EMBEDDIA

Cross-Lingual Embeddings for Less-Represented Languages in European News Media

Research and Innovation Action
Call: H2020-ICT-2018-1
Call topic: ICT-29-2018 A multilingual Next generation Internet
Project start: 1 January 2019 Project duration: 36 months

D5.1: Datasets, benchmarks and evaluation metrics for multilingual text generation (T5.4)

Executive summary
In this document we describe the datasets that have been identified as suitable for the work in the scope of WP5, 'Multilingual Text Generation', the available evaluation metrics for the tasks of WP5, together with their strengths and weaknesses, the academic state of the art with regard to benchmarking work done in WP5, and our conclusions regarding the evaluation of the work conducted therein. We identify that the corpus-based evaluation metrics used in state of the art natural language generation research have problems with the setting of WP5: some make fundamental assumptions that are questionable in this multilingual context and the rest assume existence of very large datasets of human-written texts that are not available and would be prohibitively expensive to procure in a multilingual setting. To this end, we propose an alternative evaluation schema based on intrinsic human evaluations of the documents.

Partner in charge: UH

Project co-funded by the European Commission within Horizon 2020 Dissemination Level		
PU	Public	PU
PP	Restricted to other programme participants (including the Commission Services)	–
RE	Restricted to a group specified by the Consortium (including the Commission Services)	–
CO	Confidential, only for members of the Consortium (including the Commission Services)	–

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825153

Deliverable Information

Document administrative information	
Project acronym:	**EMBEDDIA**
Project number:	**825153**
Deliverable number:	**D5.1**
Deliverable full title:	**Datasets, benchmarks and evaluation metrics for multilingual text generation**
Deliverable short title:	**Datasets and evaluation for text generation**
Document identifier:	**EMBEDDIA-D51-DatasetsAndEvaluationForTextGeneration-T54-submitted**
Lead partner short name:	**UH**
Report version:	**submitted**
Report submission date:	**30/09/2019**
Dissemination level:	**PU**
Nature:	**R = Report**
Lead author(s):	**Leo Leppänen (UH), Carl-Gustav Lindén (UH)**
Co-author(s):	**Khalid Alnajjar (UH)**
Status:	**_ draft, _ final, X submitted**

The EMBEDDIA Consortium partner responsible for this deliverable has addressed all comments received. Changes to this document are detailed in the change log table below.

Change log

Date	Version number	Author/Editor	Summary of changes made
19/08/2019	v0.1	Leo Leppänen (UH)	Submitted for UH-internal review
20/08/2019	v0.2	Leo Leppänen (UH)	Clarifications based on UH-internal review
22/08/2019	v0.3	Leo Leppänen (UH)	Clarified the conclusions based on internal feedback
29/08/2019	v0.4	Leo Leppänen (UH)	Added Section 2.4 on a corpora of statistical reports from Statistics Finland
30/08/2019	v1.0	Leo Leppänen (UH)	Additional clarification of language based on feedback from contributors and UH-CS internal feedback
02/09/2019	v1.1	Marko Robnik-Šikonja (UL)	Internal review
11/09/2019	v1.2	Antoine Doucet (ULR)	Internal review
16/09/2019	v1.3	Leo Leppänen (UH)	Addressed internal review from MRŠ and AD
18/09/2019	v1.4	Nada Lavrač (JSI)	Quality check
19/09/2019	v1.5	Leo Leppänen (UH)	Addressed comments from NL
23/09/2019	v1.6	Leo Leppänen (UH)	Ready for submission
25/09/2019	final	Khalid Alnajjar (UH)	Prepared final version for submission
30/09/2019	submitted	Tina Anžič (JSI)	Report submitted

Table of Contents

List of abbreviations

NLG	Natural Language Generation
E2E	End-to-End
WP	Work Package
STT	Finnish News Agency, *Suomen Tietotoimisto*

1 Introduction

The main objective of the EMBEDDIA project is to develop methods and tools for effective exploration, generation and exploitation of online content across languages thereby building the foundations for the multilingual next generation internet, for the benefit of European citizens and industry using less-represented European languages. One facet of this effort is EMBEDDIA work package 5 (WP5), which is concerned with Natural Language Generation (NLG). In order to support journalists and media companies in efficiently reaching as many demographics as possible, the objective of WP5 is to design and develop news automation systems that are transferable across languages, transferable across domains, and are transparent in the NLG process. In particular, with the output of NLG that is dynamic, has narrative structures, and uses figurative and colourful language.

More specifically, WP5 will develop a self-explainable, flexible, accurate, and transparent NLG system architecture that can be transferred to new domains and languages with minimal human effort; develop tools for creation of dynamically evolving content, incorporating narrative structure and user knowledge; and develop tools for creation of figurative language and headlines. The work package consists of three primary tasks.

Task T5.1, Multilingual text generation from structured data, will adapt NLG technology for the requirements of news generation. The task will develop mechanisms for (i) determining what is interesting or important in the given data and deciding what to report, and for (ii) rendering that information in an accurate manner (iii) in multiple languages.

Task T5.2, Multilingual storytelling and dynamic content generation, will develop a novel method for automatically organising news articles based on the domain of the article.

Task T5.3, Creative language use for multilingual news and headline generation, will make the generated texts more varied and colourful by generating creative expressions, especially in headlines. We will find similar terms and metaphors by finding analogous terms in different contexts using context-dependent embeddings. A special focus will be on cross-cultural metaphors.

The purpose of this deliverable – D5.1 'Datasets, benchmarks and evaluation metrics for multilingual text generation' – is to provide an overview of the potential datasets needed to fulfill the aforementioned tasks (Section 2), the methods that can be used to evaluate the implemented systems (Section 3) as well as to identify benchmarks against which to measure the success of the proposed approaches (Section 4).

2 Datasets

Natural language generation is the process of turning some input (usually not in natural language) into natural language (such as English) text (Gatt & Krahmer, 2018). The work conducted within WP5 revolves around T5.1, which is concerned with an NLG subtask known as 'data-to-text NLG', where the system's input is assumed to be in some structured format. This is in opposition to other types of NLG systems which might, for example, ingest images or video to produce captions. Tasks 5.2 and 5.3 then support this primary task by focusing on how to improve the structure and creativity of the produced texts and can thus be seen as parts of this larger data-to-text problem.

In relation to such NLG tasks, data has three roles. First, the system needs *input data*. In the case of data-to-text NLG this is usually some database or other structured dataset that the system ingests and transforms into text. In the case of WP5, we want this dataset to allow the output of the system to be news reports. Second, data is used for training machine learning components that take part in the NLG process. Here, the specific contents of such a dataset are highly dependent on what is being trained. In the case of an 'end-to-end' system, for example, a neural network that *is* the whole of the NLG system, the dataset contains pairs of system inputs and outputs the system is expected

to produce. Third, the data is used to evaluate the quality of the NLG system. Here, the data can be system inputs which are used to produce documents that are then evaluated by humans (discussed further in Section 3.2); alternatively, data consists of pairs of input and expected output that enable the use of automated metrics that investigate whether the system produces output that is similar (for various definitions of 'similar') to the expected output. These methods are discussed in Section 3.1.

