts (a tenor and a vehicle) and returns the salient features among them. Metaphor Magnet is based on a knowledge base of stereotypical associations, obtained using Google 3-grams and web searches with suitably designed linguistic patterns.

Galvan *et al.* (2016) generate metaphors based on categorizations of concepts and adjectival properties associated with them, as provided by the *Thesaurus Rex* web service (Veale and Li 2013). Their method takes the tenor as input, picks one of its properties at random, and then identifies a vehicle that highlights that property. The vehicle identification starts by finding a suitable category: one that is (strongly) associated to both the tenor and the property. A concept falling in the selected category and with a strong association to the selected property is then chosen as the vehicle.

Xiao and Blat (2013) propose a method for generating pictorial metaphors for advertisements. Their approach takes a concept and a list of adjectival properties to express, and uses multiple knowledge bases, for example, word associations and common-sense knowledge,[d] to find concepts with high imageability. The found concepts are then evaluated against four metrics, namely affect polarity, salience, secondary attributes, and similarity with the tenor. Concepts with high rank on these measures are considered apt vehicles to be used metaphorically.

In contrast to the direct metaphor generation methods above, we employ a metaphor interpretation model to identify apt metaphors that are more likely to result in the desired meaning. The interpretation model, Meta4meaning (Xiao, Alnajjar, Granroth-Wilding, Agres, and Toivonen 2016), uses corpus-based word associations to approximate properties of concepts. Interpretations are obtained by considering salience of the properties of the tenor and the vehicle, either their aggregation or difference.

3.2. Computational generation of slogans

Strapparava, Valitutti, and Stock (2007) propose a "creative function" for producing advertising messages automatically. The function takes a topic and a familiar expression as input, and modifies the expression by substituting some words with new ones related to the given topic. In the process, they use semantic and emotional relatedness along with assonance measures to identify candidate substitutes. This approach is motivated by the "optimal innovation hypothesis" (Giora 2003). The hypothesis states that optimal innovation is reached when novelty co-exists with familiarity, which encourages the recipient to compare what is known with what is new, resulting in a pleasant surprise effect.

Özbal, Pighin, and Strapparava (2013) introduce a framework called *BrainSup* for creative sentence generation. The framework generates sentences such as slogans by producing expressions with content semantically related to the target domain, emotion, and color, and some phonetic properties. Using syntactical treebanks of existing sentences as sentence skeletons and syntactical relations between words as constraints for possible candidate fillers, Özbal *et al.* have employed beam search to greedily fill in the skeletons with candidates meeting the desired criteria.

Using *BrainSup* as a base, Tomašič *et al.* (2014) and Tomašič, Žnidaršič, and Papa (2015) propose an approach for generating slogans without any user-defined target words by extracting

[c] http://ngrams.ucd.ie/metaphor-magnet-acl/.

[d] ConceptNet: http://www.conceptnet.io.

keywords from the textual description of the target concept. Their evaluation criteria are different from *BrainSup*'s evaluation, and they use genetic algorithms instead of beam search.

The approach proposed by Żnidaršič, Tomašič and Papa (2015) employs case-based reasoning where actual slogans written by humans (not their syntactical skeletons) were reused with some modifications in a different context as a new slogan (cf. the approach of Strapparava, Valitutti, and Stock (2007) earlier in this section). The approach commences by retrieving slogans related to the textual description of the input concept using semantic similarities. Slogans are then transformed by replacing content words in them with words from the concept description while satisfying existing part-of-speech (POS) tags.

The *Bislon* method by Repar, Martinc, Znidarsic, and Pollak (2018) produces slogans based on cross-context associations, so-called bisociations (Koestler 1964), and prosody features (alliteration, assonance, consonance, and rhyme). The method accepts three types of input—a set of documents, *Metaphor Magnet* terms (Veale and Li 2012), or domain-specific terms—for both the main concept and the bisociated one. Keywords are automatically extracted from the input and then expanded using a word-embedding model. To generate slogans, the method uses existing slogans as skeletons and fills them with candidate words that match the POS tags of the placeholders. The method ranks slogan candidates based on their relevance to the input and their semantic cohesion as estimated by a language model. Finally, the top slogan candidates are suggested to the user.

In terms of slogan generation in languages other than English, Yamane and Hagiwara (2015) propose a method for producing Japanese taglines related to the input theme and keywords specified by the user. The method generates slogan candidates from a large-scale *n*-gram corpus containing words related to the input. The candidates are then assessed on three aspects: (1) the relatedness of words, (2) grammaticality (based on POS *n*-grams), and (3) novelty (based on combinations of words). The highest scoring candidates are output to the user. Another approach for producing Japanese slogans is proposed by Iwama and Kano (2018).

Figure8 by Harmon (2015) generates metaphorical sentences for a given tenor. Five criteria were considered in the generation process: clarity, novelty, aptness, unpredictability, and prosody. The system selects a property and searches for a suitable vehicle to express it. Thereafter, it composes sentences to express the metaphor by filling in hand-written templates of metaphorical and simile expressions.

Persuasive messages are not only used in slogans, but news headlines also employ them a lot to encourage the audience to read the article (Fuertes-Olivera *et al.* 2001). Gatti *et al.* (2015) have demonstrated how well-known expressions (such as slogans) can be utilized to produce interesting news headlines. Their headline generation process extracts keywords from a news article and then alters man-made slogans based on semantic similarities, dependency statistics, and other criteria, resulting in catchy news headlines.

The method proposed in this paper differs from existing methods for slogan generation in a couple of important aspects. First, it focuses on a specific marketing message, that is, generating slogans for a product while expressing a specific, given adjectival property. In contrast, many of the above methods just create a figurative expression about the given concept without concern for a specific property. Second, the property is to be expressed indirectly via a metaphor, and the metaphor is further automatically generated for the given task. While the above methods often produce metaphoric expressions, they exercise less control over what the metaphor actually expresses. *Bislon* (Repar *et al.* 2018) is an exception: the user is expected to give a bisociated concept which could effectively act as a metaphorical vehicle. Additionally, in this paper we examine several internal evaluation functions used by our method, in order to gain insight into their value in generation of metaphorical slogans.

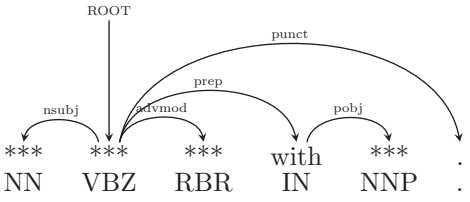

Fig. 1. An example of a skeleton constructed from Visa's slogan: "Life flows better with Visa."

4. Method

Recall our goal: the user (or another software component, if this method is part of a larger system) specifies a concept T and a property P, and the method should suggest slogans related to T. These slogans should associate concept T with property P, preferably in a metaphoric manner, and use prosodic features and avoid negative expressions.

In a nutshell, the slogan generation process involves the following steps that will be detailed in the following subsections:

0. Construction of slogan skeletons, that is, slogan templates that have empty placeholders to be filled in with suitable words (Section 4.1).

Skeleton construction is performed just once to obtain a data set of skeletons for later use. All the following tasks are performed at runtime for the given concept T and property P.

1. *Metaphor generation* ($(T, P) \mapsto v$): Given a concept T and a property P, identify a suitable metaphorical vehicle v to associate the concept and property metaphorically (Section 4.2).
2. *Slogan search space definition* ($(T, v, s) \mapsto \{\mathcal{E}_i\}$): Given a concept T, a vehicle v and a (random) skeleton s, identify sets of words that can potentially be used to fill in the placeholders in skeleton s, in order to obtain grammatical slogan expressions $\mathcal{E}_i$ related to concept T and vehicle v (Section 4.3).
3. *Slogan filtering* ($\{\mathcal{E}_i\} \mapsto \{\mathcal{E}_j\} \subseteq \{\mathcal{E}_i\}$): Given candidate slogan expressions $\mathcal{E}_i$, filter out those lacking internal cohesion or with negative sentiment (Section 4.4).
4. *Internal slogan evaluation* ($f_d(T, P, \mathcal{E}_j) \to \mathbb{R}$): Given a concept T, a property P and a candidate slogan expression $\mathcal{E}_j$, evaluate the quality of the slogan along various dimensions f_d (Section 4.5).
5. *Finding good slogans*: Given the slogan search space $\{\mathcal{E}_j\}$ and the internal evaluation dimensions f_d, carry out actual slogan expression generation and optimization to search the space for slogans $\mathcal{E}_j$ with high $f_d(\mathcal{E}_j)$ (Section 4.6).

4.1. Construction of slogan skeletons

The slogan generation method reuses skeletons, that is, syntactical structures, extracted from existing slogans. Figure 1 shows the skeleton generated from *Visa*'s slogan "Life flows better with Visa." Skeletons are to be filled in with appropriate words, so that a slogan results, as will be described in the following subsections.

A slogan skeleton is a parse tree of a sentence where content words are replaced with a placeholder "***" and where grammatical relations between words and POS tags are maintained. A grammatical relation connects a word (called dependent) to its head word (called governor) with a specific type of relation. The POS tags are based on the Penn Treebank tag set (Marcus, Santorini, and Marcinkiewicz 1993).

Skeletons are constructed once and stored for later use in the slogan generation process. In order to construct a set of skeletons, for the experiments described in this paper we initially obtain 40 well-known good and modern slogans.[e]

We then manually preprocess the slogans to increase parsing accuracy. The first preprocessing step is converting capitalized words into lower case, except the first word and any recognized named entities. This step reduces misclassifications of verbs, adverbs, and adjectives as nouns (e.g., the adverb *differently* in *Red Lobster*'s slogan "Seafood Differently."). Slogans tend to be informal; therefore, we convert words with the suffix *VERB*-in' into *VERB*-ing, in the second step. As a result of the preprocessing phase, *KFC*'s slogan "Finger Lickin' Good." becomes "Finger licking good."

Subsequently, we convert the 40 slogans into parse trees using *spaCy* (Honnibal and Montani 2017). Skeleton candidates are obtained from the parse tree simply by keeping stop words but substituting all content words with placeholders ("***"). Here we use stop words lists from *NLTK* (Bird, Klein, and Loper 2009).

We then keep only those skeletons that can be meaningfully used to generate novel slogans: an acceptable skeleton must have at least two placeholders, and the fraction of placeholders over all tokens must be at least 40%. These choices are made to avoid trivial adaptations, since slogans that are recognizable variations of other slogans are not likely to be good for branding. As a result, *Reebok*'s slogan "I am what I am." will not be reused: it contains no content words, only stop words, so the skeleton would be identical to the original slogan. Several slogans can also produce identical skeletons, for example, *Volkswagen*'s "Think Small." and *Apple*'s "Think Different." In total, the 40 slogans produce 26 unique skeletons (cf. Table 9).

4.2. Computational generation of metaphors

The method aims to identify apt metaphorical vehicles that highlight a given property P in the given target concept (tenor) T. An example of such input is $T = computer$ and $P = creative$. The vehicle identification step does not depend on the skeleton used.

The method begins by retrieving nouns associated with the input property P using two resources: *Thesaurus Rex* (Veale and Li 2013) is used to obtain general nouns such as *coffee* or *flower*, while the resource by Alnajjar *et al.* (2017) provides human categories such as *actor*, *lawyer*, or *politician*. The former will be used for generating general metaphors and the latter for personifications. Given a property, both resources provide a ranked list of nouns associated to this property. As the quality of their results vary, we empirically decided to use only the top 10% of each type, in order to obtain the nouns most strongly related to the given property P. These nouns are used as vehicle candidates.

