No Time Like the Present: Methods for Generating Colourful and Factual Multilingual News Headlines

Khalid Alnajjar
khalid.alnajjar@helsinki.fi

Leo Leppänen
leo.leppanen@helsinki.fi

Hannu Toivonen
hannu.toivonen@helsinki.fi

Department of Computer Science and HIIT, University of Helsinki, Finland

Abstract

News headlines are the main method for briefly providing a summary of the news article and attracting an audience. In this paper, we experiment with different existing methods for producing colourful expressions and news headlines computationally, in a practical setting. Our case study is conducted by modifying an automated journalism system that generates multilingual news in three languages, namely English, Finnish and Swedish. We adapt existing methods for creative headlines and figurative language generation into the headline generation process of the system, modifying them to work in a multilingual setting. We conduct our evaluation by asking online judges to assess the original titles produced by the unmodified system and those enhanced by the methods described in this paper. The results of the evaluation suggest that the presented methods increase the creativity of existing headlines while maintaining their descriptiveness.

Introduction

The interest in automated journalism has increased in the past years, driven by the ability to produce tailored stories cost-effectively even for small audiences, i.e., the so-called long tail effect. Current methods for automated news generation typically utilize linguistic templates written by journalists, and fill them in with appropriate information from structured data sources. Template-based methods give strong control over the output generated by the system and ensures conveying the message as intended. However, news produced by these approaches tend to be repetitive and sound mechanistic.

Headlines are an essential part of the news. They must relate to the news article and briefly describe it while motivating readers to visit and read the article. Automated journalism systems aim to produce informative headlines, but not colourful ones requiring creativity.

In this paper, we experiment with different existing methods for creating colourful expressions and with their use in template-based news headlines. We seek for a balance between creativity and factuality. Because of the latter, we build on an existing template-based system that produces factual headlines; for the former, we generate creative expressions and add them to the factual headlines.

For our case study, we use a modified version of *Valtteri*[1] (Leppänen et al. 2017), an automated journalism system, as the baseline. *Valtteri* generates election news about the 2017 Finnish municipal election results in three languages, English, Finnish and Swedish. For the scope of this work, we focus on two languages only: English and Finnish.

Creativity, such as use of figurative language, is something human journalists consider to be one of their strengths when compared to automated journalism systems (van Dalen 2012). Creativity is missing from most, if not all, automated journalism systems is creativity. This also applies to *Valtteri*.

Inspired by previous research on generating figurative language (Veale and Li 2013; Alnajjar et al. 2017) and creative headlines (Lynch 2015; Gatti et al. 2015), we present two methods which add a creative touch to news headlines generated by the automated journalism system. The methods are developed to operate in a multilingual setting. The first method finds a suitable well-known phrase (e.g. movie title) to be presented to the reader as a catchy title (i.e. it draws attention) along with the factual message. The other method injects figurative phrases (e.g. similes and metaphors) into headlines, depending on the polarity of the news. We exploit recent research in word cross-lingual embeddings, permitting us to project knowledge from English, with rich linguistic resources, into a less-resourced one, i.e. Finnish.

In our evaluation, we crowdsourced the assessment of the headlines to online judges (acting as the audience). We asked them to evaluate the original headline produced by *Valtteri* and the new altered headlines produced by the methods described in this paper in order to test the applicability of these methods in a practical scenario. The judges were asked to assess aspects such as informativeness, correctness and catchiness, to measure the effects of figurative modifications on the original headlines. Because of the availability of crowdsourcing workers, the current evaluation is conducted on English headlines only and Finnish is left for future work.

This paper is structured as follows. We begin by reviewing related work on headline generation. Thereafter, we describe the *Valtteri* system and how the creative component is attached to the system. We then elucidate the methods employed by us to convert the headlines generated by *Valtteri*

[1] https://www.vaalibotti.fi/

into more colourful ones. The evaluation details are then provided, followed by the results. Lastly, we discuss the results and conclude this work.

Related Work

Previous research on headline generation is extensive, covering different approaches based e.g. on rules, statistics, summarization or machine learning. In this section, we briefly describe the most relevant work.

Hedge Trimmer (Dorr, Zajic, and Schwartz 2003), a rule-based method for headline generation, decides which words are to be retained and which to be pruned from the news article. Their rules are linguistically motivated and based on analyzing human-made headlines written in English. Building such rules is tedious, especially when dealing with multilingual news articles. Wang, Dunnion, and Carthy (2005) extended the work by introducing a C5.0 decision tree classifier for predicting which words to include in the title.