In some cases, depending on both the technical aspects of the system being built, as well as its *domain* (i.e. the type of text being produced), a singular dataset can be used for all three purposes. In other cases, this is not possible and separate datasets need to be used for the different purposes. In Section 2.1, we outline the requirements that the EMBEDDIA context imposes on the datasets. In the following subsections, we outline various datasets we have identified to fulfill those requirements.

2.1 Requirements for datasets

To be useful, a natural language generation system needs a dataset that concurrently fulfills multiple qualitative and quantitative requirements. Table 1 describes the requirements imposed on the data in terms of the data quality assessment framework described by Pipino, Lee, and Wang (2002).

The constraints and requirements for a dataset arrive from two major directions. First, the data must be suitable for NLG in general and be of sufficiently high complexity to enable scientific research. Second, the data must allow the demonstration of the technology's suitability for journalistic use. This requirement manifests primarily as requirements for the veracity, believability, lack of errors, etc.

Related to the journalistic requirements, data management is not considered a core function of journalism or newsrooms. Rather, it is something used in the business side of news media for handling consumer data, for instance subscriptions or reader engagement metrics. While exceptions do exist (Magnusson, Finnäs, & Wallentin, 2016), research by Halevy and McGregor (2012) points to challenges - such as lacking data literacy, safety and privacy or standardisation - in data management. Compared with other industries, news media lacks an organizational culture of maintaining projects over time, which is a core feature of data management (Stavelin, 2014). These factors suggest that a successful news automation technology should work with datasets that require minimal data management, i.e. public datasets that are usable as-is.

Finally, we identify the requirement for enabling multilingual generation as somewhat orthogonal to previous requirements. While it does not significantly affect the scientific side (technically, it is possible to do multilingual generation from even hyper-local data), it has a strong interplay with the journalistic requirements. Namely, the requirement for the journalistic potential should preferably be fulfilled in all the languages and locales relevant to the project. As such, the dataset must either be non-local or alternatively contain local data for all the relevant locales.

When seeking datasets concerning creative content (e.g., humorous news headlines) to be used for natural language generation, defining a set of required aesthetics to exist in the content and match the overall goal is essential. These aesthetics should not be limited to creativity aspects but should also cover all the necessary elements of the type of content. As an example of aesthetics for humorous news headlines, factualness and grammatical correctness of the headlines could be the minimum required elements. Regarding the creativity aesthetics, having a pun (e.g., a word substituted from the original headline based on a rhyme) in these headlines could be the desired creativity aesthetic. Additional aesthetics could be considered depending on the NLG system (e.g., parallel text), content (e.g., non-offensive headlines and headlines that fit the context of the article), type of creativity (e.g., rhetorical expressions such as metaphors) and so on. The requirements mentioned earlier (e.g., openness and multilinguality) also contribute to the quality of datasets of creative content.

Table 1: Requirements for the dataset underlying NLG in terms of the Pipino et al. (2002) framework.

Dimension	**Definition (Pipino et al., 2002)** 'The extent to which...'	**Applicability**
Accessibility	... data is available, or easily and quickly retrievable	Preferably public, free of charge and online; should not require significant preprocessing; machine-readable
Appropriate Amount	... the volume of data is appropriate for the task at hand	Amount should be large in order to highligh advantages of automated processing
Believability	... data is regarded as true and credible	Very high due to the journalistic context
Completeness	... data is not missing and is of sufficient breadth and depth for the task at hand	High requirement for completeness due to setting
Conciseness	... data is compactly represented	Low; data is transformed into an internal format
Consistent	... data is presented in the same format	Very high, as constant changes in format would necessitate constant system modifications
Ease of Manipulation	... data is easy to manipulate and apply to different tasks	Low; most manipulation is done on internal data representations that are not affected by source format
Free-of-Error	... data is correct and reliable	Very high due to the setting
Interpretability	... data is in appropriate languages, symbols, and units, and the definitions are clear	Very high
Objectivity	... data is unbiased, unprejudiced and impartial	Very high
Relevancy	... data is applicable and helpful for the task at hand	Low; the specifics of the generation task are driven by the available data
Reputation	... data is highly regarded in terms of its source or content	Very high
Security	... access to data is restricted appropriately to maintain its security	Data should be as open as possible and thus require as little security as possible
Timeliness	... the data is sufficiently up-to-date for the task at hand	High requirement from the journalistic view; less-significant w/r/t scientific effort
Understandability	... data is easily comprehended	Very high
Value-Added	... data is beneficial and provides advantages from its use	Journalistically data should contain potential for finding newsworthy phenomena, scientifically value-added is largely determined by other factors

2.2 The Eurostat dataset

Discussions with the media partners involved in the EMBEDDIA project, conducted during the project's workshop to elicit the needs of the media partners, held in Tallinn, Estonia March 2019 (M3), identified the Eurostat, the statistical office of the European Union, as a potential source for suitable data. Eurostat offers a large amount of structured datasets[1] pertaining to a wide variety of domains about the European Union, both in the aggregate and for each individual member state. In addition to being usable as general system input, the Eurostat dataset is applicable to evaluation in settings where known-good output texts are not required. That is, it can be used to produce texts that are then evaluated manually by humans.

In terms of the dataset requirements set forth above, the provided data is highly *accessible*, being available online, publicly and for free, in a machine-readable format. The total *amount of data* is large, thus enabling the work conducted within the scope of WP5 to highlight the strengths of automation. At the same time, the individual tables that make up the database are sized so as to be *understandable* by a human. The units used are also sufficiently *intepretable* with high-quality explanatory texts available to aid in interpretation. As such, it still enables us to verify the work manually.

As a provider of data, Eurostat as the statistical office of the European Union has a good *reputation* and can be viewed as highly trustworthy. As such, the provided data is *believable* and can be assumed to be *objective* and *free of errors*. In fact, significant errors or biases in the Eurostat data would be potentially newsworthy by themselves. The data provided by the Eurostat encompasses a significant amount of settings and is very *complete* in the sense that very little data is missing. When data is missing, it is clearly marked as such. Data is provided in a highly standard TSV format and is thus very *consistent* and suitably *easy to manipulate*.

In journalistic terms, the dataset encompasses a significant amount of domains (e.g., economy, population, social conditions, quality of life, immigration, equality, climate change etc.) that have high journalistic potential, thus making the dataset *relevant* for our use. It is updated reasonably often, thus also fulfilling our need for *timeliness*. Altogether, these factors of journalistic potential result in potential for *added value*, thus enabling the data to highlight the possibilities facilitated by the technologies being developed. At the same time, the dataset suffers from the drawback that while the domains are 'newsworthy' in the sense that news often talk about them, they are relatively 'dull' compared to other topics covered in the news. Furthermore, it is not guaranteed that interesting (newsworthy) insights will be found in the dataset.