For example, nouns most strongly associated with $P = creative$ are *{painting, music, . . ., presentation}* and *{artist, genius, poet, . . ., dancer}* in the categories of general and personal nouns, respectively.

The vehicle candidates are not necessarily good vehicles, however, if they do not result in the intended metaphorical interpretation. We therefore use a separate metaphor interpretation model, *Meta4meaning* (Xiao *et al.* 2016), to assess the vehicle candidates in the context of tenor T.

Meta4meaning accepts two nouns as input, a tenor T and a (candidate) vehicle v, and produces a ranked list of possible interpretations for the corresponding nominal metaphor "[tenor] T is [vehicle] v." In other words, *Meta4meaning* outputs a list of properties that it predicts the metaphor to assign to the tenor T via vehicle v. These are not necessarily the properties most strongly associated to vehicle v and they also depend on the tenor T (see later in this section).

[e] Retrieved from http://www.advergize.com/advertising/40-best-advertising-slogans-modern-brands/2/ on 24 October 2016.

We keep a vehicle candidate v only if the desired property P is among the top 50 interpretations of the respective metaphor with tenor T. This ensures that the intended interpretation of the metaphor is feasible. The threshold of 50 top interpretations is chosen based on the results of Xiao *et al.* (2016), which indicate a recall of about 0.5 of human interpretations.

Our implementation of the *Meta4meaning* metaphor interpretation model (Xiao *et al.* 2016) uses the semantic model ω described next to obtain measures of association between nouns and properties (based on word embedding). Following *Meta4meaning*, the method interprets the potential metaphors by considering the shared associations between the tenor and vehicle, and calculating the "combined metaphor rank" metric on them (cf. Xiao *et al.* 2016). In a nutshell, a property is considered a likely interpretation of the metaphor if either the property is strongly associated with both the tenor and the vehicle (as measured by the product of association strengths), or the property has a much stronger association to the vehicle than to the tenor. This metric highlights associations based both on semantic similarities and on salience imbalance between vehicle and tenor. Additionally, since metaphors are asymmetrical, we remove a vehicle candidate if the intended interpretation P is not better in the intended metaphor "T is [a] v" than in the reverse metaphor, that is, "v is [a] T."

Continuing our example, by interpreting all the vehicle candidates in the context of the tenor $T = computer$ and keeping only those for which *creative* is among the top interpretations, we obtain vehicles *{art, drama, . . ., exhibition}* and *{genius, artist, . . ., inventor}* for the general and human categories, respectively. Finally, we merge the two lists of potential vehicles into one list.

To our knowledge, this proposed method is the first for generating metaphors based on their interpretations.

Semantic model ω

We construct a simple semantic model in order to find words that are semantically related to a given word, and to measure the semantic relatedness between two given words. This semantic model is used in several parts of the slogan construction method, not just metaphor generation as described earlier in this section.

We follow the approach described for *Meta4meaning* (Xiao *et al.* 2016) in building the semantic model ω. We obtain co-occurrence counts of words in *ukWaC*[f] (Baroni, Bernardini, Ferraresi, and Zanchetta 2009), a 2 billion word web-based text corpus. Co-occurrences are constrained by sentence boundaries and a window of ± 4 words. We limit the vocabulary of the model to the most frequent 50,000 words, excluding closed class words. We then convert co-occurrence counts to a relatedness measure by employing the log-likelihood measure of Evert (2008) while capping all negative values to zero. Finally, we normalize relatedness scores using L1-norm following McGregor *et al.* (2015). As a result, an ambiguous word (e.g., *bank*) can be related to semantically different words (e.g., *money* and *river*). The semantic model does not aim to handle polysemy in any informed manner.

Examples of words related to the concept *computer* in the semantic model ω include *{system, software, network, skill, . . ., workstation}*.

4.3. Search spaces for filling in skeletons

When producing slogan expressions, the method considers one skeleton s at a time, for the given concept T and vehicle v. The relationship to property P comes (metaphorically) via words related to vehicle v. Throughout this paper, we use vehicles generated by the metaphor generation process described above, but vehicle v could be input manually as well.

[f] http://wacky.sslmit.unibo.it.

To instantiate a skeleton, the method constructs sets of words that can be used as potential fillers for each placeholder i in skeleton s. It starts by identifying the *grammatical space* $\mathcal{G}_i$ consisting of all words that have the POS and grammatical relations matching placeholder i in skeleton s. Similar to the approaches by Özbal, Pighin, and Strapparava (2013) and Tomašič, Žnidaršič, and Papa (2014), we build a repository of grammatical relations, that is, of pairs of words that occur in each grammatical relationship to each other. The repository is built once, and is then used to identify $\mathcal{G}_i$ at runtime by retrieving words that match the relevant relations from the repository. To construct the repository, we parse the entire *ukWaC* corpus using *spaCy* and store all grammatical relations observed along with their frequencies. We retain grammatical relations with frequencies at least 50 to remove rare and noisy cases. The process yields 3,178,649 grammatical relations, which are publicly available (Alnajjar 2018).

We then further identify those grammatical words that are also related either to the input concept T or the vehicle v, according to the semantic model ω described above. This set of related and grammatical words is the *related space* $\mathcal{R}_{i,T,v}$, or just $\mathcal{R}_i$ for short when the concept T and vehicle v are clear in the context. In order to identify the related space, the method obtains those words in $\mathcal{G}_i$ that are either within the k words most strongly related to concept T, or within the k words most strongly related to vehicle v. In our case, k was empirically set to 150. Since abstraction tends to be required in processing metaphors (Glucksberg 2001), we only accept abstract terms related to vehicle v. For this, we utilize the abstractness data set provided by Turney *et al.* (2011) and keep words with abstractness level at least 0.5.

Given a skeleton s, concept T and vehicle v, the search space for possible slogans consists of all feasible ways of filling each placeholder i with a word from the respective related and grammatical space $\mathcal{R}_i$. Alternatively, if the above is not feasible, grammatical (unrelated) words in $\mathcal{G}_i$ can be used as fillers.

As an example, let the skeleton s be

***_NN, ***_NN and ***_NN.

That is, three singular nouns (NN) separated by a comma and *and* (with grammatical relations omitted for simplicity). Let concept T be *computer* and vehicle v be *artist*. The grammatical space $\mathcal{G}_{i=1}$ for the first placeholder consists of all singular nouns in the grammatical repository (that satisfy all relations linked to it, such as the "punc" relation to the second token ","). Examples of filler candidates in $\mathcal{G}_1$ are *{management, talent, site, skill, . . ., health}*. The related and grammatical space $\mathcal{R}_1$ for the same placeholder is the subset of $\mathcal{G}_1$ that is related to *computer* or *artist* in the semantic model ω: {*system, skill, programming, art, designer, talent, simulation, . . .*}. A random filler word is then selected from $\mathcal{R}_1$ (e.g., *talent*) or, if the set were empty, then an (unrelated) filler is chosen at random from $\mathcal{G}_i$. This process is repeated for each placeholder, yielding slogans such as

"software, design and simulation."

and

"talent, talent and support."

4.4. Filtering criteria for slogan expressions

Not all expressions in the search space defined above are suitable as slogans. We use two criteria to filter out expressions that are not likely to be good slogans: lack of cohesion within the expression, and negative sentiment.

Semantic cohesion is measured to avoid slogans that have mutually unrelated words. We require that all content words (i.e., words used in the placeholders) are semantically related to each other, according to the semantic model ω. If any pair of content words is not related, the expression is

discarded. Alternatively, we could use a nonbinary measure of cohesion. We will return to this in the discussion.

As advertising slogans tend to be positive expressions (Dowling and Kabanoff 1996), we employ *sentiment analysis* to prevent negative sentiment. We use the sentiment classifier provided in *Pattern* (De Smedt and Daelemans 2012) to predict the sentiment polarity score of expressions. The score is a value between −1 and +1; we discard slogan expressions with a negative score.

4.5. Internal evaluation dimensions for slogan expressions

With the spaces $\mathcal{R}_i$ and $\mathcal{G}_i$ and the filtering criteria above, we have defined a space of possible slogans. Still, some expressions in the space are likely to be better slogans than others, and we next define four internal evaluation dimensions that the slogan generator can use. Our hypothesis is that the dimensions are useful ones, and we will test this hypothesis empirically in the experimental section.

The four dimensions are (1) target relatedness, that is, relatedness to concept T and property P, (2) language, (3) metaphoricity, and (4) prosody. Each dimension can be further composed of multiple sub-features.

4.5.1. Target relatedness (to concept T and property P)

Slogan expressions generated according to the above-defined constraints relate to concept T and property P to varying degrees. By construction, the search space favors content words that are related to concept T or vehicle v, but property P is not considered directly because we want to encourage this relation to be metaphoric. Given that a slogan eventually intends to connect property P to concept T, it seems natural to measure and possibly maximize the relationship of the slogan expression to the target input, that is, both concept T and property P.

Formally, we measure semantic relatedness $f_{rel}(\mathcal{E}, w)$ between a slogan expression $\mathcal{E}$ and a single target word w as the mean relatedness

$$f_{rel}(\mathcal{E}, w) = \frac{\sum_{t \in c(\mathcal{E})} \omega(t, w)}{|c(\mathcal{E})|} \tag{1}$$

where $c(\mathcal{E})$ is the set of content words (i.e., filler words in placeholders) in slogan expression $\mathcal{E}$ and $\omega(t_i, w)$ is a score given by the semantic relatedness model ω. The internal evaluation dimension of *relatedness* (to concept T and property P) is computed as a weighted sum of the semantic relatedness of the slogan expression to T and to P. The weights are given in Table 1. (The other three dimensions are also computed as weighted sums of their sub-features; all weights are given in the table.) We chose to give relatedness to P a higher weight as the search space already consists of words related to the concept T.

4.5.2. Language

Skeletons, with their grammatical relations and POS tags, aim to ensure that slogan expressions produced with them are likely to be grammatically correct. However, these constraints are not sufficient to guarantee correctness. We resort to a simple statistical method, bigrams, to obtain an alternative judgment, in the form of a likelihood of the slogan expression in comparison to a large corpus. In addition, under the language dimension, we also consider surprisingness (rarity) of the individual words in the expression.

We build a probabilistic language model using bigram frequencies provided with the *ukWaC* corpus. A slogan with higher probability according to the language model is more likely to be grammatically correct as its bigrams appear more frequently in the *ukWaC* corpus. Employing

Table 1. The weights assigned to each sub-feature in the four internal evaluation dimensions

Dimension	Feature	Weight
Relatedness	$f_{rel}(\mathcal{E}, T)$	0.4
	$f_{rel}(\mathcal{E}, P)$	0.6
Language	$Prob(\mathcal{E})$	0.8
	$f_{unusual}(\mathcal{E})$	0.2
Metaphoricity[a]	$f_{metaph\text{-}maxrel}(\mathcal{E}, T, v)$	0.5
	$f_{metaph\text{-}diffrel}(\mathcal{E}, T, v)$	0.5
Prosody	$f_{rhyme}(\mathcal{E})$	0.4
	$f_{alliteration}(\mathcal{E})$	0.4
	$f_{assonance}(\mathcal{E})$	0.1
	$f_{consonance}(\mathcal{E})$	0.1

[a]In case the value of this dimension is negative (i.e., when a word in the expression $\mathcal{E}$ is related to the concept/tenor T more than to the metaphorical vehicle v), it is capped to zero.

bigrams, in contrast to trigrams or higher n-grams, gives the method a greater degree of freedom in its generation; higher n-grams would improve the grammar of the generated expressions but would tie them to expressions in the original corpus.