Zajic, Dorr, and Schwartz (2002) use a Hidden Markov Model to generate news headlines for a news story by having the model capture keywords from the beginning, i.e. first paragraphs, of the story. A Viterbi Decoding algorithm is then applied to headlines generated by the model to find the most representative headline. Additionally, four decoding parameters are imposed to ensure the quality of the generated headline, namely: (1) a length penalty, to keep headlines within the 5 to 15 word length limits, (2) a position penalty, to give a higher penalty to words appearing later in the story, (3) a string penalty, to encourage neighbouring words and (4) a gap penalty, to reduce the distance between selected words. Another statistical approach (Colmenares et al. 2015) uses sequence prediction methods for learning how humans craft headlines. Given a story, their model classifies whether a certain token in the story should be in the headline or not. In the case of a token being classified as in-headline, their method considers various features regarding the text of the story, the token (e.g. parts-of-speech tags and name-entities) and the constructed headline at each stage. Other statistical-based research on headline generation has been conducted by Banko, Mittal, and Witbrock (2000), Knight and Marcu (2002), Wan et al. (2003) and Unno et al. (2006).

Summarization-based techniques treat the problem of headline generation as producing a one sentence digest of the article (Morita et al. 2013; Martins and Smith 2009; Filippova 2010). Summarization techniques tend to extract and then compress sentences existing in the new article, which results in reusing words/phrases existing in the article. Furthermore, deep learning models have also been employed in the generation of headlines by learning how to summarize a certain text (Ayana et al. 2016). Such models require sufficiently big training data sets which can be prohibitively large for some scenarios.

A way of expressing headlines in various styles is to learn different ways of talking about the same news article. Wubben et al. (2009) have proposed a way of grouping news articles from different sources based on the content similarity. Using the different ways of writing a headline for a certain topic, a machine translation model could be trained to learn how to paraphrase headlines (Wubben, Bosch, and Krahmer 2010). *HEADY* (Alfonseca, Pighin, and Garrido 2013), on the other hand, performs event pattern clustering and generates a headline for an unseen news article by inferring headlines based on the events in it.

The above approaches do not consider an important aspect of news headlines, which is catchiness. To our best knowledge, catchiness in news headlines generation is addressed only in the work by Lynch (2015) and Gatti et al. (2015).

Lynch (2015) proposed a system for adding a well-known phrase (e.g. songs, films ...etc) as a prefix to an existing title. The added phrase is intended to catch the attention of the readers and increase search engine optimization. The system extracts keywords from an article, clusters and expands them. Then, it pairs keywords from distinct clusters if they co-occurred in a corpus of 5-grams. Using a pseudo-phonetic string matching algorithm and semantic similarity measurement, the system finds and ranks well-known phrases suitable for the pair. Lastly, it embeds the matched well-known phrase in the existing headline.

In the method described by Gatti et al. (2015), titles are given a creative touch by blending them with well-known expressions. Their headline generation process extracts keywords from the input news article. Thereafter, the method finds existing well-known phrases that are semantically similar to the existing headline and the article. These phrases are then modified by altering a word in them that satisfies a semantic similarity threshold, and lexical and syntactic constraints.

Despite the advances in automated headline generation, research on generating catchy and diverse headlines for automated journalism is scarce, especially in a multilingual setting with less-resourced languages.

Adding Creativity to Valtteri Headlines

Valtteri (Leppänen et al. 2017; Melin et al. 2018) is a multilingual system for automated journalism, reporting on the 2017 Finnish municipal elections. The system follows a data-driven approach to generate news while ensuring certain requirements, e.g. accuracy (i.e. factual and not misleading) of the produced news.

We add the creativity component to the system at a central stage of the pipeline, immediately after the aggregation process. It has access to the data and the selected templates to be used in the news. The component can alter the content of the news article produced along with its headline.

Inspired by existing research on computational linguistic creativity and creative headline generation, we implement two methods for producing colourful headlines. The methods are:

1. **Phrase-copying**: We find and insert a suitable well-known phrase into a factual headline (Lynch 2015; Gatti et al. 2015).
2. **Figurative-injection**: We generate figurative expressions using linguistic patterns and knowledge-bases of stereotypical properties of nouns (Veale and Li 2013; Alnajjar et al. 2017), and insert them into existing headlines.

For our use case, these methods should be incorporated in the automated journalism system and they should work

in multiple languages. To achieve this, in case the required linguistic resources are not available for Finnish, we resort to pre-trained and aligned multilingual word embeddings ζ (Bojanowski et al. 2017; Joulin et al. 2018). In these models, a vector representation of a word in a certain language (e.g. *king* in English, ζ_{en}) should roughly point to the same semantic direction in another model (e.g. *kuningas* in Finnish, ζ_{fi}) and vice versa. With the help of these aligned models, we can exploit available linguistic creativity resources in English and project them into Finnish.