For development purposes, we have selected a subset of the dataset as a starting point (Table 2). For all tables, data is available from the aggregated Euro area (EA11-2000, EA12-2006, EA13-2007, EA15-2008, EA16-2010, EA17-2013, EA18-2014, EA19), the aggregated European Union (EU6-1958, EU9-1973, EU10-1981, EU12-1986, EU15-1995, EU25-2004, EU27-2007, EU28-2013), and individual countries (including some non-EU reference countries) i.e., Belgium, Bulgaria, Czechia, Denmark, Germany (until 1990 former territory of the FRG), Estonia, Ireland, Greece, Spain, France, Croatia, Italy, Cyprus, Latvia, Lithuania, Luxembourg, Hungary, Malta, Netherlands, Austria, Poland, Portugal, Romania, Slovenia, Slovakia, Finland, Sweden, United Kingdom, Iceland, Norway, Switzerland, North Macedonia, Serbia, Turkey and United States.

This subset should demonstrate the suitability of the technology for working not only with the full Eurostat dataset, but also demonstrate the ability of the methods to work with similar more local datasets provided by various national statistical agencies such as Eesti Statistika[2] in Estonia, Državni zavod za statistiku[3] in Croatia, Statistični urad Republike Slovenije[4] in Slovenia, Institut national de la statistique

[1] https://ec.europa.eu/eurostat/data/database
[2] https://www.stat.ee/
[3] https://www.dzs.hr/
[4] https://www.stat.si

Table 2: Selection of Eurostat database tables used for development purposes.

Table	Content
ei_cphi_m	Harmonized customer price indices for various categories by country. Presented monthly from 1996 to 2019.
ilc_di03	Mean and median income by age and sex. Data is available on the level of countries expressed in euros, national currencies and the purchasing power standard. Data available by-year from 1995 to 2018.
tepsr_sp310	Household out-of-pocket expenditure on healthcare as a percentage of total current health expenditure. Data available by-year from 2005 to 2017.

et des études économiques[5] in France, Tilastokeskus[6] in Finland and Office for National Statistics[7] in the United Kingdom.

2.3 Finnish News Agency Archive 1992-2018

The Finnish News Agency, STT, has published a corpus (STT, n.d.) of newswire articles. The corpus consists of some 2.8 million Finnish language newswires sent to media companies between 1992 and 2018. The contents of the corpus are in the industry-standard NewsML-G2 format[8]. The corpus is accessible, in a restricted manner, to third parties through the Finnish Language Bank.

As the dataset is not structured, it cannot be used as a system input for news production. Similarly, since the contents are not produced from the data we intend to use as input, namely the Eurostat data, it cannot be used directly for evaluation. This limits the data to be used as training data for any machine learning task. Furthermore, it can be used qualitatively to direct certain development efforts as general examples of journalist-produced news articles.

As such, we intend to use the STT dataset for quantitative analysis directing work, and as training data for any suitable machine learning tasks.

2.4 Corpora of statistical reports from Statistics Finland

Statistics Finland is the official statistical agency of the state of Finland. It has produced altogether some 15 000 statistical reports, total, in three languages, namely Finnish, Swedish and English. These texts, ranging back to 2005, describe important factors of the datasets collected by Statistics Finland. We have obtained a copy of these articles under the Creative Commons license.

Unfortunately, while potentially very interesting, the corpus suffers from several practical limitations. First, the texts are not well-aligned with individual tables of data. Rather, they are linked to larger themed collections of multiple tables of data, where it is not clear which parts of the linked datasets were used in the production of the texts. Second, it's not clear whether the writers conducted additional analysis by, e.g., deriving numbers that are not present as-is in the underlying data and the corpus is too large to verify this manually. Third, the texts are not aligned temporally with the data. That is, while each text links to the relevant tables of data, the linked tables have been updated since the production of the text. Furthermore, while the corpus as a whole is relatively large (over 15 000 documents), it spans

[5] https://www.insee.fr
[6] http://stat.fi/
[7] https://www.ons.gov.uk/
[8] https://iptc.org/standards/newsml-g2/

Table 3: A table summarizing the available dataset on creative headlines.

Dataset	Language	Description
unfun.me	English	A dataset of 9159 satirical headlines obtained from `theonion.com` and 9000 serious headlines collected from 9 news websites, along with their URLs pointing to the news articles. The dataset contains rating for these headlines whether they are saterical or serious based on human judgments. Furthermore, online users have participated in a task to convert 1191 satirical headlines into (a total of 2801) serious-looking ones. The seriousness of these satirical-serious pairs of headlines were then rated by human evaluators on a binary scale.
Humicroedit	English	A dataset released for the goal of training a model to automatically predicting the funniness of headlines. The dataset contains 9653 headlines, their modified humourous versions and the average funniness score they received on a 4-point Likert scale. Only one token (e.g., a word or named entity) was changed in the original headlines to make them funny. The dataset also contains 9381 pairs of two altered headlines (having the same original headline) and a score representing which of them is funnier, intended for building a model that predicts which of two input headlines is funnier.
STT's creative headlines	Finnish	A dataset of 84 creative and 24 non-creative manually picked and annotated Finnish news headlines by the Finnish News Agency STT from (STT, n.d.). In addition to the "creative" categorization, STT has provided a detailed explanation for why the creative headlines are considered interesting and suitable.

three languages and multiple different themes. As a result, very few texts exist *per language per table of data*. Together, these factors significantly limit the ways this corpus can be employed.

Given the above considerations, we will not use datasets from Statistics Finland as primary data, but keep them in reserve in case the need arises.

2.5 Corpora of creative headlines

Creativity is an important part of news, as it gives the journalists the ability to express the news' novelty while keeping the articles interesting to the readers. As WP5 focuses on natural language generation of news, we consider textual datasets that contain some kind of creative writing and relate to the news domain. Creative writing in news can be exhibited in various forms (e.g., satirical, humorous, puns, metaphorical etc). As creativity is a complex phenomena to fully understand or model computationally, recent academic research on creativity and news concentrated on introducing creativity to headlines. Here, we describe and discuss three datasets on creative headline, 1) unfun.me (West & Horvitz, 2019), 2) Humicroedit (Hossain, Krumm, & Gamon, 2019), and 3) creative Finnish headlines annotated by STT. Table 3 provides a brief comparison between the datasets and the remainder of this section describes them in detail.

2.5.1 unfun.me

The unfun.me dataset by West and Horvitz (2019)[9] contains satirical news headlines collected from the satirical newspaper The Onion[10]. The headlines are modified by human participants to convert them into serious-looking headlines with minimal changes, which yields a corpus of parallel satirical-serious news headlines. To evaluate the dataset, human evaluators examined whether the satirical headlines, their altered versions, and other real non-satirical headlines are real headlines or not. The intent behind creating such a dataset was not to allow generation of humorous headlines but rather to analyze humour in headlines and to support building computational models for detecting humour. Nonetheless, the analysis and dataset could assist in constructing a creative headline generation system by, e.g., learning what to change to make a headline humorous.