Surprisingness is the other feature we consider in the language dimension, inspired by Özbal, Pighin, and Strapparava (2013). We measure how infrequent, that is, unusual, the individual words in the slogan are

$$f_{unusual}(\mathcal{E}) = \frac{\sum_{t \in c(\mathcal{E})} \frac{1}{freq(t)}}{|c(\mathcal{E})|} \tag{2}$$

where $freq(t)$ is the absolute frequency of word t in the *ukWaC* corpus, and where word t is ignored in the computation (both nominator and denominator) if its frequency $freq(t)$ is zero. While such words could be surprising, they also add noise, so we consider it safer to ignore them. In case no content word appears in the corpus, the surprisingness score is defined to be zero; that is, we conservatively consider the expression not to be surprising. The weights assigned to these sub-features when representing the entire language dimension were set empirically (cf. Table 1).

4.5.3. Metaphoricity

By construction, slogan expressions in the defined search space are encouraged to be metaphorical, but their degree of metaphoricity varies. We define two functions that aim to measure some aspects of metaphoricity in the produced slogan expressions. In these functions, we use both the concept/tenor T and the metaphorical vehicle v used in the construction of the expression.

The first function, $f_{metaph\text{-}maxrel}$, considers the strongest relationships between any of the content words $t \in c(\mathcal{E})$ in slogan $\mathcal{E}$, and the tenor T and the vehicle v:

$$maxrel(\mathcal{E}, w) = \max_{t \in c(\mathcal{E})} \omega(t, w) \tag{3a}$$

$$f_{metaph\text{-}maxrel}(\mathcal{E}, T, v) = maxrel(\mathcal{E}, T) \cdot maxrel(\mathcal{E}, v) \tag{3b}$$

where $\omega(\cdot)$ is a score given by the semantic relatedness model. When this function has a value larger than zero, then the slogan contains a word that is related to the concept/tenor and a (possibly same) word that is related to the vehicle. The larger the value, the more related these words are to the concept/tenor and vehicle. Obviously, a slogan that is not (strongly) related to both the concept T and the vehicle v can hardly be metaphorical in the intended manner.

The other metaphoricity function, $f_{metaph\text{-}diffrel}$, checks whether the slogan expression $\mathcal{E}$ contains a word t that is strongly related to the metaphorical vehicle v but *not* to the concept/tenor T. The hypothesis is that such a word t is more likely to force a metaphorical interpretation of the expression, in order to connect t to the concept/tenor T. For instance, let the tenor T be *car* and the vehicle v be *dancer*, and let candidate content words related to *dancer* be *stage* and *street*. The expression "cars of stage" is much more likely to have a metaphorical interpretation than the expression "cars of street," since the word *stage* used in the former is *not* related to cars. Function $f_{metaph\text{-}diffrel}$ is introduced to measure and encourage this metaphoricity arising from words t related to the vehicle v but *not* to the concept/tenor T as follows:

$$f_{metaph\text{-}diffrel}(\mathcal{E}, T, v) = \max_{t \in c(\mathcal{E})} (\omega(t, v) - \omega(t, T)) \tag{4}$$

The internal dimension of metaphoricity is obtained as the sum of the two sub-features, that is, they are given equal importance.

4.5.4. Prosody

In our work, we consider four features of *prosody*: rhyme, alliteration, assonance, and consonance. For this, we make use of *The CMU Pronouncing Dictionary* (Lenzo 1998) to analyze repeated sounds in words. *The CMU Pronouncing Dictionary* is a mapping dictionary from English words to their phonetic translations. While the dictionary is limited by its vocabulary, the vocabulary is relatively extensive as it contains over 134,000 words.

Let $\varphi(t)$ be *CMU*'s function which returns the sequence of phonemes in a given text (word t or slogan $\mathcal{E}$), and let *vowels* be the set of (phonetic transcriptions of) vowels. $\mathbb{1}_X$ is an indicator function that returns 1 if X is true and 0 otherwise.

Equation (5a) is for counting the total number of occurrences of phoneme *pho* in slogan $\mathcal{E}$. We only consider sounds repeated at least three times (Equation (5b)).

$$count_{phoneme}(\mathcal{E}, pho) = \sum_{t \in \mathcal{E}} \sum_{p \in \varphi(t)} \mathbb{1}_{p=pho} \tag{5a}$$

$$count_{phoneme \geq 3}(\mathcal{E}, pho) = \begin{cases} count_{phoneme}(\mathcal{E}, pho), & \text{if } count_{phoneme}(\mathcal{E}, pho) \geq 3 \\ 0, & \text{otherwise} \end{cases} \tag{5b}$$

We implement the *assonance* and *consonance* functions by considering the total relative frequency of vowels or consonants, respectively, that are repeated at least three times:

$$f_{assonance}(\mathcal{E}) = \frac{\sum_{pho \in vowels} count_{phoneme \geq 3}(\mathcal{E}, pho)}{|\{\varphi(\mathcal{E})\}|} \tag{6a}$$

$$f_{consonance}(\mathcal{E}) = \frac{\sum_{pho \notin vowels} count_{phoneme \geq 3}(\mathcal{E}, pho)}{|\{\varphi(\mathcal{E})\}|} \tag{6b}$$

For *alliteration* and *rhyme*, we count the number of word pairs that share their first or last phonemes, respectively, regardless of their quality and stress. For simplicity, syllables are not taken into account. Denoting the first phoneme in a word t by $\varphi(t)_0$ and the last by $\varphi(t)_{-1}$, the measures are as follows:

$$f_{alliteration}(\mathcal{E}) = \frac{\sum_{t_i,t_j \in \mathcal{E}, t_i \neq t_j} \mathbb{1}_{\varphi(t_i)_0 = \varphi(t_j)_0}}{|\mathcal{E}|} \tag{7a}$$

$$f_{rhyme}(\mathcal{E}) = \frac{\sum_{t_i,t_j \in \mathcal{E}, t_i \neq t_j} \mathbb{1}_{\varphi(t_i)_{-1} = \varphi(t_j)_{-1}}}{|\mathcal{E}|} \tag{7b}$$

4.6. Algorithm for finding good slogans

We employ genetic algorithms to find good slogans in the above-described space of possible expressions, given a skeleton s, related words $\mathcal{R}_i$, and grammatical words $\mathcal{G}_i$ for each placeholder i, as well as the filtering criteria and internal evaluation dimensions described above. We use Deap (Fortin *et al.* 2012) as the evolutionary computation framework. Next, we use μ to denote the size of the population, G the number of generations to produce, and $Prob_m$ and $Prob_c$ the probability of the mutation and crossover, respectively.

As an overview, the algorithm first produces an initial population of slogan expressions ("individuals") and then evolves it over G iterations. Starting with the initial population, the employed $(\mu + \lambda)$ evolutionary algorithm produces λ number of offspring by performing crossovers and mutations according to the respective probabilities $Prob_m$ and $Prob_c$. The algorithm then puts the current population and offspring through a filtering process (described below). The population for the next generation is produced by evaluating the current population and the offspring, and then selecting μ number of individuals. The evolutionary process ends after the specified number of generations. Details of the process will be given in the following paragraphs.

Initial population. Given a skeleton s, related words $\mathcal{R}_i$, and grammatical words $\mathcal{G}_i$, the algorithm produces a new individual (i.e., slogan expression) as follows. It begins by filling the placeholder with the most dependent words to it, usually the root. The algorithm attempts to randomly pick a related word from $\mathcal{R}_i$. If, however, the set is empty, that is, there are no related and grammatical words that can be used in the placeholder, a grammatical word is randomly picked from the set $\mathcal{G}_i$. The algorithm repeats the above steps to fill in the rest of the placeholders, always taking into account the conditions imposed by the already filled words. If the method fails to locate a suitable filler for a placeholder also in $\mathcal{G}_i$, the individual (expression) is discarded and the filling process starts over with a new individual. The process above is repeated until the desired number of individual expressions is generated, serving as the initial population.

Mutation, crossover, and filtering. Our algorithm employs one type of mutation which substitutes filler words in placeholders. The probability of producing an offspring by mutation is $Prob_m$. In the substitution, the mutation operation follows a similar process as for the initial population to find a related and grammatical word for the placeholder. For instance, mutating the slogan

"talent, talent and support."

begins by turning a random content word back into a placeholder (e.g., "talent, ***_NN and support.") and then filling the placeholder with a new word from the relevant space $\mathcal{R}_i$. A new variant of the slogan results, such as

"talent, design and support."

The algorithm applies a one-point crossover on two individuals with probability $Prob_c$; that is, any pair of individuals is crossed over with probability $Prob_c$. As an example, a crossover of the two slogans

"work, skill and inspiration."
"talent, design and support."

after the third token would yield

"work, skill and support."
"talent, design and inspiration."

The resultant newly generated child expressions are put through a grammatical check, verifying that the filler word in each placeholder i is in the grammatical space $\mathcal{G}_i$ also when considering the other content words that may have changed meanwhile. A failure of the grammatical check, for any of the two children, results in their disposal while parent expressions are kept in the population.

All offspring are filtered based on lack of internal cohesion, or negative sentiment, as described in Section 4.4. Additionally, mutation and crossover may produce duplicate slogans; once a new generation is produced, the filtering process also removes any duplicates.

Fitness functions and selection. The genetic algorithm uses the four internal evaluation dimensions defined in Section 4.5 as its fitness functions: (1) target relatedness, (2) language, (3) metaphoricity, and (4) prosody.

Some of the evaluation dimensions are conflicting in nature. For instance, the target relatedness dimension favors words related to the target concept T and property P, while the metaphoricity dimension favors words related to concept T and the metaphorical vehicle v. A single ranking method for selection, based on some linear combination of the dimensions, would not allow different trade-offs between the evaluation dimensions. Instead, our selection process involves the nondominant Non-dominated Sorting Genetic Algorithm II (NSGA-II) algorithm (Deb *et al.* 2002) that looks for Pareto-optimal solutions, that is, solutions that cannot be improved any further without degrading at least one of the internal evaluation dimensions. This approach supports diversity among multiple, potentially conflicting objectives.

5. Empirical evaluation

We carried out human evaluations of both slogans and metaphors generated by the method. Metaphors were evaluated on their own since their generation method is novel, and since metaphors have a central role in the slogan generation process. The evaluations were carried out as crowdsourced surveys on Crowdflower.[g] Crowdsourcing allowed us to gather large amounts of judgments of metaphors and slogans and to carry out quantitative analysis on them. We targeted our surveys to the following English-speaking countries: the United States, the United Kingdom, New Zealand, Ireland, Canada, and Australia.

As input to the slogan generation system, we used concept–property pairs and let the system generate slogans for them. Given that the space of possible concept–property pairs for slogan generation is not closed, and that no obvious distribution exists from which to draw a representative sample of concept–property pairs, we resorted to manually selecting a diverse collection of 35 concept–property pairs (Table 2). These pairs were inspired by Xiao and Blat (2013) and defined by the authors of this paper to represent a range of different concepts and different properties, including both typical ("chocolate is sweet") and less typical associations ("computer is creative"). The aim is to use this set as a proof of concept across a range of slogan generation tasks; the results obviously are specific to this data set. The concept–property pairs were chosen before the tests described in the following were carried out, so they have not been cherry-picked to give good results. From the 35 concept–property pairs, we generated 212 metaphors and subsequently 684 slogans. Each slogan and metaphor was evaluated through CrowdFlower.

[g] www.crowdflower.com.