The following sub-sections describe the two methods for colourful headline generation in-depth.

Phrase-copying: Insertion of Well-Known Phrases

Inspired by the research by Lynch (2015) and Gatti et al. (2015), we implement a method for finding and inserting well-known phrases into headlines produced by *Valtteri*. The results have the form "*phrase*: *headline*", c.f. Table 1 for examples. Juxtapositioning the phrase with the headline is expected to catch the attention of viewers and motivate them to click on the headline to read the news article, while keeping the factual content of the original headline intact. For this to work as intended, the method should find a well-known phrase that matches the original headline.

We use two types of well-known phrases: proverbs and movie titles. Proverbs for each language are extracted from `wikiquote.org`[2]. Regarding movie titles, we use the dataset of movies provided by IMDB[3]. We restrict the dataset to movies with more than 100,000 votes, to exclude generally unfamiliar titles. As these titles are in English and we desire to know how they are known to people in other languages, we query *Wikipedia* with the movie title in English and retrieve the title of its corresponding Finnish Wikipedia article. As an example, the movie title "Harry Potter and the Philosopher's Stone" is known to Finns as "Harry Potter ja viisasten kivi".

We perform a preprocessing step on the collected phrases to clean and expand them. The process commences by striping punctuation and any parentheses including the content in them to omit some explanations given in the proverbs. We also removed phrases containing more than 5 words to avoid lengthy headlines that could distract the audience. Some movie titles separate a general title and a subtitle by a colon or a dash; we include in our dataset both the short version (before the colon/dash) and the long version (all of the text).

In total, the database of well-known phrases contains 1,744 and 1,322 phrases in English and Finnish, respectively. We denote this database by P.

In order to identify a well-known phrase that matches the headline, two aspects are checked: 1) semantic similarity (or relatedness) between the phrase and the headline, and 2) prosody of the phrase and the headline. Semantic similarity is used for coherence of the resulting combination, while prosody is evaluated to increase catchiness of the result.

We employ a greedy algorithm to match phrases to a given headline H. For each phrase ρ in P, the method computes the cosine semantic similarity between individual words w_1, w_2 in ρ and H, using the corresponding language model ζ_l, where l is either 'en' or 'fi', as follows:

$$sim_{words}(w_1, w_2, t, l) = \begin{cases} \zeta_l(w_1, w_2), & \text{if } \zeta_l(w_1, w_2) \geq t \\ 0, & \text{otherwise} \end{cases} \quad (1)$$

$$sim_{phrases}(H, \rho, t, l) = \sum_{w \in H} \sum_{k \in \rho} sim_{words}(w, k, t, l) \quad (2)$$

In equations 1 and 2, t is a threshold for the minimum semantic similarity desired. We empirically set t to 0.3. While increasing the threshold t would find phrases that are more semantically similar to the headline, it would also narrow the space of possible solutions, especially for languages other than English. If a phrase has received a semantic similarity score on Equation 2 greater than 0, it is considered to be similar to the headline.

When a phrase ρ is found to be semantically similar to headline H, the method computes four prosody features between the matched phrase and the headline. These features are assonance, consonance, alliteration and rhyme. We utilize the *espeak-ng tool*[4] to acquire the International Phonetic Alphabet (IPA) transcriptions of words. The tool supports producing IPA for multiple languages, including English and Finnish. For each word pair in $\rho \times H$, the method evaluates whether the pair has phonetic similarity of any of the four prosody features. Then, each prosody feature is aggregated to phrase-level by computing its average score over the word pairs in $\rho \times H$.

Finally, we aggregate all four features into one number by obtaining a weighted sum of their phrase-level scores. We assigned to rhyme and alliteration weights of 40% each, and to consonance and assonance weights of 10% each, based on empirical testing.

Given a headline H, the method considers phrases ρ in P in a random order, computing the above measures of semantic similarity and prosody. The method keeps progressing until it finds ten well-known phrases that have a positive semantic similarity and a positive phonetical similarity with headline H. Among the ten phrases, the method then picks the phrase with the highest prosody score. The magnitude of the semantic similarity is not considered, since the template-based headline generation method tends to give the highest semantic similarity to the same phrases; for prosody, there is more variation based on how the template has been instantiated.

Finally, the selected phrase is inserted into the headline. For headline examples generated by this method, see Table 1.