2.5.2 Humicroedit

Humicroedit (Hossain et al., 2019) is a dataset of actual news headlines and their modified alternatives, made funny by online human workers. Workers were requested to change only one word to facilitate understanding how a simple change could transform a headline to be funny. The authors have evaluated the changes by asking different online workers to examine whether they are funny on a 4-point Likert scale following Hossain, Krumm, Vanderwende, Horvitz, and Kautz (2017). Similarly to unfun.me, this dataset could be used within WP5 to allow the creative NLG system make intelligent decisions about what to replace and, potentially, what would be a funny replacement.

2.5.3 Creative Finnish headlines annotated by STT

The previously mentioned datasets regarding creative headlines are in English and, in the case of Humicroedit, do not have their corresponding article (i.e. no context). Furthermore, they do not have the journalist's point of view or any kind of a justification for why a headline is supposed to be funny. Creating a creative headline generation system solely based on these data might result in generating something humorous but not necessarily apt, factual, or even useful for journalists and readers. Due to this, the Finnish News Agency, STT, has provided UH-CS with a small corpus of headlines considered creative by their writing staff. In addition, the corpus contains a short description about the reasoning behind the creativity of each headline and why it is appreciated. The corpus also contains a selection of headlines that are considered particularly non-creative. This corpus is intended to be used within the scope of WP5 by carefully studying it and incorporating the views of journalists in constructing a system that knows when is it suitable to introduce creativity (e.g., not in sensitive articles talking about war), and what kind of creativity is more suitable in different contexts such as puns and metaphors.

3 Evaluating natural language generation systems

Evaluation of natural language generation is a notoriously difficult problem, as demonstrated by the wide variety of automated and non-automated evaluation metrics proposed and used in the literature. In the following sections, we overview both automated and non-automated (i.e. human) evaluation methods for NLG research.

[9] Publicly available at https://github.com/epfl-dlab/unfun
[10] https://www.theonion.com/

3.1 Automated evaluation metrics

A recent survey of the field of natural language generation by Gatt and Krahmer (2018) identified that the basic problems causing the present state of difficult evaluation are variable inputs and multiple possible outputs. Of these, the first is more related to benchmarking (i.e. comparing one system to another), whereas the second applies to evaluation more generally.

By the multiplicity of outputs, Gatt and Krahmer (2018) refer to the observation that for any generation task there are multiple potentially valid answers: even extremely simple pieces of information can be expressed in natural language in a multitude of completely valid ways. As the length of the generated text increases, the variability of 'correct' solutions also continues to increase.

This results in a situation where, in the case of non-human evaluation, relatively complex metrics have to be employed. The underlying assumption behind most of these metrics is that a *candidate* – in case of NLG, the text generated by the system – is good if it is close to some human-created *reference* text. The variability of the outputs (as discussed above) is then accounted for by virtue of having a corpus which consists of a multitude of 'correct' references for each generation task, i.e. for the candidate. That is, they assume that each input is associated with a large set of possible and correct outputs, that together are a representative sample of all good outputs. This latter requirement for a large corpus consisting of $1 - to - n$ input-output pairs is often prohibitive, especially as the complexity and length of the generated documents increases.

The pre-existing datasets for complete NLG systems, also known as end-to-end datasets, tend to be highly domain-specific, mostly with very 'neat' and clearly defined domains such as restaurants, e.g., the E2E NLG Challenge dataset (Novikova, Dušek, & Rieser, 2017), and the BAGEL dataset (Mairesse et al., 2010); sports, e.g., boxscore (Wiseman, Shieber, & Rush, 2017); or weather, e.g., WeatherGov (Liang, Jordan, & Klein, 2009), and SUMTIME, a corpus of outputs from the SUMTIME system (Reiter, Sripada, Hunter, Yu, & Davy, 2005). We are not aware of any such corpus for the relevant datasets that we intend to employ for the task, and producing such a dataset (considering the intended length and information complexity of the produced texts) is not feasible at this stage. Despite the problems identified above, including the lack of a suitable reference corpus, we have identified several automated metrics that we can employ if such a corpus becomes available later via the efforts of some third party.

It is notable that generation of news falls into a category of NLG tasks where the user is potentially exposed to a large amount of outputs that should be varied in language, while retaining correctness with regard to the information content and certain stylistic constraints. In such a scenario, the quality of a piece of output is not a static value but rather dependent on the outputs seen previously by the user. We are not aware of any automated metrics that account for this 'moving target' nature of desired output.

3.1.1 BLEU

BLEU (Papineni, Roukos, Ward, & Zhu, 2002), or BiLingual Evaluation Understudy, is, in broad terms, a modification of the standard precision metric used in information retrieval.

Formulated in terms of a *candidate* sentence produced by a software and a set of *references* produced by humans, the standard precision would be applied as a metric of similarity by taking the fraction of words (or n-grams, sequences of n words) in the candidate that also appear in the references. This trivial formulation, however, fails due to the simple fact that it can be gamed by using short candidates consisting of only very common words. As an example, the sentence '*the the the*' has precision of 1, the maximum, when compared to the sentence '*the newspaper that Harry bought home turned out to be from yesterday*' as all words in the candidate appear in the reference.

BLEU makes modifications to account for these two major problems. The basis of the BLEU score is the *modified unigram precision*. Assuming we have a function $Count(\omega_k^n, c)$ which counts how many times

some n-gram ω_k^n appears in some sentence c, we define a clipped version

$$Count_{clip}(\omega_k^n, c, References) = \min\left(Count(\omega_k^n, c); \max_{r \in References}(Count(\omega_k^n, r))\right), \quad (1)$$

which provides the number of times the n-gram appears in the candidate, clipped by the maximum number of times it appears in any single reference. The modified unigram precision is then obtained by calculating the sum of the clipped counts over all unigrams ω_k^1 (words) in the candidate over the total length of the candidate:

$$p(c, References) = \frac{\sum_{\omega_k^1 \in c} Count_{clip}(\omega_k^1, c, References)}{|c|} \quad (2)$$

This idea is then expanded to n-grams (sequences of n words) for any n and multi-sentence documents in the following manner:

$$p_n(Candidate, References) = \frac{\sum_{sentence \in Candidate} \sum_{\omega_k^n \in sentence} Count_{clip}(\omega_k^n, sentence, References)}{\sum_{sentence' \in Candidate} \sum_{\omega_k^n \in sentence'} Count(\omega_k^n, sentence')}. \quad (3)$$

A weighted average of this measure for various n (usually $1 \geq n \geq 4$ with equal weights w_n for all n) is then combined with a *brevity penalty* to penalize candidates that are very short compared to the references. This brevity penalty is calculated as

$$BP(Candidate, References) = \begin{cases} 1 & \text{if } |Candidate| > effective\_length(References) \\ c^{1-r/c} & \text{if } |Candidate| \leq effective\_length(References) \end{cases}, \quad (4)$$

where $|c|$ is the length of the candidate and $effective\_length(References)$ is, for multi-sentence cases, obtained by 'summing the best match lengths for each candidate sentence in the corpus' (Papineni et al., 2002).