Table 2. The 35 concept–property pairs used to evaluate the methods

Concept	Properties
book	wise, valuable
chocolate	healthy, sweet
computer	creative, mathematical, powerful
painting	creative, majestic, elegant
car	elegant, exotic, luxurious
university	diverse, valuable
coke	sweet, dark
museum	ancient, scientific
love	wild, beautiful, hungry
professor	old, wise, prestigious, smart
newspaper	commercial, international
paper	white, empty, scientific
politician	powerful, dishonest, persuasive, aggressive

Each property is used individually with the respective concept.

By design, a main goal of the slogan generation method proposed in this paper is to produce metaphoric slogans. Given the central role of metaphors for the method, we first evaluate the metaphor generation component. Discussion of the results is deferred to Section 6.

5.1. Evaluation of metaphor generation

As described in Section 4.2, the metaphor generation method is based on a metaphor interpretation model; that is, the method looks for an apt vehicle such that the interpretation of the resulting metaphor is as close to the intended meaning as possible. In this evaluation, we compare these generated apt vehicles to various baselines.

Given the 35 inputs in Table 2, the method produced 53 apt vehicles, that is, vehicles that are considered by the method to highlight the input property P in the input concept/tenor T. Out of these vehicles, 31 are general nouns and 22 are human. Tables 3 and 4 list ten random examples of generated vehicles in both classes, respectively (column "Generated Apt Vehicles").

For each generated apt vehicle, we generated three matching baseline vehicles without the metaphor interpretation model:

- A *strongly related* vehicle is selected at random among the same top 10% of nouns associated to property P as considered by the metaphor generation method (cf. Section 4.2), but under the constraint that it is not considered apt by the generation method.
- A *related* vehicle is selected randomly among the bottom 90% of nouns associated with property P.
- A *random* vehicle is picked from those nouns that are not associated at all with property P.

Given that we have two classes of vehicles, general and human, we picked the baseline vehicles always from the same class as the apt vehicle. Baseline vehicles for the random examples are also given in Tables 3 and 4.

Table 3. Random examples of vehicles in the class of general nouns, both the apt vehicle generated by the method and three baseline vehicles

Input		Generated	Baselines		
Tenor	Property	Apt vehicle	Strongly related	Related	Random
book	valuable	purse	image	ginger	metal
painting	elegant	velvet	tuberose	aluminum	gps
car	elegant	scarf	tuberose	mahogany	mold
professor	smart	refrigerator	dolphin	weapon	pomfret
computer	creative	poet	performance	speech	bittersweet
professor	old	tractor	printer	beads	timber
politician	Aggressive	bullying	wrestling	skateboarding	ambulance
chocolate	Healthy	colon	herb	aorta	tantrism
museum	Ancient	latin	brachiopod	universe	crocodile
love	beautiful	art	line	moonstone	deerskin

Table 4. Random examples of vehicles in the class of humans, both the apt vehicle generated by the method and three baseline vehicles

Input		Generated	Baselines		
Tenor	Property	Apt vehicle	Strongly related	Related	Random
book	wise	father	judge	brother	marker
museum	scientific	scientist	computer	technologist	apartment
computer	powerful	king	tyrant	mogul	grief
politician	powerful	monster	emperor	thug	temple
professor	wise	king	father	politician	executive
coke	sweet	mother	friend	mistress	cinema
coke	dark	demon	terrorist	spy	travel
paper	scientific	scientist	computer	philosopher	hexachlorophene
professor	old	child	king	invalid	tendon
love	wild	cat	warrior	pirate	orator

Given the 53 generated apt vehicles and three baselines for each of them, we obtained a total of 212 metaphors to evaluate. For the evaluation, we represented each of them as a nominal metaphor of the form "*T* is [a/n] *v*" (e.g., "computer is an artist"). We then asked judges if the metaphor expresses the intended property (that computer is creative). The judges used a five-point Likert scale where 1 indicates strong disagreement and 5 strong agreement. The order of metaphors was randomized for each judge. Ten judges were required to evaluate every metaphor.

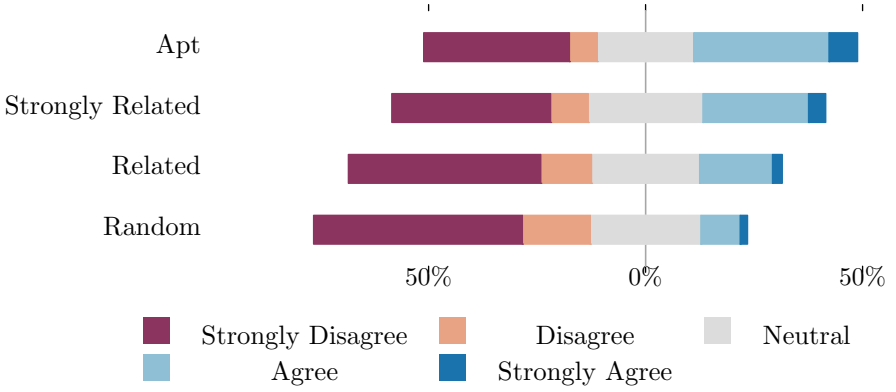

Fig. 2. Success of metaphor generation: agreement that the generated metaphor expresses the intended property

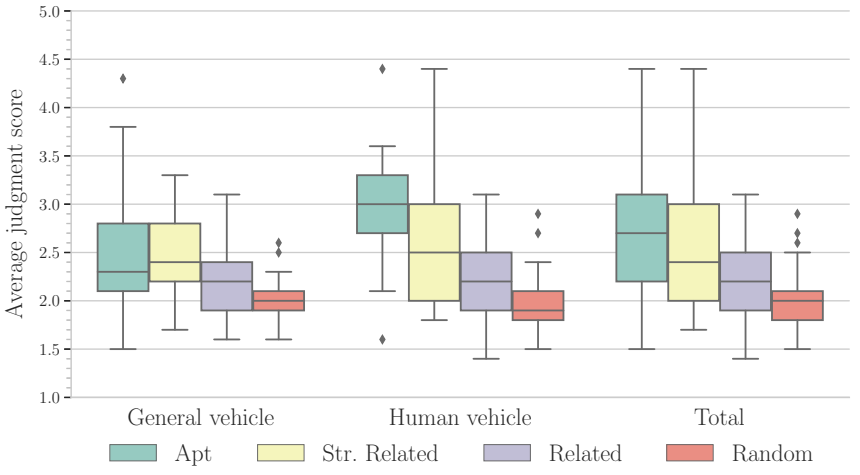

Fig. 3. Distributions of mean judgements over metaphors with different types of vehicles (apt vehicles used by the method, strongly related baseline, related baseline, and random baseline). Results are given separately for general and human classes of vehicles, as well as for their combination ("Total"). Plots indicate the median, fist and third quartiles and 95% intervals.

A summary of results is given in Figure 2 in the form of a diverging bar chart illustrating the percentages of judgments on the Likert scale for each type of vehicles tested (the generated apt vehicle, and the baselines of strongly related, related, and random).

We can observe that apt vehicles performed best, followed by the baseline vehicles in the order of strength of relatedness to the property. Overall, judges agreed or strongly agreed 38% of the time that nominal metaphors constructed with apt vehicles expressed the intended property. On the other hand, metaphors where the vehicle was strongly associated with the property (but not apt according to the method) were successful in 28% of the cases. The corresponding agreements are even lower for (non-strongly) related vehicles, 19%, and non-related vehicles, 11%.

Figure 3 shows the distributions of mean judgements over the metaphors generated. The first group of bars is for metaphors with general vehicles, the second group with human vehicles, and the third group represents their union. Table 5 provides the respective numbers.

Based on the results, we can observe that apt and also strongly related vehicles of the human class performed best. Their median scores of 3.0 and 2.5, respectively, also outperform apt general vehicles (median 2.3). Within the group of general vehicles, apt and strongly related vehicles performed best.

Table 5. Five-number summaries (median, first and third quartiles, minimum and maximum values) of the mean judgments of metaphors

	Generated				Baselines			
	Apt vehicle		Str. related		Related		Random	
General	2.3		2.4		2.2		2.0	
vehicles	2.1	2.8	2.2	2.8	1.9	2.4	1.9	2.1
($n = 31$)	1.5	4.3	1.7	3.3	1.6	3.1	1.6	2.6
Human	3.0		2.5		2.2		1.9	
vehicles	2.8	3.3	2.0	3.0	1.9	2.5	1.8	2.1
($n = 22$)	1.6	4.4	1.8	4.4	1.4	3.1	1.5	2.9
Total	2.7		2.4		2.2		2.0	
($n = 53$)	2.2	3.1	2.0	3.0	1.9	2.5	1.8	2.1
	1.5	4.4	1.7	4.4	1.4	3.1	1.5	2.9

n denotes the number of metaphors evaluated; the number of individual judgments is tenfold.

The combined results (group "Total") suggest that the generated apt vehicles outperform the baselines. A statistical test validates this observation. Nonparametric permutation test shows that the mean judgment of apt vehicles is statistically significantly higher than the mean judgment of strongly related vehicles, $P = 0.0074$ (one-tailed).

5.2. Evaluation methodology for slogan generation

We next evaluate the generated slogans. The primary goal is to identify whether the proposed method is capable of producing expressions suitable for the task, that is, feasible as advertising slogans. A secondary goal is to investigate the effects of the evaluation dimensions of the genetic algorithm on the produced slogans. With this, we hope to shed light on computational criteria for future slogan generation methods. The evaluation setup for slogan generation is the following.

For every triplet of concept T, property P, and (apt) vehicle v obtained from the metaphor generation stage, we randomly select two skeletons. In our experiments, we have a set of 26 skeletons to choose from; the number of skeletons applied per input is here limited to two for simplicity of experimental design. In real applications, a wider selection would provide more variation.

One skeleton at a time is filled in by the genetic algorithm. We empirically set the following values for parameters of the genetic algorithm: $\mu = \lambda = 100$, $G = 25$, $Prob_c = 0.4$, $Prob_m = 0.6$.

We selected multiple slogans for evaluation from the final population produced by the genetic algorithm, in order to study the effects of various evaluation dimensions on the quality of slogans. As described in Section 4.5, there are four **internal evaluation dimensions**: (1) *relatedness* of the slogan to the concept and the property given as input, (2) *language*, (3) *metaphoricity*, and (4) *prosody*. Because these dimensions are partially mutually contradictory, we evaluate slogans that have different trade-offs between them. For the experiments of this paper, we used three **selection methods** for slogans:

- *Balanced* dimensions: A randomly selected slogan that has a positive value on several internal evaluation dimensions. In addition to requiring that all four dimensions are positive, we also try the cases where this requirement is relaxed either for prosody or for metaphoricity.

- A *maximized* dimension: A slogan with the maximum value on one of the four dimensions, regardless of other dimensions.
- *Minimized* dimensions: A random slogan with the lowest values on all four dimensions (relatedness, language, metaphoricity, and prosody, considered in order).

This selection yielded 684 slogans to be evaluated. The balanced selection failed for some cases because no slogan in the generated population met the selection criteria.

In order to represent the slogans in a uniform, slogan-like style, we detokenize them using *NLTK*, capitalize the words in them, and add a full stop in the end.

We asked five judges to evaluate each selected slogan on a five-point Likert scale based on the following five aspects or **judgments**: (1) the *relatedness* of the slogan to the title (i.e., input concept and property), (2) the *language correctness*, (3) the *metaphoricity*, (4) the *catchiness, attractiveness and memorability*, and (5) the *overall quality* of the expression as a slogan.