[2] English: `https://en.wikiquote.org/wiki/English_proverbs_(alphabetically_by_proverb)`
Finnish: `https://en.wikiquote.org/wiki/Finnish_proverbs`

[3] `https://datasets.imdbws.com/`

[4] `https://github.com/espeak-ng/espeak-ng`

#	Baseline	Phrase-copying	Figurative-injection
(1)	Most seats go to The Centre Party of Finland in Kangasniemi	Legends of the Fall: Most seats go to The Centre Party of Finland in Kangasniemi	Most seats go to The Centre Party of Finland, the free queen, in Kangasniemi
(2)	Biggest vote gains for The Green League in Kuopio	Alls well that ends well: Biggest vote gains for The Green League in Kuopio	Biggest vote gains for The Green League –the lovely god– in Kuopio
(3)	Biggest gains for The Christian Democrats across Lapin vaalipiiri	The Running Man: Biggest gains for The Christian Democrats across Lapin vaalipiiri	Biggest gains for The Christian Democrats, as powerful as a soldier, across Lapin vaalipiiri
(4)	Second largest gains for The Christian Democrats in Rovaniemi	The Transporter: Second largest gains for The Christian Democrats in Rovaniemi	Second largest gains for The Christian Democrats –the king– in Rovaniemi
(5)	The Finns Party lose three seats in Jyväskylä	To each his own: The Finns Party lose three seats in Jyväskylä	Like a spy, The Finns Party lose three seats in Jyväskylä

Table 1: Five examples of generated headlines from an existing headline by the two presented methods in this paper, in English.

Figurative-injection: Generation of Figurative Language

The figurative-injection method inserts figurative language (e.g. metaphors and similes) into existing headlines. See the column 'Figurative-injection' of Table 1 for examples of headlines generated by this method. We next describe the method.

If the given headline has polarity with respect to the main entity in the headline, a political party or candidate in our case, then the method adds a figurative comparison to an adjective and common noun that is stereotypically associated with the polarity. The aim is that this comparison indirectly attributes properties to the entity of the headline, thereby emphasizing the polarity in a creative, figurative way.

Given that the automated journalism system works with structured data and given templates, we can directly associate polarities with the templates and values used to populate them, and avoid the need for automated polarity analysis of headlines. The polarity is determined by inspecting the reported result (i.e. the gains or losses of votes and seats) in the headline, while taking negations into account. In the cases where the headline states that an entity has received a positive result (e.g. majority of votes, biggest gains ... etc) or negative result (e.g. no seats, lose *X* seats ... etc) it is classified accordingly; otherwise, it is considered to be neutral. Neutral headlines are not modified by this method.

Identification of suitable adjectives and common nouns proceeds in three steps, performed once as a pre-processing step. First, we have manually listed seed nouns that match the election domain (e.g. win, success; loss, defeat). Second, we use corpus-based methods to identify adjectives associated to the seed nouns (e.g. heroic; tragic). Third, we identify common nouns that are stereotypically associated to these adjectives, using an existing knowledge base. We next detail these steps.

First, we manually define a set of seed nouns describing each of the polarities:

- *positive*: win, gain, accomplishment, success, achievement
- *negative*: loss, defeat, failure

Second, using the seed words, we mine stereotypical properties related to them. We observe trigrams in Google N-Grams (Brants and Franz 2006) that match the linguistic pattern "a/n * *SEED*", where SEED is any of the seed words, as conducted in previous research by Veale and Li (2013). We retrieve the adjectival properties that occur at the wildcard position ("*") in such trigrams. We then use the resource by Alnajjar et al. (2017) to prune out noisy and non-adjectival relations (e.g. "a *3-5* win"). Examples of mined properties for the two categories are: "a *heroic* achievement" and "a *tragic* loss".

Some positive adjectives can be associated with negative situations (e.g. "a *great* loss"). We use the polarity function provided in *Pattern* library (De Smedt and Daelemans 2012) to predict the polarity of adjectives, and we filter out any adjectival property that has a polarity which does not match the intended classification.

Third, the method looks for suitable metaphorical nouns (common nouns in our case) that are strongly associated with the desired properties. For this, we use a tested dataset κ of nouns and their weighted stereotypical properties (Alnajjar et al. 2017). An example of a noun and its stereotypical properties along with their weights is *King*: {powerful: 1563, successful: 1361, ... etc}.

Given a headline to modify, the method now has access to knowledge of which properties describe a positive or negative situation and which nouns are well-known to possess these properties. The method then searches for a suitable metaphorical noun to be introduced in the headline. It does so by iterating over all the properties describing the situation and the common nouns in κ to find out which nouns are associated to many of these properties. In the process, the method keeps track of all these nouns and how strongly they are related to the relevant properties in knowledge-

base κ. Thereafter, the nouns are sorted based on the sum of their association weights. A random noun having a total weight above the third quartile of weights is selected to be the metaphorical noun. A random stereotypical property of the selected noun is then chosen while ensuring that it meets two constraints: 1) it is strongly associated with the noun (i.e. in the top 50%) and 2) it describes the situation. The selected noun and its property will be used, in the remainder of this method, to construct a figurative expression.