Combining these gives us the final BLEU score:

$$BLEU(Candidate, References) = BP(Candidate, References) \times \exp\left(\sum_{n=1}^{N} w_n \log p_n(Candidate, References)\right). \quad (5)$$

While BLEU is a standard evaluation metric in machine translation, it has a complicated status in NLG. While a significant proportion of NLG papers do report BLEU scores when possible, questions have been raised about its suitability and robustness. Empirical evidence based on metastudies of the metric's correlation with human judgements question the validity of using BLEU outside of machine translation (Belz & Reiter, 2006; Reiter, 2018). Even within the MT community, concerns have been raised about the comparability of reported BLEU scores due to differences in parametrisation and processing (Post, 2018).

Given this status, we will report BLEU scores if some other third party produces the required evaluation data. At the same time, we do not intend to produce said data ourselves as it would be prohibitively costly compared to the expected strength of the obtained evaluation.

3.1.2 ROUGE

ROUGE, or Recall-Oriented Understudy for Gisting Evaluation (Lin, 2004), is a family of related evaluation metrics that all are related to the standard recall metric as used in information retrieval. It is often used in conjunction with BLUE, similar to how both recall and precision are often reported together in information retrieval.

The simplest of the ROUGE metrics is Rouge-N, which is defined by Lin (2004) as

$$\text{Rouge-N} = \frac{\sum_{S \in References} \sum_{\text{n-gram} \in S} Count_{match}(\text{n-gram})}{\sum_{S \in References} \sum_{\text{n-gram} \in S} Count(\text{n-gram})}, \tag{6}$$

with $Count_{match}$ being the 'the maximum number of n-grams co-occurring in a [candidate] and a set of [references]' (Lin, 2004). Intuitively, this counts the fraction n-grams present in the references that also appear in the candidate.

A common variant of the ROUGE-N metric is ROUGE-L, which is an F-score of the recall and precision based on the longest common subsequence $LCS(r, c)$ of words between the candidate text c and the reference text r. Below, β is used to set recall to have β times as much weight as precision. That is, with $\beta = 1$, both are equally as important but with $\beta = 2$ recall is twice as important as precision.

$$R_{lcs} = \frac{LCS(r, c)}{|r|} \tag{7}$$

$$P_{lcs} = \frac{LCS(r, c)}{|c|} \tag{8}$$

$$\text{ROUGE-L} = \frac{(1 + \beta^2) R_{lcs} P_{lcs}}{R_{lcs} + \beta^2 + P_{lcs}} \tag{9}$$

In case of multiple references, the maximal value for any reference r is taken.

Another common variant of ROUGE is ROUGE-S, which simply replaces both R_{lcs} and P_{lcs} with $SKIP_n(X, Y)$ which is the number of skip-n-gram matches between X and Y while also replacing the reference and candidate lenghts $|r|$ and $|c|$ with $\binom{|r|}{n}$ and $\binom{|c|}{n}$ respectively. Here, a skip-n-gram is an n-gram, i.e. sequence of n words, that is not necessarily continuous but is in the correct order. This means that both '*A B*' and '*A C*' are valid skip-2-grams of the sentence '*A B C*'. At the same time, '*C A*' is not a valid skip-2-gram of the same sentence since the ordering of the words in the original sentence is not respected.

Like BLEU, significant concerns have been raised about the limitations of ROUGE as a metric for evaluating NLG (Belz & Reiter, 2006; Novikova, Dušek, Curry, & Rieser, 2017). Namely, ROUGE scores do not necessarily correlate well with human judgements, especially in situations where the systems aim for high variability of language (Belz & Reiter, 2006) and can fail to distinguish between high and medium quality systems (Novikova, Dušek, Curry, & Rieser, 2017). At the same time, it has retained a status of a *de-facto* standard (together with BLEU) and is almost always reported if calculation is possible in the first place. Given this status, like with BLEU, we will report ROUGE scores if some other third party produces the required evaluation data. At the same time, we do not intend to produce said data ourselves as it would be prohibitively costly compared to the expected strength of the obtained evaluation.

3.1.3 METEOR

Another metric commonly used in NLG literature is METEOR (Lavie & Agarwal, 2007). It is based on aligning words between the candidate and the references in a fairly complex manner. While METEOR has been used extensively for evaluation in e.g., English, it suffers from a series of downsides. Namely, it depends on the use of language-specific stemmers, lists of synonym sets and numeric hyperparameters. Notably, the use of stemmers makes it a questionable fit for very synthetic languages, such as Finnish, which conveys meaning by adding suffixes and inflection. For such languages, stemming would remove a high fraction of the information content in a sentence, analogous to removing function words from an English language sentence. The requirement for synonym sets makes METEOR difficult to set up for smaller languages where large synonym sets are not available. Finally, to our knowledge, the standard METEOR is only available pre-tuned for English, Czech, German, French, Spanish, and Arabic. As such, we do not intend to use the baseline version of METEOR for the evaluations.

More recently, a 'universal' version of the METEOR metric has also been published (Denkowski & Lavie, 2014). The universal version of METEOR learns the required parameters from a corpus and a phrase-table. A much newer metric than BLEU and ROUGE, the limitations of METEOR universal are not as well explored and the evidence of superiority over BLEU and ROUGE is mixed (see Gatt and Krahmer (2018) for discussion). Furthermore, the amount of training data required for this universal version to function properly is not clear: the original work describing the method does not explicitly address the question and the manual of the software library simply states a need for 'enough data to build a standard phrase-based machine translation system'[11]. Thus, for the low resource languages used within this project, it would be unclear to which degree a certain universal METEOR score would reflect failures of parameter tuning and paraphrase learning and to what degree the actual performance of the system. As such, we do not intend to use the universal version of METEOR for evaluation.

3.1.4 CIDEr

CIDEr, or Consensus-based Image Description Evaluation (Vedantam, Lawrence Zitnick, & Parikh, 2015), observes the cosine similarity of sentence vectors obtained by weighing the n-grams of the sentences with their Term Frequency - Inverse Document Frequency (TF-IDF).

A technical description of CIDEr is skipped due to a possible problem with the CIDEr in the context of the languages used within this project. Namely, the standard CIDEr procedure, as described by Vedantam et al. (2015), includes a preprocessing step in which all words in both the candidate and the references are stemmed. We are not aware of any studies that investigate the effects of this preprocessing in the context of highly *synthetic* languages, meaning languages that convey meaning by adding suffixes and inflecting words. For such languages, stemming would remove significant amounts of information. As such, CIDEr's suitability for languages such as Finnish is questionable and we do not intend to employ it in our evaluation.

3.2 Human evaluation

The difficulties with automated evaluation of NLG are well-acknowledged within the field and as such human evaluations are common. Human evaluations are divided into two sub-categories: intrinsic and extrinsic evaluation schemas. We will describe both of these separately.