These judgments and the internal evaluation dimensions described above consider similar aspects. With this design, we intend to measure how well the internal evaluation dimensions are reflected in the output, as well as to test how they contribute to the overall quality of the generated slogans.

To simplify some of the analyses next, we consider the overall quality of an individual slogan to be *good* if the mean judgment is above 3 for the question "Overall, this is a good slogan." In some of the analyses, we also do such dichotomization to the other judgments.

For a comparison of computer-generated slogans to professionally crafted ones, we ran a similar survey with slogans produced by professionals for past advertising campaigns. We use http://www.textart.ru/[h] due to its consistent structure of listing slogans and wide coverage of slogans from different categories. The corpus includes additional information regarding slogans such as the name of the brand and its category (e.g., pizza or university). In the experiment, we use 100 random slogans obtained from the above site. In order to reduce the effect of familiarity of the brand on the evaluation, we manually substituted product and brand names with the text "ProductName." We also had to adjust the first evaluation question about relatedness: due to the lack of explicit input concepts and properties in the human-made slogans, we used the product's category (provided in the database) as the target concept T and removed the property P from the question. We required 10 judges to provide their opinions on each human-made slogan and thus received a total of 1000 judgments.

It is worth noting that a direct comparison between the results of computer-made slogans and human-made ones is not feasible. First, the two evaluations are not identical (e.g., missing the adjectival properties from evaluated human-made slogans, nonequivalent number of judges, and nonidentical judges); second, some artificial constraints were enforced during computational slogan production (e.g., computer-made slogans were restricted to two skeletons). It is also good to keep in mind that generated slogans are intended as slogan candidates for brainstorming. Nevertheless, juxtaposing the results for computer-generated and existing slogans can give useful insights.

5.3. Overview of results for slogan generation

As concrete examples of what the experimental setup produced, Table 6 shows some generated slogans, both more and less successful ones.

Figure 4 gives the distributions of judgments on the overall suitability of slogans, and on their catchiness. Slogans created by professionals stand out, as expected, but the generated slogans fair well, too. The judgments are centered around 3 and have a relatively wide distribution, indicating

[h] Collected on 24 October 2016.

Table 6. Examples of generated slogans

Concept	Property	Vehicle	Output
computer	creative	artist	"Talent, Skill And Support."
computer	creative	artist	"Follow Questions. Start Support."
computer	creative	poet	"Work Unsupervised."
computer	creative	poet	"Younger Than Browser."
car	elegant	dancer	"The Cars Of Stage."
painting	creative	literature	"You Ca N't Sell The Fine Furniture."
politician	persuasive	orator	"Excellent By Party. Speech By Talent."
politician	dishonest	thief	"Free Speech."
politician	aggressive	predator	"Media For A Potential Attack."

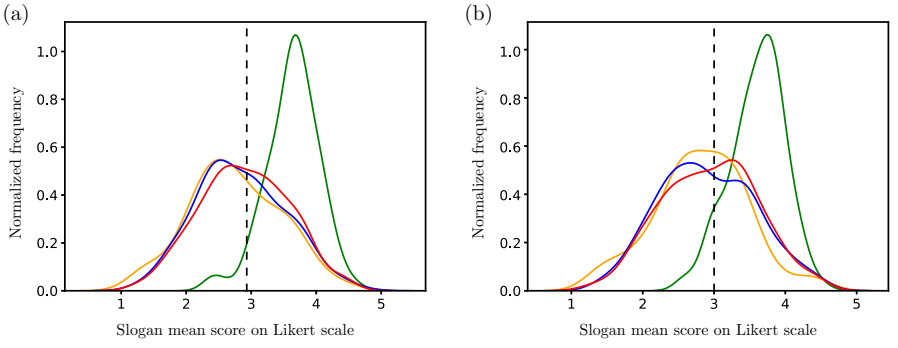

Fig. 4. Distributions of judgments for overall quality and catchiness for generated slogans (*balanced* in red, *maximized* in blue, and *minimized* in orange) and expert-written slogans (in green). (The graphs show distributions over slogans, where each slogan is represented by its mean score.). (a) Overall quality. (b) Catchiness.

that while most slogans are neutral in the Likert scale, there are also some relatively good and some relatively poor ones.

A comparison between different selection methods indicates that balanced slogans contain somewhat more suitable ones (i.e., with scores larger than 3) than the other selection methods. This observation is similar for catchiness (Figure 4(b)) and for other judgments (not shown).

Table 7 provides a numerical summary of the performance of slogans with regard to all judgments. We observe that the balanced selection performs best in all judgments and the minimized selection worst by a clear margin. A comparison across different judgments in Table 7 shows that language correctness received the best scores, followed by relatedness, catchiness, and, finally, metaphoricity.

In total, over all selection methods, 35% of generated slogans were judged to be suitable (and 39% of the balanced slogans). The input that resulted in most suitable slogans was *computer–powerful*, with 13 suitable slogans out of 20 generated for it. On the other hand, input *newspaper–international* had the least number of successful slogans, 1 out of 12. This means that the method has generated at least one successful slogan for each input, even though we only used two random skeletons for each input.

Table 7. The percentage of slogans being judged as successful with respect to different aspects

Selection method	Relatedness (%)	Language (%)	Metaphoricity (%)	Catchiness (%)	Overall (%)
Balanced ($n = 466$)	48	52	39	44	39
Maximized ($n = 389$)	45	49	39	40	35
Minimized ($n = 104$)	28	38	28	36	32
Expert ($n = 100$)	94	98	84	89	92

A slogan is considered successful if the respective mean score is greater than 3.

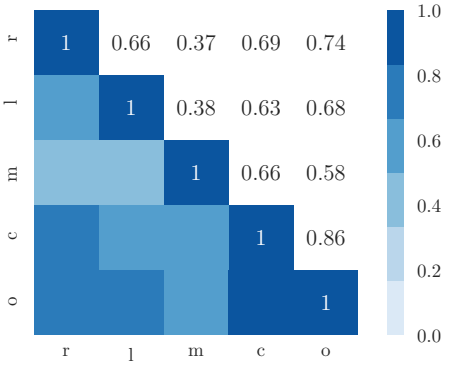

Fig. 5. Pearson correlation coefficient of judgments on human-made slogans between the five questions: (r)elatedness, (l)anguage, (m)etaphoricity, (c)atchyness, and (o)verall quality.

5.4. Human judgments and evaluation dimensions

In this paper, we decided to focus on four different aspects of slogans: relatedness, language (correctness), metaphoricity, and catchiness/prosody. How do these four aspects relate to the overall suitability of slogans?

We measured all correlations between the four human judgments and the overall quality using human-made slogans (Figure 5). Correlations of the four judgments with the overall quality are strong (line and column "o" in the figure), ranging from 0.86 for catchiness to 0.58 for metaphoricity. This suggests that all four aspects contribute to the overall quality of slogans, especially catchiness and relatedness.

Correlations between the four judgments tend to be strong as well, over 0.5, except for correlations between metaphoricity and relatedness (0.37), and between metaphoricity and language correctness (0.38).

Overall, the high levels of correlation between the four judgments and the overall suitability suggest that all the four aspects should be balanced rather than only maximizing some of them. This is in line with the observation made above that a balanced selection produces better slogans.

Human judgments versus internal evaluation dimensions. Above we established that catchiness, relatedness, language, and metaphoricity are all important factors in slogans. How well do the respective internal evaluation dimensions correlate with the judgments in the survey, that is, does the method optimize the right things?

Here, we consider the sets of successful and unsuccessful slogans with respect to each human judgment type separately, and compute the mean values of the corresponding internal evaluation dimension in both sets.

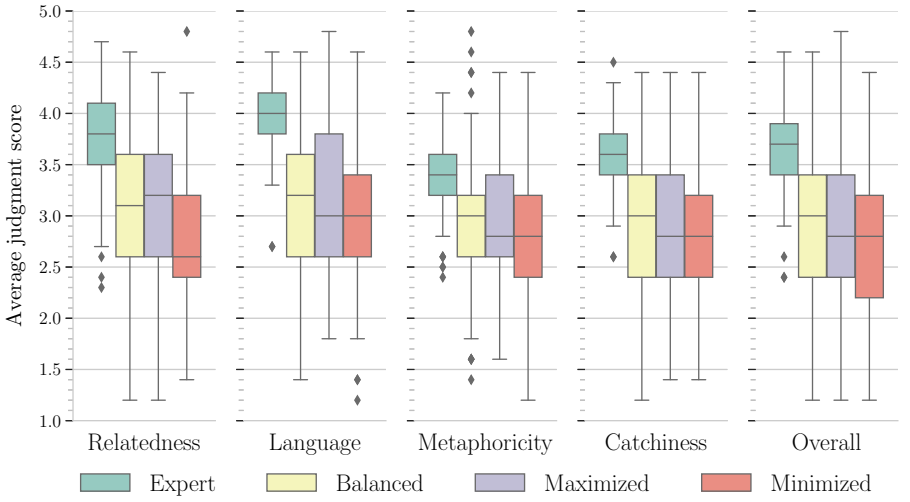

Fig. 6. Distributions of mean judgments of slogans, for expert-written as well generated ones with different selection methods (balanced, maximized, or minimized internal dimensions). Results are given separately for different human judgments (relatedness, language, metaphoricity, catchiness, and overall quality). For each judgment, the "maximized" results shown are for the case where the corresponding internal evaluation dimension was maximized by the method; the "overall" case is their aggregation. Plots indicate the median, first and third quartiles, and 95% intervals.

Permutation tests indicate statistically significant associations between the internal evaluation dimensions and the respective human judgments for relatedness ($P = 10^{-6}$), for metaphoricity ($P = 0.0033$), and for prosody/catchiness ($P = 0.046$), but not for language (correctness) ($P = 0.84$).

5.5. Different slogan selection criteria

We next consider the different selection methods (balanced, maximized, or minimized internal evaluation dimensions) as well as different human judgments of the respective slogans. An overview of the results is given in Figure 6 while more details are available in Table 8. The general overview corresponds to the observations above.

Looking at the overall judgments (last group on the right in Figure 6), we notice—as before—that slogans with balanced dimensions tend to be appreciated more than slogans with a single maximized dimension. The first four groups look at the four specific human judgments, and the "maximized" results are always given for the case where the corresponding internal evaluation dimension has been maximized. Except for the relatedness dimension (first group on the left), balancing all four dimensions actually produced better results than maximizing the respective single dimension.

Pairwise statistical permutation tests between the three groups of selection methods (balanced, maximized, minimized), for differences in the mean of the overall judgments, indicate that the balanced selection is statistically significantly better than the minimized selection ($P = 0.029$, one-tailed). These statistics confirm that slogans with balanced values on multiple dimensions improve the suitability of slogans over the case where they are minimized.

Existing, expert-written slogans stand out again with a clear margin. They received a median judgment of 3.7 for being good slogans, compared to 3.0 for the balanced computer-generated slogans. Among the different judgments of expert-written slogans, language correctness received the highest scores and had the smallest variation.

Expert-written slogans are considered to be metaphoric with a median score of 3.4, which is closer to neutral than the other judgments. At the same time, the human judgment, where

Table 8. Five-number summaries of mean judgments of slogans, grouped by different selections.