The knowledge-base κ and the linguistic pattern used to find adjectival properties are in English but we desire to generate figurative language in multiple languages. To overcome this obstacle, we employ aligned word embedding models between multiple languages (English and Finnish) as follows. When the method is requested to generate a figurative expression for a language other than English, it begins by using the trigrams and knowledge available in English to find suitable a suitable noun and property. Once a noun and a property are selected, the method obtains their vector representations in the English model. These vectors are then projected into the other aligned model (i.e. Finnish). We consider the closest word to the projected vector as the representation of the word in the other language.

To realize a figurative expression using the selected metaphorical noun and property, we hand-crafted a set of figurative templates in both languages, given in Table 2. For each template, we define whether the template should be injected in the headline before or after the name of the entity. Depending on the position of the entity's name in the headline, a random figurative template is chosen.

English	Finnish	Position
, as *PROPERTY* as [a\n] *NOUN*,	, *PROPERTY* kuin *NOUN*,	after
, the *NOUN*,	, *NOUN*,	after
–the *NOUN*–	–*NOUN*–	after
, the *PROPERTY NOUN*,	, *PROPERTY NOUN*,	after
–the *PROPERTY NOUN*–	–*PROPERTY NOUN*–	after
Like [a\n] *NOUN*,	Kuin *NOUN* konsanaan,	before
Like [a\n] *PROPERTY NOUN*,	Kuin *PROPERTY NOUN*,	before

Table 2: Hand-crafted figurative templates in English and Finnish to be injected in existing headlines. The position column indicates whether the template should be injected before or after the entity name.

Finally, the chosen template gets filled with the selected noun and property. To ensure producing grammatically correct metaphorical expressions, we use *Pattern* to reference nouns and properties correctly, for English. Regarding Finnish, we analyze and inflect the projected words in the Finnish space into the nominative form, if necessary, using *UralicNLP* (Hämäläinen 2019) and *Omorfi* (Pirinen 2015).

Evaluation

We asked online judges on `figure-eight.com` to evaluate both the baseline (non-creative) headlines produced by *Valtteri* and the modified (creative) headlines by the methods described above. As Finnish is not supported by the crowdsourcing platform, we only evaluated English headlines at this stage.

Our evaluation dataset is constructed as follows. We randomly selected a pair of a location and an entity in Finland and passed them to *Valtteri* to obtain the news article covering the election results of the entity in that location, in English. For locations, we only considered the ones on country, district or municipality levels, to exclude news for small areas. In case the reported news by *Valtteri* was classified to be neutral in its polarity, then another random pair was selected. This process was repeated until we had 100 news articles.

The headline of each generated news article was then passed to the creativity component, which generated two modified headlines using the two presented methods. Table 1 shows examples of headlines generated by the methods.

Overall, the evaluation dataset contains 300 English headlines: 100 from the baseline system and 100 generated by both methods. We asked 10 online judges to evaluate each headline. Judges were given a brief description of the task, and the first paragraph of the news story generated by *Valtteri*. They were then asked to evaluate the headline on a 5-point Likert scale against the following claims:

1. The headline is descriptive of the article.
2. The headline is grammatically correct.
3. The headline is catchy.
4. The headline is creative.
5. The headline can be considered offensive.
6. The headline is generated by a computer.