Extrinsic human evaluations are evaluation schemes that observe whether the output of the system results in humans doing some right action. For example, Reiter, Robertson, and Osman (2003) investigated how effective computer-personalized letters were in encouraging smokers to stop smoking. This effect was then compared to the effect a non-personalized letter had. Another study (Portet et al., 2009) gave nurses and doctors automatically generated summaries of patients' statuses in a neonatal intensive care unit and asked them what their next actions would be based on said summaries. These actions were then compared to a gold standard obtained from experts given access to the data underlying the summary.

In the case of generating news stories, it is not clear what such an extrinsic task would be. There is no specific 'thing' the user is expected to do after interacting with the system. We have only identified two possible extrinsic tasks. The first of these is memorizing information, meaning we would test whether the users have, after reading the news article, retained the knowledge presented to them. Such a task is, however, inherently unrealistic. News stories are not intended to be a learning material and are not designed to maximize recall of information. It is perfectly reasonable for a user to read a news story, be satisfied with what they read and not recall the specifics. As such, any test of specific knowledge would be unreasonably strict.

[11] METEOR readme, accessible at `https://www.cs.cmu.edu/ alavie/METEOR/README.html`

The second task we have identified is testing *understanding*. The setup here would be similar to the memorization task, but the user would be able to consult the text while answering the questions. The problem with this task lies in defining suitable questions to test understanding: the questions would need to be non-trivial to produce usable results. At the same time, they cannot test for knowledge outside of the presented text. One potential solution to this task would be to elicit the questions from third parties by exposing them to the text and asking them to produce questions. The problem then becomes that if the system is actively misleading, the producers of the questions, too, would be misled. At the same time, the producers of the questions cannot be exposed to the underlying data due to it's size and also because they would then likely end up producing questions that are not answerable based on the textual document alone. Due to these difficulties, we do not currently plan on conducting an extrinsic evaluation of understanding either.

Intrinsic human evaluations refer to evaluation settings where humans are presented with the output of the system and then asked questions about it. Such evaluations are common when evaluating computer-generated news articles, having been conducted e.g., by Clerwall (2014), van der Kaa and Krahmer (2014), Graefe, Haim, Haarmann, and Brosius (2018), Jung, Song, Kim, Im, and Oh (2017), and Melin et al. (2018). For NLG in general, common questions asked during such evaluations include questions about the fluency, readability, correctness, pleasantness, etc. (Gatt & Krahmer, 2018). We overview the relevant literature that evaluated computer-generated news articles, observing specifically their evaluation setup, in Section 4.

It is notable that the sizes of EMBEDDIA working languages present a possible problem with regard to human evaluation, namely in finding a sample of native speakers that enables statistically sound evaluations. Although at this point somewhat old (considering the topic) and focused on only a single platform, the research of Pavlick, Post, Irvine, Kachaev, and Callison-Burch (2014) studied the language demographics of the workers available from Amazon's Mechanical Turk. They characterize most of the smaller languages relevant for WP5 as having few workers that work slowly, albeit with high quality. In their 2013 book, Eskenazi, Levow, Meng, Parent, and Suendermann (2013) note that they were only able to elicit 139 judgements for a language task in Canadian French, a language with some 7 million speakers. These results align with our experiences in trying to use international crowdsourcing platforms to judge Finnish language content.

3.3 Evaluation of news text structure

We are not aware of any metrics for evaluating the structural quality of news articles. The work conducted within the scope of task T5.2 will thus be evaluated alongside the work conducted within task T5.1 using extrinsic human evaluations by having online judges evaluate versions of the produced news stories both with and without the improved structuring. By showing the judges texts that only differ in the content structuring, we can isolate the effect of the improved methods developed within the scope of T5.2. The same consideration regarding qualitative and quatitative evaluation of T5.1 apply to this task as well.

3.4 Evaluation of creative NLG systems

When evaluating a creative system, there are multiple perspectives one could consider to assess its creativity. Most notably, one can evaluate not only the *output* of the system, but also the system itself. Multiple researchers in the computational creativity community have proposed different frameworks to evaluate the creativity of computational system on an abstract level (Jordanous, 2012, 2016; Colton, Charnley, & Pease, 2011; Colton, 2008; Linkola, Kantosalo, Männistö, & Toivonen, 2017). When evaluating a NLG system for generating creative headlines, these frameworks could be utilized as they provide us with a direction of what to consider when assessing computational creative systems.

In terms of evaluating the output of a creative headline generation system, we are unaware of any automated evaluation metrics to measure the creativity of written text. Given the subjective nature of creativity, a rational approach for evaluating creative text is to conduct an intrinsic human evaluation by asking human evaluators to provide their opinions on the system's output. However, it is crucial to define the questions in a manner that reflects the nature of the output and all of its desired qualities. As for the case of the creative headlines, the questions should address qualities of headlines (e.g., their factualness, grammaticality and so on) along with creativity criteria (e.g., their humorousness).

4 Benchmarks for multilingual text generation

As noted above, evaluation of NLG is not only difficult due to the multitude of possible 'correct' outputs for any individual generation task, but also due to variability of inputs. The variability of input means that different systems employ wildly varying types and formats of inputs without any clear standardization. This lack of standardisation presents an obstacle for comparing a proposed system against previous state of the art and even makes identification of 'state of the art' difficult.

The output from NLG systems can be perfectly readable and understandable texts, but the quality and the reading experience does not match texts produced by humans in terms of nuance and variability (Diakopoulos, 2019). However, there is very limited research on the perception of automated news. Users' thoughts on automated journalism have been described in only a few papers up to now.

Using Swedish test subjects, Clerwall (2014) measured the perception of automated articles in English, in a setting where the news source (human-written or computer-written) was not declared to the test users. No significant differences in users' perceptions of the texts were found, except that the human-written news got more positive ratings for the 'pleasant to read' descriptor. However, this study did not use the same data for the human versus the machine-written news, making the comparisons problematic. Furthermore, the online judges were only subjected to one text. As such, the results would be unlikely to reflect any differences in the variability of the produced texts, e.g., whether all the computer-generated texts are very similar whereas the human-produced texts vary more.

Van der Kaa and Krahmer (2014) examined the user perception of computer-written news articles with by-lines potentially manipulated to falsely state a human author, on the dimensions of expertise and trustworthiness. In this evaluation with Dutch content and Dutch-speaking respondents, they found no strong differences in perceived expertise nor in trustworthiness, amongst regular news consumers. Notably, since there were no human-produced control articles, the results of the study must be carefully interpreted. Mainly, the results should not be taken to mean that human-produced and computer-generated texts were seen as equally trustworthy, but rather that masking the identity of a computer 'author' does not affect the perceived expertise and trustworthiness at least in the case of these specific articles.