		Relatedness		Language		Metaphoricity		Catchiness		Overall	
Selection method	*n*										
Balanced dimensions											
		3.1		3.2		3.0		3.0		3.0	
pos(*r*, *l*, *m*, *p*)	262	2.6	3.6	2.6	3.6	2.6	3.2	2.4	3.4	2.5	3.4
		1.2	4.6	1.4	4.6	1.4	4.8	1.2	4.4	1.2	4.6
		3.0		3.0		3.0		2.8		2.8	
pos(*r*, *l*, *m*)	93	2.4	3.6	2.6	3.6	2.6	3.4	2.4	3.4	2.4	3.4
		1.2	4.8	1.6	4.6	1.6	4.4	1.4	4.6	1.4	4.4
		3.0		3.2		3.0		2.8		2.8	
pos(*r*, *l*, *p*)	111	2.4	3.6	2.8	3.6	2.4	3.4	2.4	3.3	2.4	3.3
		1.4	4.4	1.6	4.4	1.6	4.2	1.2	4.4	1.4	4.4
A maximized dimension											
		3.2		3.2		3.0		2.8		2.9	
max(*r*)	100	2.6	3.6	2.8	3.6	2.5	3.4	2.4	3.5	2.4	3.4
		1.2	4.4	1.6	4.8	1.2	4.0	1.4	4.4	1.2	4.4
		2.8		3.0		2.8		2.8		2.8	
max(*l*)	105	2.4	3.4	2.6	3.8	2.4	3.2	2.4	3.4	2.2	3.2
		1.2	4.4	1.8	4.8	1.6	4.2	1.4	4.4	1.4	4.8
		3.0		3.0		2.8		2.8		2.8	
max(*m*)	88	2.5	3.4	2.6	3.4	2.5	3.4	2.4	3.4	2.4	3.4
		1.4	4.6	1.6	4.6	1.6	4.4	1.4	4.4	1.4	4.4
		3.0		3.2		3.0		2.8		2.8	
max(*p*)	96	2.4	3.6	2.6	3.7	2.4	3.4	2.4	3.4	2.4	3.2
		1.2	4.2	1.4	4.4	1.4	4.8	1.4	4.4	1.4	4.6
Minimized dimensions											
		2.6		3.0		2.8		2.8		2.8	
min(*r*, *l*, *m*, *p*)	104	2.4	3.2	2.6	3.4	2.4	3.2	2.4	3.2	2.2	3.2
		1.4	4.8	1.2	4.6	1.2	4.4	1.4	4.4	1.2	4.4
Expert-written slogans											
		3.8		4.0		3.4		3.7		3.7	
	100	3.5	4.1	3.8	4.2	3.2	3.6	3.4	3.8	3.4	3.9
		2.3	4.7	2.7	4.6	2.4	4.2	2.6	4.5	2.4	4.6

Letters in the *Selection method* column reflect the four evaluation dimensions: **r**elatedness to input, **l**anguage, **m**etaphoricity, and **p**rosody. *pos*(·) denotes a positive value on all mentioned dimensions, while *min*(·) and *max*(·) indicate that the given dimension is minimized or maximized, respectively. The number of slogans evaluated is expressed as *n*.

Table 9. Slogan skeletons used in this paper, in a simplified form without grammatical relations

Skeleton (without dependency structure and trailing period)	Metaphorical origin	Good slogans	Total of slogans	Success rate
***_NOUN ,_PUNCT ***_NOUN and_CCONJ ***_NOUN	Yes	50	85	0.59
***_VERB ***_NOUN ._PUNCT ***_VERB ***_ADV	No	18	31	0.58
***_ADJ by_ADP ***_NOUN ._PUNCT ***_NOUN by_ADP ***_NOUN	Yes	24	47	0.51
***_VERB the_DET ***_ADJ ***_NOUN	Yes	10	23	0.43
***_NOUN for_ADP a_DET ***_ADJ ***_NOUN	Yes	19	46	0.41
***_VERB[a] ***_NOUN	Yes	7	18	0.39
***_VERB the_DET ***_NOUN to_PART ***_NOUN	No	5	13	0.38
***_ADJ than_ADP ***_NOUN	No	8	22	0.36
***_VERB ***_ADJ	No	7	20	0.35
The_DET ***_ADJ ***_NOUN is_VERB ***_NOUN	Yes	8	23	0.35
***_PROPN ***_VERB ***_ADJ	No	2	6	0.33
***_NOUN ***_NOUN ._PUNCT ***_VERB ***_NOUN	No	9	27	0.33
The_DET ***_NOUN of_ADP ***_NOUN	Yes	7	21	0.33
The_DET ***_ADJ ***_NOUN on_ADP ***_NOUN	Yes	13	40	0.33
***_NOUN never_ADV ***_VERB out_ADP of_ADP ***_NOUN	Yes	11	38	0.29
***_VERB your_ADJ ***_NOUN do_VERB the_DET ***_NOUN	Yes	13	48	0.27
You_PRON ca_VERB ***_ADV ***_VERB the_DET ***_ADJ ***_NOUN	No	8	31	0.26
***_VERB ***_NOUN ***_NOUN	No	6	24	0.25
***_PROPN ***_ADV	No	4	18	0.22
***_VERB ***_NOUN the_DET ***_NOUN over_ADV	No	3	16	0.19
***_NOUN ***_VERB and_CCONJ ***_VERB and_CCONJ ***_VERB	No	1	6	0.17
It_PRON ***_VERB ***_NOUN	No	3	19	0.16
Between_ADP ***_NOUN and_CCONJ ***_NOUN ***_VERB ***_NOUN	Yes	2	13	0.15
***_VERB[b] ***_NOUN	Yes	2	14	0.14
I_PRON ***_VERB ***_VERB it_PRON	No	1	12	0.08
***_NOUN ._PUNCT It_PRON ***_VERB a_DET ***_NOUN ***_NOUN	Yes	1	23	0.04

Skeletons are ordered by their success rates, that is, the ratio of suitable results to produced results.
[a]In base form.
[b]In present participle form.

computer-made slogans are closest to expert-made ones is metaphoricity. This is natural: on the one hand, metaphoricity is not a strong requirement for successful (expert-written) slogans; on the other hand, the method of this paper encourages the use of metaphors in slogans.

5.6. Differences between skeletons

Finally, we consider performance differences between skeletons. Table 9 shows all skeletons used in these experiments, along with the numbers of total and successful slogans generated from them (as per mean human judgment greater than 3). Best skeletons produced successful slogans for more than half of the time, whilst for the worst ones, less than one slogan in ten was successful. The absolute numbers of produced and successful slogans also vary, suggesting that some skeletons are easier to instantiate than others.

The method described in this paper aims to produce metaphoric slogans by construction. Are skeletons extracted from existing metaphoric slogans better at producing metaphorical slogans?

One half of the 26 skeletons originate from metaphorical slogans (cf. Table 9); 38% of slogans generated from them were considered metaphorical, compared to 31% for slogans generated from the other skeletons. In total, 35% of all generated slogans were considered to be metaphorical.

These results indicate that generating slogans using skeletons extracted from metaphorical slogans has a higher potential to produce metaphorical slogans as well. On the other hand, the proposed method appears to be capable of generating metaphorical slogans also from nonmetaphorical skeletons, even if the success rate in this respect is modest, around one third.

6. Discussion

Metaphor generation. To the best of our knowledge, our metaphor construction method is the first one based on a metaphor interpretation model. The experimental results indicate that this is beneficial: metaphorical vehicles that are more likely to have the desired interpretation, in the context of the given tenor, outperformed vehicles selected solely based on their strong association to the target property.

Nonetheless, the metaphor interpretation model only gives partial information on how a metaphor is comprehended. For instance, two examples of apt vehicle candidates produced for expressing that a *computer* is *creative* are *poet* and *music*. The interpretations can be quite different: the former suggests that a computer can produce creative artifacts, while the latter suggests that the computer is a creative artifact itself. This question is partially related to the ambiguity of the word *creative*.

The experiments show that personal vehicles (such as *poet* above) produced on average better metaphors than general nouns (such as *music*), and the effect was relatively strong (cf. Table 5). What kind of vehicles are more effective varies across slogans and the role of the vehicles in them. However, personal vehicles probably are more likely to assign human properties to the tenor, and possibly, this tends to make the metaphors better. Further analysis is required to assess the impacts of each type, given that we have utilized two different resources which could have affected the results.

While salience imbalance and similarities between the vehicle and tenor are approximated through the metaphor interpretation model, additional criteria could be considered to further assess the aptness of generated metaphors. Examples of such criteria are the ontological distance between the concepts, concreteness of the vehicle, and the novelty of the metaphor.

Skeleton-based slogan production. In the experiments, the number of good slogans, that is, slogans with a mean score greater than three, ranged from 1 to 13 per input. We consider this to be a strong result: each input resulted in at least one good slogan. This was despite artificial limitations in our experimental setting; in particular, we used only two slogan skeletons for each input, out of our pool of 26 skeletons. This limitation was introduced for ease of experimentation only, and in real use of the method in supporting ideation of slogans, a larger set of skeletons obviously should be used. This would increase not only the number of better slogans, but also the variety of slogans produced.

Our 26 skeletons varied a lot in their productivity and success rates (Table 9). The fraction of successful slogans among those generated from a single skeleton varied from 59% to 4%. It is not obvious where these differences come from. While simple expressions are easier to generate, they are not necessarily better slogans. According to our results (Table 9), the length or complexity of the skeleton is not directly reflected in its success rate. This topic, among others, deserves further study and should be considered in practical use of the method.

Regardless of the success rate of generated slogans, some skeletons are harder to instantiate than others. The slogan generation method ensures that the grammatical relations encoded in skeletons are obeyed (see Figure 1 for an example). Sometimes, however, the method is not able to instantiate a skeleton. Obviously, the grammatical complexity of a skeleton constraints the number of ways it can be filled in. The method may run into a dead end also because of its preference for related words. Recall that when a placeholder i is being filled in a skeleton, the method identifies the set $\mathcal{G}_i$ of words grammatically consistent with the words already in other placeholders of the skeleton, and its further restriction $\mathcal{R}_i \subset \mathcal{G}_i$ to words related to the target concept and property given as input. The method resorts to grammatical words in $\mathcal{G}_i$ if $\mathcal{R}_i$ is empty, and problems materialize when $\mathcal{G}_i$ is also empty. It would be possible to remedy the dead-end problem without giving up the grammatical constraints: increasing the sizes of the related spaces would provide more (related) alternatives for fillers earlier in the process, potentially leading to more (grammatical) alternatives also later on. The downside of this would be decreased relatedness of slogans to the target concept and property. This option is worth exploring further, however, since relatedness can be—and already is—measured and optimized as one of the internal evaluation dimensions.

Further variation in skeletons and slogans could potentially be obtained by generating new skeletons automatically. One could try to linguistically analyze both slogan and non-slogan expressions manually or by machine learning to highlight their differences (cf. Yamane and Hagiwara 2015; Repar *et al.* 2018; Alnajjar 2019), and then generate novel slogan-like skeletons. We leave this for future research.

Internal evaluation dimensions and human judgments. The empirical tests indicated statistically significant associations between the internal evaluation dimensions and the corresponding human judgments for relatedness, metaphoricity, and prosody/catchiness. The result suggests that these internal evaluation dimensions could be given a larger role in the design of the method. For instance, a wider selection of slogans could potentially be obtained by removing the strict coherence requirement that all words in a slogan must be related to each other (cf. Section 4.4). Instead, the method could rely more on the existing evaluation dimensions, and a measure of internal coherence could be added as a new one.

The correlation between internal evaluation and human judgment was not significant for language. This reflects the design of the method: the search space has a lot of variation in terms of relatedness, metaphoricity, and prosody, while the language is strongly bounded by the grammatical constraints of the skeletons. In addition, the internal evaluation dimension of language combines language correctness and surprise, while human judges were only asked about language correctness.