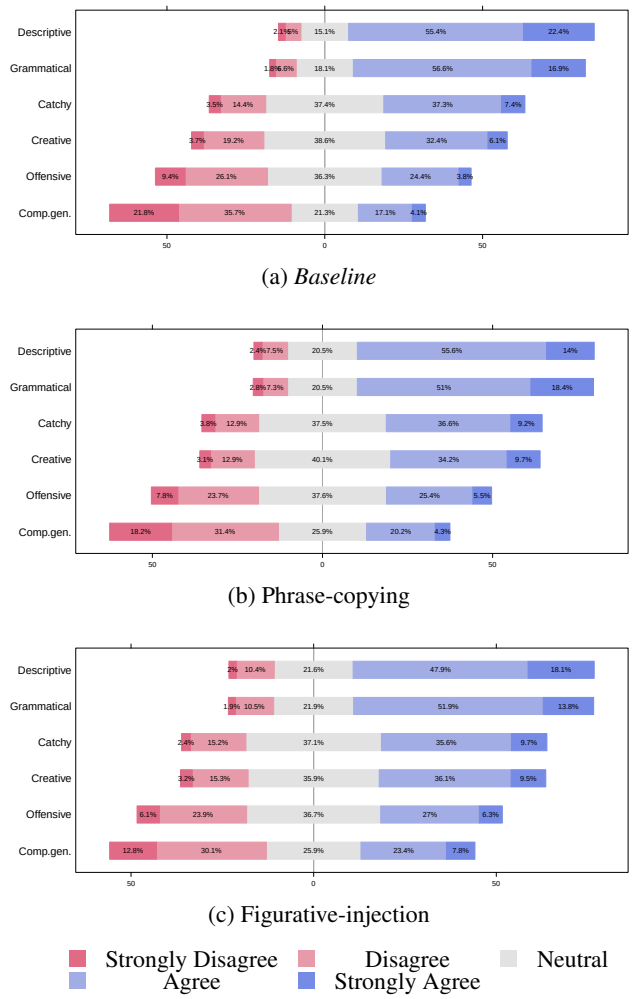
Some of these perspectives are from the prior research by Lynch (2015) and they should be self-explaining.

The quality control mechanism enforced in crowdsourcing was that a minimum of 10 seconds was spent in answering questions about five headlines, in order to eliminate spammers that answer them randomly. We did not apply other measures since the questions and interpretations are subjective and do not have correct answers.

Results

The evaluation process resulted in 3,000 unique judgments from crowdsourcing, 1,000 for each type of headlines. Table 3 shows the mean and standard deviation of judgments received on each question for the three types of headlines, and Figure 1 gives the diverging bar charts for the answers. We next look at the results for each property assessed.

Descriptive From the results, it appears that the three types of generated headlines were considered to be descriptive on average (i.e. $\mu_x > 3$). Despite all versions of the headline having the same factual message present, the headlines produced by *Valtteri* (the baseline) were judged to be the most descriptive. This difference is statistically significant.



(a) *Baseline*

(b) Phrase-copying

(c) Figurative-injection

Figure 1: Diverging bar charts illustrating the percentage of judgments received on each question for the three types of headlines.

Grammatical The results concerning the grammaticality of produced headlines is similar to the results on their descriptiveness. That is, all methods produced grammatically correct headlines on average, with headlines produced by the baseline being statistically significantly the most grammatically correct.

Catchy In terms of the catchiness of the headlines, both non-baseline methods have slightly improved the catchiness of original headlines. The difference, however, is not statistically significant.

Creative Regarding the creativity of headlines, the phrase-copying method is perceived to be the most creative, on average. The judges deemed both non-baseline methods to be more creative than the baseline (with statistical significance), as both methods have increased the agreements by approximately 6%.

Offensive Headlines produced by the non-baseline methods are more likely to produce offensive headlines. The difference between the baseline and the proposed methods is statistically significant. However, headlines are generally neutral and not offensive (i.e. $\mu_x \leq 3$).

Generated Headlines produced by the non-baseline methods are considered to be computer-generated more often than the ones generated by the baseline methods, to a statistically significant degree. However, headlines produced by all variations could pass as being written by humans as the majority of judges believed that they are not generated by computers.

Discussion

The aim of the proposed methods was to add creative language to news headlines, in order to add variation to them and to make them more interesting for readers.

According to our empirical results, the proposed methods indeed improved the creativity of the original headlines produced by *Valtteri*. This shows that the methods had some success in making the headlines more creative.

By adding creative elements, we also aimed to make the headlines more catchy. Here the methods were only slightly successful: catchiness was improved marginally. This result shows that creativity does not necessarily improve catchiness in the case of headline generation.

The modified headlines lost some of the descriptiveness of the original headlines, indicating that the added elements did not match the contents of the headline or the news story. In the case of the phrase-copying method, the main problem seems to be that despite our aim to choose phrases that are semantically related to the original headline, the added phrases can still be poorly chosen. Our measure of semantic similarity considers relations between individual words, but does not in any way take into account the meanings or mental images of the phrases as a whole. Adding a phrase with polarity matching the polarity of the headline could help, but more work is needed to make better use of well-known phrases given their rich, cultural meanings and interpretations. For the figurative-injection method, the result suggests that the selection of nouns and adjectives, but also the design of the templates used to inject figurative expressions, should be improved.