Building on the results of van der Kaa and Krahmer (2014), Graefe et al. (2018) performed evaluations on the impact of the actual and declared source (human-written, computer-written) of the news, on three dimensions: credibility, readability, and journalistic expertise. For this study, they developed a measure for content perception using 12 items and performed tests in an all-German context by varying the actual news source. Their study found that computer-written articles were rated as more credible and higher in terms of expertise than the human-written articles. Regarding readability, human-written articles were rated significantly higher. However, the finding that the declared source has an impact on perception taints these ratings, as computer-written articles were rated substantially higher on readability if declared as written by a journalist. As with the van der Kaa and Krahmer (2014) study, Graefe et al. subjected each judge to only a single computer-generated article. As such, their results should be interpreted as a 'best-case' scenario where any effects caused by possible lack of variety in language would not have yet manifested.

The impact of declared source appears to vary between countries. Jung et al. (2017) found that in

South Korea articles attributed to a computer received higher ratings than those attributed to a human. While South Korea ranked 25th of 26 developed countries in that news trustworthiness survey, Finland stood out as the clear number one, with 65 % agreeing with the statement 'you can trust most news most of the time' (Newman, Fletcher, Kalogeropoulos, Levy, & Nielsen, 2016). Findings from a mixed European nationality study – although with $\frac{3}{4}$ being Germans – by Wölker and Powell (2018), showed that credibility perceptions of computer generated news can be considered equal to human-created content. That study found that for special topics like sports, automatically generated news can even be perceived as more credible than a human-written story.

To our knowledge, there is only one study (Melin et al., 2018) referencing user perception evaluations for automatically generated textual news content in the contexts or languages where the EMBEDDIA project operates. In this study, 152 users were asked to evaluate the output of texts in Finnish generated by a NLG system called Valtteri. Each evaluator rated six preselected computer-generated articles, four control articles written by journalists, and four computer-generated articles of their own choice. All the articles were evaluated along four dimensions, inspired by work of Sundar (1999): credibility, liking, quality, and representativeness. As expected, the texts written by Valtteri received lower ratings than those written by journalists, but overall the ratings were satisfactory (avg. 2.9 vs. 4.0 for journalists on a 5-point scale). Valtteri's best rating (3.6) was for credibility.

It should be noted that several of these studies present a false dichotomy of texts produced by journalists or machines in terms of production effort. In reality, humans are very much in the loop since the templates used for 'automated' creation of texts in NLG systems are all handcrafted by journalists, a potentially laborious process that can involve careful choosing of wordings, synonyms and expressions. For example, Associated Press is using some two dozen manually-written templates for financial reporting (Lindén, 2017). When German NLG service provider Retresco in 2018 created 8,000 templates for Bundesliga, they were also written by journalists, which was an effort that lasted several months. Similarly, any system based on automatic extraction of templates is dependent on previous human effort to produce the texts from which the templates are extracted.

This means that it does not make that much sense to compare stories written by 'humans' or 'machines' without a consideration of this context. However, it is obvious that text templates written by humans to be used at massive scale should differ in style and tone (more objective, neutral) from stories written as standalone pieces of quality journalism that are supposed to deeply engage the reader. Producing more complex news articles requires complex systems that either employ opaque statistical methods (i.e. machine learning, neural networks) that are problematic for the news industry, or alternatively are transparent but costly to set up (Melin et al., 2018). We are aware of no effort that has attempted to quantify the interplay of human effort in system-setup with output quality.

Given the problems present with automated and extrinsic human evaluations in the news generation domain, we believe that intrinsic human evaluations present the best method for evaluation. At the same time, the details of the evaluation setups in other studies vary. Furthermore, the systems employed in the studies are in multiple different sub-domains of news, none exactly like the expected domain to be investigated in WP5. Here, the question then becomes how to conduct the evaluation to ensure interpretability of the results with respect to other studies within the field of automatically generated news articles. As such, while any results obtained can be compared to the results presented by the cited studies, they are not applicable as benchmarks to be beaten.

The majority of NLG research on automated news generation focuses on producing descriptive natural text that conveys an intended message. Recently, researchers (Lynch, 2015; Gatti, Özbal, Guerini, Stock, & Strapparava, 2015; Alnajjar, Leppänen, & Toivonen, 2019) have been studying methods for making news colourful and creative, in addition to the efforts in collecting and making datasets on the topic available (West & Horvitz, 2019; Hossain et al., 2019). Lynch (2015) has developed a system for adding a well-known phrase (e.g., song, movie titles etc) as a prefix to a headline, where the juxtaposition of the two expressions is intended to make the generated headlines catchy. The internal process of the system searches for well-known phrases that match the headline semantically and ranks them. The method proposed by Gatti et al. (2015) finds well-known phrases that are semantically related

to the original headline and replaces a word in them to make them relate to the context of the news article, while satisfying a semantic similarity threshold, and lexical and syntactic constraints. Notably, these methods alter an original (i.e. human written) headline to give it a creative touch. The research by Alnajjar et al. (2019) incorporates these methods along with a method for inserting metaphorical expressions in an automated journalism system and tests their effects on the output.

Given that the nature of creative content is subjective, the generated texts of such systems are evaluated by humans and, in some cases, compared to a baseline. To the best of our knowledge, there are not any benchmarks for creative news generation. As the interest in the topic is increasing and more datasets are being made available, benchmarks of potentially sub-tasks of the NLG process (e.g., detecting whether a change in a headline is humorous or not) might be available in the future (e.g., benchmarks of a shared task such as SemEval-2020 Shared Task 7 on assessing the funniness of edited news headlines). Despite that, direct comparisons of benchmarks (e.g., accuracy of detecting humour or survey scores by human judges) in the computational creativity field is usually inapplicable as there is no gold standard for creativity. Hence, evaluation scores whether they are computationally calculated (e.g., accuracies of models for detecting humour etc) or obtained from humans (e.g., crowdsourcing) are given as a reference and not as a score to beat as they are based on human opinions.

Testing the produced headlines – and other content – by exposing it to the users of EMBEDDIA media partners as part of their standard (online) coverage seems enticing from the perspective of the obtained data and feedback. This approach, however, is ethically problematic especially in terms of work-in-progress systems and presents potential legal liabilities to the media partners. Discussions on what kinds of testing setups are legally and ethically feasible will be conducted with the media partners as the produced systems become more mature and we have a better understanding of their performance and their specific weakenesses.

Where possible, we will make qualitative comparisons about the technical capabilities, usability and flexibility of different systems presented in academic writing and used in the industry. It is notable, however, that scientific works in the field rarely comment on the 'soft' factors of the presented methods in a way that allows for interpretable and actionable comparisons. Similarly, most industry systems are kept secret to preserve a competitive advatange and thus comparisons there, too, are very difficult.

5 Conclusions

Natural language generation employs data in multiple roles. First, structured data is needed to act as the input of the system. Second, pairs of structured data and human-written texts based on that data can be used for evaluation and for training of 'end-to-end' NLG systems. Third, a wide variety of different datasets can be used to train machine learning components for use as sub-components of NLG systems.