In human judgments of the four aspects of generated slogans, language correctness received better scores than relatedness, catchiness, and metaphoricity (Table 7). This speaks in favor of the grammatical constraints and their maintenance throughout the method, even if the internal language dimension was not able to reliably measure the remaining variance in quality, and even though some skeletons were not so productive due to the constraints. At the same time, more creative slogans could potentially be produced by dropping strict grammar constraints. This could, however, result in too many poor expressions, and automated assessment of their quality would be difficult.

The relatively low performance of generated slogans with respect to metaphoricity (Table 7) is somewhat surprising, given that the method is specially constructed to use metaphor. However,

by design, the method does not enforce all slogans to be metaphoric. Rather, they are encouraged to be metaphoric by primarily using words in $\mathcal{R}_i$ related to the concept or the vehicle, and by the internal metaphoricity evaluation dimension. As mentioned above, the correlation between the internal evaluation dimension and the human judgment of metaphoricity was statistically significant, allowing for optimization of metaphoricity in the results.

Looking at correlations between human judgments of different aspects of slogans (Figure 5), we observed that correlations tended to be high but that correlation between metaphoricity and relatedness was relatively low (0.37). This is probably explained by the introduction of a metaphoric vehicle and words associated to it, which decreases associations to the input concept. (Nevertheless, metaphoricity has a strong positive correlation with catchiness and overall suitability of slogans.)

Despite the abovementioned statistically significant relation between internal evaluation and relatedness, metaphoricity, and prosody/catchiness, maximizing just one internal dimension seems to only have some correspondence to the respective human judgment (Table 8). This confirms the broader observation that better slogans are obtained by a balanced mix of several internal dimensions than by a single one. High correlations between the human judgments (Figure 5) suggest that those aspects are intertwined and cannot be easily optimized in isolation.

Finally, while we have observed statistically significant associations between the degree of metaphoricity as measured by the internal dimension and by human judgment, there is no guarantee that generated slogans convey the intended metaphor and the intended property. It would be interesting to analyze the human interpretations of metaphors, both in their nominal form (i.e., purely as metaphors) and in the produced slogans. Such evaluation probably should involve open questions asking the judges to give their interpretations. Obtaining answers of sufficiently high quality could be difficult in crowdsourcing, and quantitative assessment of the answers would be difficult, too.

Resources and parameters. The method proposed in this paper makes use of multiple linguistic resources and tools, and limitations in their scopes and functionalities can have an impact on the slogans generated. The resources include well-known corpora (e.g., *ukWaC*) and tools (e.g., *NLTK* and *spaCy*), but also the more novel metaphor interpretation model *Meta4meaning* by Xiao *et al.* (2016). Metaphor interpretation is a difficult and ambiguous task, and misinterpretations by *Meta4meaning* are not unlikely, potentially resulting in metaphors conveying a meaning different from the intended one. This issue is also related to polysemy, which is not directly dealt with by our methods. Additionally, given that slogans are short and not even full sentences, NLP tools might fail in parsing them. Such failures result in building skeletons with incorrect grammatical relations, eventually affecting the generated expressions.

The slogan generation method takes multiple parameters that could be tuned to achieve better results. For instance, computation of the semantic model ω alone (cf. Section 4.2) takes parameters such as window width and frequency limits; the genetic algorithm likewise takes many parameters. More central to the slogan generation method are issues like the number of related words to consider when filling in skeletons (cf. discussion above). Reducing the number would likely result in generating fewer yet better slogans, while an increase would produce a larger variety of slogans including ones that are less related to the given concept and property. As relatedness to the advertised product and the desired property is important for slogans, the former approach seems more promising, especially if a larger selection of skeletons is used to ensure that a variety of slogans is produced.

Effects of Randomness. In this paper, there are two major uses of randomness: in generation of metaphors and slogans, and in empirical evaluation of the generation system.

Starting with empirical evaluation, selecting a random sample from the output produced by the system is a common practice in evaluation of generative methods. In our evaluation, we have

used random artifacts of several types, for example, strongly related, related, and random vehicles, as well as slogans with balanced, maximized, and minimized evaluation dimensions, in order to shed light on how the method works and what affects the quality of the output. Regarding generation of metaphors and slogans, we have two notes. First, randomness mostly takes place within stochastic search/optimization algorithms: during its operation, the method makes random decisions, but it also evaluates the decisions and either pursues the most promising ones, or selects the better ones for the next phases. Overall, the operation thus is not arbitrary while randomness is used as part of the method. Second, in most cases random selection is informed, not blind. For instance, the method carries out random selection among the top vehicles, or among the most strongly associated words, in order to provide variation and to avoid relying too heavily on computational estimates. Because of this stochasticity, we have evaluated a large number of artifacts from different aspects, to reduce random effects in the results.

Additionally, there is one major random choice in the paper: selection of which skeletons to use. As discussed earlier, we only use two random skeletons for each input, in order to make the empirical tests of this paper feasible.

Crowdsourcing and evaluation. In our evaluations, we have used a crowdsourcing platform to judge metaphors and slogans. We chose to obtain opinions of ordinary people, rather than advertising experts, because they are much easier to reach. A hard-to-mitigate risk of crowdsourcing subjective tasks that do not have unique answers is scammers, that is, users who abuse the system by answering tasks very fast, possibly just randomly, in order to maximize their income. Given that scammers add noise to the data, the signals that were detected statistically despite the noise are likely to be reliable. However, some associations may have remained undiscovered due to the noise.

Another problem with crowdsourcing was that we could not assume that the judges know and understand linguistic concepts such as metaphor, semantic relatedness, and prosody. We aimed to craft the questions in a manner that would be simple to understand and answer, but regardless of our best efforts, it is infeasible for us to verify that the judges have actually understood the task fully and answered accordingly.

The purpose of the proposed method is to act as an ideation tool for professionals when constructing slogans. We did not evaluate the method in this use case, but it would be relevant to assess if the method can actually inspire professionals. This would involve recruitment of professionals willing to test the method, further development of the method to a user-friendly tool, and design of the experimental setup. In sustained use, the tool could additionally monitor its use by the professionals, slogans selected/saved, adjustments to parameters, etc., and then estimate the relationships between parameters, internal evaluation dimensions, and the satisfaction of slogans by the users.

Given the difficulty of assessing the method with professionals, a more practical evaluation could compare generated slogans to those written by amateurs. Additional task-related relevance could be obtained by using both generated and amateur-written slogans for further ideation and development (by amateurs), and seeing how different initial slogans fare in mutual comparison.

Creativity. Generation of slogans is a creative task involving "production of a novel and appropriate response, product, or solution to an open-ended task" (Amabile 2012). It would be interesting to assess the creativity of the method, or the creativity of the slogans and metaphors produced. The field of computational creativity (Colton and Wiggins 2012; Xiao *et al.* 2019) offers conceptual tools for this. A full discussion is outside the scope of this paper, but Jordanous (2012) describes a procedure consisting of defining what creativity means in the application at hand and then deriving evaluation metrics. To instantiate the evaluation methodology to slogan generation, the example by Alnajjar and Hämäläinen (2018) could be followed, as it evaluates a related creative task.

7. Conclusions

In this paper, we have introduced a method for generating metaphorical slogans computationally, given a concept to produce a slogan for, and a property to be associated to the concept. As a subcomponent of the approach, we have also proposed a novel method for generating metaphors.

The slogan generation method uses skeletons, that is, templates with empty placeholders and grammatical constraints between them. We have described how skeletons can be extracted automatically from existing slogans, how possible (metaphorical) filler words are identified, and how the resulting slogan candidates can be assessed using four internal evaluation dimensions. We have used a genetic algorithm to construct slogans with a multi-objective fitness function based on the evaluation dimensions.

The metaphor generation method uses a metaphor interpretation model to identify metaphorical vehicles that are likely to result in the intended interpretation. To the best of our knowledge, this is the first metaphor generation method based on an interpretation model rather than just generation heuristics.

We have evaluated the proposed method and its various components using crowdsourcing. Our empirical findings can be summarized as follows:

- The method produced at least 1 good slogan for each input and up to 13 for some. Significant increase can be expected when using more skeletons instead of the two (out of 26) used per input in our experiments.
- Catchiness, relatedness to the target concept, language correctness, and metaphoricity correlate with the overall quality of slogans ($r = 0.86$, 0.74, 0.68, and 0.58, respectively), based on the evaluation on expert-made slogans. Further, the internal evaluation measures defined in this paper for relatedness, metaphoricity, and prosody/catchiness are related to the corresponding human judgments to a statistically significant degree ($P = 10^{-6}$, 0.0033, and 0.046, respectively). These results imply that it is possible to computationally measure—and thus, optimize—three criteria that contribute to the overall quality of slogans.
- Best slogans are obtained, on average, when the four internal evaluation dimensions are balanced. Maximizing just one of them tends to produce inferior results, often also for the maximized aspect.
- The productivity and success rate of individual skeletons varies considerably. The best skeletons produced an order of magnitude more slogans than poorer ones, and they produced good slogans for more than half of the time. By using the better skeletons only, the average overall quality of generated slogans can be increased considerably.
- Regarding metaphor generation on its own, using the metaphor interpretation model gives, on average, better metaphors than a corresponding method without it. Further, personal vehicles tend to produce better metaphors than vehicles of the general class.

This work has taken steps toward automated generation of metaphorical slogans, and toward generation of metaphors based on their interpretations. We hope that the methods described in this paper and our empirical observations earlier in this section help others build even better metaphor and slogan generation systems.

In future work, we will adapt the ideas presented in this paper to generation of other creative expressions. We are especially interested in producing short, catchy texts in a given textual context, such as creating attractive headlines (Gatti *et al.* 2016; Alnajjar, Leppänen, and Toivonen 2019) for automatically generated news texts (cf. Bouayad-Agha *et al.* 2012; Leppänen *et al.* 2017). Slogans tend to have no textual context, making their generation a more isolated task. Having a context adds complexity to the task, but also provides clues to completing the task. We also plan to expand

the current setting with exactly one concept and one property to handle cases of multiple concepts and details (e.g., in the news domain, comparing election results of two parties in a given city).

We are also interested in multilingual settings. While the current work only considers English, the key ideas hold for many other languages for which similar tools and resources are available.

Finally, another future direction could be altering existing texts to include some metaphoricity, extending current word-substitution-based methods for generation of creative language (Toivanen *et al*. 2012; Valitutti *et al*. 2016). After identifying a metaphorical reference topic for a given text, the method could be adjusted to replace verbs and adjectives in the text with content words from the space related to the reference topic, while maximizing the metaphoricity dimension.

Acknowledgments. We would like to thank the anonymous reviewers for their helpful comments.

Financial support. This work has been supported by the Academy of Finland under grant 276897 (CLiC) and by the European Union's Horizon 2020 programme under grant 825153 (Embeddia).

References

Alnajjar K. (2018). The 12 million most frequent English grammatical relations and their frequencies. https://doi.org/10.5281/zenodo.1255800.

Alnajjar K. (2019). *Computational Analysis and Generation of Slogans*. *Master's Thesis*, Helsingin yliopisto, Helsinki, Finland.

Alnajjar K., Hadaytullah H. and Toivonen H. (2018). "Talent, Skill and Support." A method for automatic creation of slogans. In *Proceedings of the 9th International Conference on Computational Creativity (ICCC 2018), Salamanca, Spain*. Association for Computational Creativity, pp. 88–95.