The modified headlines also lost some of their grammatical correctness. This is somewhat surprising for the phrase-copying method whose results consist of a well-known phrase and the original headline. Technically speaking, one would expect these to be grammatically about equally correct with the original headlines. A possible explanation is that the decrease in perceived grammatical correctness is influenced by poor matching of the added phrase and the original headline, as discussed above. An alternative cause is that the judges did not recognize all "well-known" phrases and therefor did not see the (grammatical) point in the generated headline. In the case of the figurative-injection method, the result implies again that the templates used to inject figurative expressions should be improved for grammatical fluency.

	Descriptive		Grammatical		Catchy		Creative		Offensive		Comp.gen.	
	μ_x	SD	μ_x	SD	μ_x	SD	μ_x	SD	μ_x	SD	μ_x	SD
Baseline	**3.91**	0.87	**3.80**	0.86	3.31	0.93	3.18	0.94	**2.46**	1.13	**2.87**	1.01
Phrase-copying	3.75*	0.93	3.71*	0.88	**3.35**	0.95	**3.35***	0.93	2.61*	1.12	2.97*	1.01
Figurative-injection	3.70*	0.95	3.65*	0.91	**3.35**	0.93	3.33*	0.95	2.83*	1.15	3.04*	1.00

Table 3: The mean μ_x and standard deviation SD of judgments received for each type of generated headlines on the six questions. The best result for each question appears in boldface.
* The value is statistically significantly different ($p < 0.05$) from the value for the baseline headline (non-parametric permutation test with one hundred million repetitions, one-tailed, not corrected for multiple testing).

We also assessed whether the modified headlines are more likely to be offensive than the original headlines. This turned indeed to be the case. By inspecting the headlines which were considered to be the most offensive, we noticed that they were usually negative expressions generated by the figurative-injection method. By construction, the method compares a party or person to a common noun, and therefor negative analogs easily become offensive to the involved party. The two most offensive headlines are 1) "No seats for The Christian Democrats, the thief, in Nousiainen" and 2) "Like a fool, The Finns Party drop most seats in Mynämäki". This result and examples highlight that care needs to be taken when using automated creativity methods to talk about persons (or parties), in order to avoid unintentional offensive expressions. The proposed methods could be modified to reduce the chances of producing offensive outputs as follows: 1) introduce a dictionary of taboo words to filter out risky words or well-known phrases containing them and 2) use a lower threshold when searching for metaphorical nouns, in order to allow for a wider selection of (safe) words. A better but bigger change would be to produce figurative comparisons to the events in the news, such as loss of seats, rather than to the persons or parties involved. Nevertheless, the final output cannot be guaranteed to be safe for production unless it is verified by a human.

Finally, the modified headlines generated by the proposed methods were recognized to be computer-generated more often than the original (computer-generated) headlines. This suggests that the methods to select and inject materials need to be improved, as the eventual goal is produce headlines that appear less computer-generated than the baseline method.

Conclusion

In this paper, we have presented methods for modifying an existing headline generation method, in order to give the headlines a creative touch. The methods work by inserting well-known phrases or figurative language in the headline templates. In our use case, we extended the headline generation method of *Valtteri*, a system that generates news reports on the 2017 Finnish elections in English, Finnish and Swedish. We also described how the methods can utilize cross-lingual links between Wikipedia articles and aligned multilingual word embedding models in order to take advantage of English resources when producing Finnish headlines, but this aspect was not evaluated due to lack of crowdsourcing workers.

Our empirical evaluation using English headlines generated by the proposed methods shows that they made the headlines more creative, and also slightly more catchy, but at the same time we observed a decrease in how descriptive and grammatically correct the headlines are.

In future work, we plan to improve the methods to select and inject materials to headlines, taking better into account the implied meanings of the added phrases or expressions, as well as making the results linguistically more fluent. Evaluation of Finnish headlines will help assess how well the cross-lingual aspects of the methods work. Interesting topics for future work also include automatic extraction of templates for injection of figurative expressions, and production of apt, yet ethically appropriate, figurative expressions. Finally, it would be interesting to introduce figurative language in the body of automatically generated news, not only headlines.

Acknowledgments

This work has been supported by European Union's Horizon 2020 research and innovation programme under grant agreement No 825153, project EMBEDDIA (Cross-Lingual Embeddings for Less-Represented Languages in European News Media).

References

Alfonseca, E.; Pighin, D.; and Garrido, G. 2013. Heady: News headline abstraction through event pattern clustering. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, 1243–1253. Sofia, Bulgaria: Association for Computational Linguistics.

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*, 25–32. Atlanta, United States: Georgia Institute of Technology.

Ayana; Shen, S.; Liu, Z.; and Sun, M. 2016. Neural headline generation with minimum risk training. *CoRR* abs/1604.01904.