Alnajjar K. and Hämäläinen M. (2018). A master-apprentice approach to automatic creation of culturally satirical movie titles. In *Proceedings of the 11th International Conference on Natural Language Generation, Tilburg, The Netherlands*. Association for Computational Linguistics, pp. 274–283.

Alnajjar K., Hämäläinen M., Chen H. and Toivonen H. (2017) Expanding and weighting stereotypical properties of human characters for linguistic creativity. In *Proceedings of the 8th International Conference on Computational Creativity (ICCC 2017), Georgia, Atlanta, USA*. Georgia Institute of Technology, pp. 25–32.

Alnajjar K., Leppänen L. and Toivonen H. (2019). No time like the present: methods for generating colourful and factual multilingual news headlines. In *The 10th International Conference on Computational Creativity, Charlotte, North Carolina, USA*. Association for Computational Creativity, pp. 258–265.

Amabile T. (2012). *Componential Theory of Creativity*. Working Paper No. 12–096, Harvard Business School.

Baroni M., Bernardini S., Ferraresi A. and Zanchetta E. (2009). The WaCky wide web: a collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation* **43**(3), 209–226.

Bird S., Klein E. and Loper E. (2009). *Natural Language Processing with Python*, 1st Edn. O'Reilly Media, Inc.

Bouayad-Agha N., Casamayor G., Mille S. and Wanner L. (2012). Perspective-oriented generation of football match summaries: old tasks, new challenges. *ACM Transactions on Speech and Language Processing* **9**(2), 1–31.

Burgers C., Konijn E.A., Steen G.J. and Iepsma M.A.R. (2015). Making ads less complex, yet more creative and persuasive: the effects of conventional metaphors and irony in print advertising. *International Journal of Advertising* **34**(3), 515–532.

Colton S. and Wiggins G.A. (2012). Computational creativity: the final frontier? In *Proceedings of the 20th European Conference on Artificial Intelligence, ECAI'12, Amsterdam, The Netherlands*. IOS Press, pp. 21–26.

Dahl G. (2011). *Advertising for Dummies*. Hoboken, NJ: John Wiley & Sons.

De Smedt T. and Daelemans W. (2012). Pattern for Python. *Journal of Machine Learning Research* **13**, 2063–2067.

Deb K., Pratap A., Agarwal S. and Meyarivan T. (2002). A fast and elitist multiobjective genetic algorithm: NSGA-II. *IEEE Transactions on Evolutionary Computation* **6**(2), 182–197.

Dowling G.R. and Kabanoff B. (1996). Computer-aided content analysis: what do 240 advertising slogans have in common? *Marketing Letters* **7**(1), 63–75.

Evert S. (2008). Corpora and collocations. In Lüdeling A. and Kytö M. (eds), *Corpus Linguistics. An International Handbook*, Vol. **2**. Berlin: Mouton de Gruyter, pp. 1212–1248.

Fortin F.-A., De Rainville F.-M., Gardner M.-A., Parizeau M. and Gagné C. (2012). DEAP: evolutionary algorithms made easy. *Journal of Machine Learning Research* **13**, 2171–2175.

Fuertes-Olivera P.A., Velasco-Sacristán M., Arribas-Baño A. and Samaniego-Fernández E. (2001). Persuasion and advertising English: metadiscourse in slogans and headlines. *Journal of Pragmatics* **33**(8), 1291–1307.

Galván P., Francisco V., Hervás R., Méndez G. and Gervás P. (2016). Exploring the role of word associations in the construction of rhetorical figures. In *Proceedings of the Seventh International Conference on Computational Creativity (ICCC 2016), Paris, France*. Sony CSL.

Gatti L., Özbal G., Guerini M., Stock O. and Strapparava C. (2015). Slogans are not forever: adapting linguistic expressions to the news. In *Proceedings of the 24th International Conference on Artificial Intelligence, Stanford, California, USA*. AAAI Press, pp. 2452–2458.

Gatti L., Özbal G., Guerini M., Stock O. and Strapparava C. (2016). Automatic creation of flexible catchy headlines. In *"Natural Language Processing meets Journalism"—IJCAI 2016 Workshop, New York City*, pp. 25–29.

Giora R. (2003). *On Our Mind: Salience, Context, and Figurative Language*. Oxford University Press.

Glucksberg S. (2001). *Understanding Figurative Language: From Metaphor to Idioms*. Oxford University Press.

Harmon S. (2015). FIGURE8: a novel system for generating and evaluating figurative language. In *Proceedings of the 6th International Conference on Computational Creativity (ICCC 2015), Park City, Utah, USA*. Brigham Young University, pp. 71–77.

Honnibal M. and Montani I. (2017). spaCy 2: natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing. To appear.

Iwama K. and Kano Y. (2018). Japanese advertising slogan generator using case frame and word vector. In *Proceedings of the 11th International Conference on Natural Language Generation, Tilburg University, The Netherlands*. Association for Computational Linguistics, pp. 197–198.

Jordanous A. (2012). A standardised procedure for evaluating creative systems: computational creativity evaluation based on what it is to be creative. *Cognitive Computation* **4**(3), 246–279.

Katz A.N. (1989). On choosing the vehicles of metaphors: referential concreteness, semantic distances, and individual differences. *Journal of Memory and Language* **28**(4), 486–499.

Koestler A. (1964). *The Act of Creation*. London: London Hutchinson.

Kohli C., Suri R. and Thakor M. (2002). Creating effective logos: insights from theory and practice. *Business Horizons* **45**(3), 58–64.

Lenzo K. (1998). The CMU pronouncing dictionary. http://www.speech.cs.cmu.edu/cgi-bin/cmudict.

Leppänen L., Munezero M., Granroth-Wilding M. and Toivonen H. (2017). Data-driven news generation for automated journalism. In *Proceedings of the 10th International Conference on Natural Language Generation, Santiago de Compostela, Spain*. Association for Computational Linguistics, pp. 188–197.

Marcus M.P., Santorini B. and Marcinkiewicz M.A. (1993). Building a large annotated corpus of English: the Penn Treebank. *Computational Linguistics* **19**(2), 313–330.

Mathur L.K. and Mathur I. (1995). The effect of advertising slogan changes on the market values of firms. *Journal of Advertising Research* **35**(1), 59–65.

McGregor S., Agres K., Purver M. and Wiggins G. (2015). From distributional semantics to conceptual spaces: a novel computational method for concept creation. *Journal of Artificial General Intelligence* **6**(1), 55–86.

Miller D.W. and Toman M. (2016). An analysis of rhetorical figures and other linguistic devices in corporation brand slogans. *Journal of Marketing Communications* **22**(5), 474–493.

Ortony A. (1993). *The Role of Similarity in Similes and Metaphors*, 2nd Edn. Cambridge: Cambridge University Press, pp. 342–356.

Ortony A., Vondruska R.J., Foss M.A. and Jones L.E. (1985). Salience, similes, and the asymmetry of similarity. *Journal of Memory and Language* **24**(5), 569–594.

Özbal G., Pighin D. and Strapparava C. (2013). BRAINSUP: brainstorming support for creative sentence generation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, Sofia, Bulgaria*. Association for Computational Linguistics, pp. 1446–1455.

Reece B.B., Van den Bergh B.G. and Li H. (1994). What makes a slogan memorable and who remembers it. *Journal of Current Issues & Research in Advertising* **16**(2), 41–57.

Reinsch Jr. N.L. (1971). An investigation of the effects of the metaphor and simile in persuasive discourse. *Speech Monographs* **38**(2), 142–145.

Repar A., Martinc M., Žnidaršič M. and Pollak S. (2018). BISLON: BISociative SLOgaN generation based on stylistic literary devices. In *Proceedings of the Ninth International Conference on Computational Creativity, Salamanca, Spain*. Association for Computational Creativity (ACC), pp. 248–255.

Richards I.A. (1936). *The Philosophy of Rhetoric*. London: Oxford University Press.

Strapparava C., Valitutti A. and Stock O. (2007). Automatizing two creative functions for advertising. In Cardoso A. and Wiggins G. (eds), *Proceedings of the 4th International Joint Workshop on Computational Creativity, London, UK*. London: Goldsmiths, University of London, pp. 99–108.

Toivanen J.M., Toivonen H., Valitutti A. and Gross O. (2012). Corpus-based generation of content and form in poetry. In *Proceedings of the 3rd International Conference on Computational Creativity (ICCC 2012), Dublin, Ireland*, pp. 211–215.

Tom G. and Eves A. (1999). The use of rhetorical devices in advertising. *Journal of Advertising Research* **39**(4), 39–43.

Tomašič P., Papa G. and Žnidaršič M. (2015). Using a genetic algorithm to produce slogans. *Informatica* **39**(2), 125.

Tomašič P., Žnidaršič M. and Papa G. (2014). Implementation of a slogan generator. In *Proceedings of the 5th International Conference on Computational Creativity (ICCC 2014), Ljubljana, Slovenia*. Josef Stefan Institute, 340–343.

Tourangeau R. and Sternberg R.J. (1981). Aptness in metaphor. *Cognitive Psychology* **13**(1), 27–55.

Turney P.D., Neuman Y., Assaf D. and Cohen Y. (2011). Literal and metaphorical sense identification through concrete and abstract context. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing, EMNLP '11, Edinburgh, United Kingdom*. Association for Computational Linguistics, pp. 680–690.

Valitutti A., Doucet A., Toivanen J.M. and Toivonen H. (2016). Computational generation and dissection of lexical replacement humor. *Natural Language Engineering* **22**(5), 727–749.

Veale T. and Li G. (2012). Specifying viewpoint and information need with affective metaphors: a system demonstration of the metaphor magnet web app/service. In *Proceedings of the ACL 2012 System Demonstrations, ACL '12, Jeju Island, Korea*. Association for Computational Linguistics, pp. 7–12.

Veale T. and Li G. (2013). Creating similarity: lateral thinking for vertical similarity judgments. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, Sofia, Bulgaria*. Association for Computational Linguistics, pp. 660–670.

Xiao P., Alnajjar K., Granroth-Wilding M., Agres K. and Toivonen H. (2016). Meta4meaning: automatic metaphor interpretation using corpus-derived word associations. In *Proceedings of the 7th International Conference on Computational Creativity (ICCC 2016), Paris, France*. Sony CSL.

Xiao P. and Blat J. (2013). Generating apt metaphor ideas for pictorial advertisements. In *Proceedings of the 4th International Conference on Computational Creativity (ICCC 2013), Sydney, Australia*. The University of Sydney, pp. 8–15.

Xiao P., Toivonen H., Gross O., Cardoso A., Correia J., Machado P., Martins P., Oliveira H.G., Sharma R., Pinto A.M., Díaz A., Francisco V., Gervás P., Hervás R., León C., Forth J., Purver M., Wiggins G.A., Miljković D., Podpečan V., Pollak S., Kralj J., Žnidaršič M., Bohanec M., Lavrač N., Urbančič T., Velde F.V.D. and Battersby S. (2019). Conceptual representations for computational concept creation. *ACM Computing Surveys* **52**(1), 9:1–9:33.

Yamane H. and Hagiwara M. (2015). Tag line generating system using knowledge extracted from statistical analyses. *AI & SOCIETY* **30**(1), 57–67.

Žnidaršič M., Tomašič P. and Papa G. (2015). Case-based slogan production. In Kendall-Morwick J. (ed), *Proceedings of the ICCBR 2015 Workshops, Frankfurt, Germany*. CEUR, pp. 123–130.

Cite this article: Alnajjar K and Toivonen H (2021). Computational generation of slogans. *Natural Language Engineering* **27**, 575–607. https://doi.org/10.1017/S1351324920000236