Banko, M.; Mittal, V. O.; and Witbrock, M. J. 2000. Headline generation based on statistical translation. In *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*, ACL '00, 318–325. Hong Kong: Association for Computational Linguistics.

Bojanowski, P.; Grave, E.; Joulin, A.; and Mikolov, T. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics* 5:135–146.

Brants, T., and Franz, A. 2006. *Web 1T 5-gram Version 1. Linguistic Data Consortium, Philadelphia, PA.* Philadelphia, PA.

Colmenares, C. A.; Litvak, M.; Mantrach, A.; and Silvestri, F. 2015. Heads: Headline generation as sequence prediction using an abstract feature-rich space. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 133–142. Denver, Colorado: Association for Computational Linguistics.

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

Dorr, B.; Zajic, D.; and Schwartz, R. 2003. Hedge trimmer: A parse-and-trim approach to headline generation. In *Proceedings of the HLT-NAACL 03 Text Summarization Workshop*.

Filippova, K. 2010. Multi-sentence compression: Finding shortest paths in word graphs. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, 322–330. Beijing, China: Coling 2010 Organizing Committee.

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*, IJCAI'15, 2452–2458. AAAI Press.

Hämäläinen, M. 2019. Uralicnlp: An NLP library for Uralic languages. *Journal of Open Source Software* 4(37):1345.

Joulin, A.; Bojanowski, P.; Mikolov, T.; Jégou, H.; and Grave, E. 2018. Loss in translation: Learning bilingual word mapping with a retrieval criterion. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2979–2984. Brussels, Belgium: Association for Computational Linguistics.

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

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*, 188–197. Santiago de Compostela, Spain: Association for Computational Linguistics.

Lynch, G. 2015. Every word you set: Simulating the cognitive process of linguistic creativity with the PUNdit system. *International Journal of Mind Brain and Cognition* 6(1-1).

Martins, A. F. T., and Smith, N. A. 2009. Summarization with a joint model for sentence extraction and compression. In *Proceedings of the Workshop on Integer Linear Programming for Natural Langauge Processing*, ILP '09, 1–9. Boulder, Colorado: Association for Computational Linguistics.

Melin, M.; Bäck, A.; Södergård, C.; Munezero, M. D.; Leppänen, L. J.; and Toivonen, H. 2018. No landslide for the human journalist-an empirical study of computer-generated election news in finland. *IEEE Access* 6:43356–43367.

Morita, H.; Sasano, R.; Takamura, H.; and Okumura, M. 2013. Subtree extractive summarization via submodular maximization. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, 1023–1032. Sofia, Bulgaria: Association for Computational Linguistics.

Pirinen, T. A. 2015. Development and use of computational morphology of Finnish in the open source and open science era: Notes on experiences with omorfi development. *SKY Journal of Linguistics* 28:381–393.

Unno, Y.; Ninomiya, T.; Miyao, Y.; and Tsujii, J. 2006. Trimming CFG parse trees for sentence compression using machine learning approaches. In *Proceedings of the COLING/ACL on Main Conference Poster Sessions*, COLING-ACL '06, 850–857. Sydney, Australia: Association for Computational Linguistics.

van Dalen, A. 2012. The algorithms behind the headlines. *Journalism Practice* 6(5-6):648–658.

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*, 660–670. Sofia, Bulgaria: Association for Computational Linguistics.

Wan, S.; Dras, M.; Paris, C.; and Dale, R. 2003. Using thematic information in statistical headline generation. In *Proceedings of the ACL 2003 Workshop on Multilingual Summarization and Question Answering - Volume 12*, MultiSumQA '03, 11–20. Sapporo, Japan: Association for Computational Linguistics.

Wang, R.; Dunnion, J.; and Carthy, J. 2005. Machine learning approach to augmenting news headline generation. In *Companion Volume to the Proceedings of Conference including Posters/Demos and tutorial abstracts*.

Wubben, S.; van den Bosch, A.; Krahmer, E.; and Marsi, E. 2009. Clustering and matching headlines for automatic paraphrase acquisition. In *Proceedings of the 12th European Workshop on Natural Language Generation*, ENLG '09, 122–125. Athens, Greece: Association for Computational Linguistics.

Wubben, S.; Bosch, A. v. d.; and Krahmer, E. 2010. Paraphrasing headlines by machine translation: Sentential paraphrase acquisition and generation using google news. *LOT Occasional Series* 16:169–183.

Zajic, D.; Dorr, B.; and Schwartz, R. 2002. Automatic headline generation for newspaper stories. In *Workshop on Automatic Summarization*, 78–85. Philadelphia, PA, USA: Association for Computational Linguistics.