For use in WP5, we have identified several datasets:

- A large *statistical dataset from Eurostat* is to be used as *system input*, as it fulfills acceptably both scientific and journalistic requirements.
- A corpus of *news texts from STT* is used as a starting point for research into incorporating *machine learning* components, and also as a *source of qualitative examples*. Similar datasets are to be compiled for other languages and from the other media partners if the developed methods necessitate their use.
- A small annotated corpus with both *creative and non-creative headlines from STT* is used in a *qualitative* fashion to better understand both the journalistic process in general as well as how creativity is used in journalism. Similar datasets are to be elicited from the other media partners if they become necessary.

- The *unfun.me* and *humicroedit* corpora can be used as training material for *machine learning* methods in relation to the creativity-related tasks within WP5.

For evaluating the NLG methods, including text structuring, we will conduct *an intrinsic human evaluations method*, where judges evaluate the output of case-study NLG systems on the 'Credibility', 'Liking', 'Quality', and 'Representativeness' axes. These axes, identified by Sundar (1999), are closest to a standard that research into automated news production has. In cases where human evaluations are limited by availability of online judges speaking the relevant languages, we will complement quantitative results with *qualitative analyses*. Similar intrisic human evaluation is to be used for the work on incorporating creativity into the system.

For the purposes of *automated evaluation* of NLG, no suitable dataset consisting of aligned input-output pairs in the statistical news domain has been located and producing such a dataset is prohibitively expensive. At the same time, if suitable datasets for automated evaluation are produced by some third party, we intend to use them to measure system performance using standard evaluation metrics such as BLEU (Papineni et al., 2002) and ROUGE (Lin, 2004) alongside these human evaluations. Notably, research into (creative) NLG, as applied to news production, is complicated and the setups not standardized. As such, we are not aware of any other works that would be directly applicable as quantitative benchmarks, i.e. that would allow direct comparison of numeric values obtained as evaluation results to determine which system is 'best'.

References

Alnajjar, K., Leppänen, L., & 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* (pp. 258–265).

Belz, A., & Reiter, E. (2006). Comparing automatic and human evaluation of NLG systems. In *11th Conference of the European Chapter of the Association for Computational Linguistics.*

Clerwall, C. (2014). Enter the robot journalist: Users' perceptions of automated content. *Journalism Practice*, *8*(5), 519–531.

Colton, S. (2008). Creativity Versus the Perception of Creativity in Computational Systems. In *AAAI Spring Symposium: Creative Intelligent Systems* (p. 14—20). Stanford, California, USA.

Colton, S., Charnley, J., & Pease, A. (2011). Computational creativity theory: the face and idea descriptive models. In D. Ventura, P. Gervas, D. Harrell, M. Maher, A. Pease, & G. Wiggins (Eds.), *Proceedings of the second international conference on computational creativity* (pp. 90–95). Universidad Autonoma Metropolitana / Unidad Cuajimalpa, Division de Ciencias de la Comunicacion y Diseno.

Denkowski, M., & Lavie, A. (2014). Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the ninth workshop on statistical machine translation* (pp. 376–380).

Diakopoulos, N. (2019). *Automating the News: How Algorithms Are Rewriting the Media*. Harvard University Press.

Eskenazi, M., Levow, G.-A., Meng, H., Parent, G., & Suendermann, D. (2013). *Crowdsourcing for speech processing: Applications to data collection, transcription and assessment*. John Wiley & Sons.

Gatt, A., & Krahmer, E. (2018). Survey of the state of the art in natural language generation: Core tasks, applications and evaluation. *Journal of Artificial Intelligence Research*, *61*, 65–170.

Gatti, L., Özbal, G., Guerini, M., Stock, O., & Strapparava, C. (2015). Slogans are not forever: Adapting linguistic expressions to the news. In *Proceedings of the 24th international conference on artificial intelligence* (pp. 2452–2458). AAAI Press.

Graefe, A., Haim, M., Haarmann, B., & Brosius, H.-B. (2018). Readers' perception of computer-generated news: Credibility, expertise, and readability. *Journalism*, *19*(5), 595–610.

Halevy, A. Y., & McGregor, S. (2012). Data Management for Journalism. *Bulletin of the IEEE Computer Society Technical Committee on Data Engineering*, *35*(3), 7–15.

Hossain, N., Krumm, J., & Gamon, M. (2019). "President Vows to Cut <Taxes> Hair": Dataset and Analysis of Creative Text Editing for Humorous Headlines. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (pp. 133–142).

Hossain, N., Krumm, J., Vanderwende, L., Horvitz, E., & Kautz, H. (2017). Filling the Blanks (hint: plural noun) for Mad Libs Humor. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing"* (pp. 638–647). Copenhagen, Denmark: Association for Computational Linguistics. doi: 10.18653/v1/D17-1067

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.

Jordanous, A. (2016). Four perspectives on computational creativity in theory and in practice. *Connection Science*, *28*(2), 194–216.

Jung, J., Song, H., Kim, Y., Im, H., & Oh, S. (2017). Intrusion of software robots into journalism: The public's and journalists' perceptions of news written by algorithms and human journalists. *Computers in human behavior*, *71*, 291–298.

Lavie, A., & Agarwal, A. (2007). METEOR: An automatic metric for MT evaluation with high levels of correlation with human judgments. In *Proceedings of the Second Workshop on Statistical Machine Translation* (pp. 228–231).

Liang, P., Jordan, M. I., & Klein, D. (2009). Learning semantic correspondences with less supervision. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1-Volume 1* (pp. 91–99).

Lin, C.-Y. (2004). ROUGE: A package for automatic evaluation of summaries. In *Text summarization branches out* (pp. 74–81).

Lindén, C.-G. (2017). Algorithms for journalism: The future of news work. *The Journal of Media Innovations*, *4*(1), 60–76.

Linkola, S., Kantosalo, A., Männistö, T., & Toivonen, H. (2017, 23). Aspects of self-awareness: An anatomy of metacreative systems. In A. Goel, A. Jordanous, & A. Pease (Eds.), *Proceedings of the 8th international conference on computational creativity (iccc'17)* (pp. 189–196). Georgia: Georgia Institute of Technology.

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).

Magnusson, M., Finnäs, J., & Wallentin, L. (2016). Finding the news lead in the data haystack: Automated local data journalism using crime data. In *Computation+ Journalism Symposium.*

Mairesse, F., Gašić, M., Jurčíček, F., Keizer, S., Thomson, B., Yu, K., & Young, S. (2010). Phrase-based statistical language generation using graphical models and active learning. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics* (pp. 1552–1561).

Melin, M., Bäck, A., Södergård, C., Munezero, M. D., Leppänen, L. J., & Toivonen, H. (2018). No Landslide for the Human Journalist - An Empirical Study of Computer-Generated Election News in Finland. *IEEE Access*, *6*, 43356–43367.

Newman, N., Fletcher, R., Kalogeropoulos, A., Levy, D. A., & Nielsen, R.-K. (2016). Digital news report 2016. *Reuters Institute for the study of Journalism*.

Novikova, J., Dušek, O., Curry, A. C., & Rieser, V. (2017). Why We Need New Evaluation Metrics for NLG. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing* (pp. 2241–2252).

Novikova, J., Dušek, O., & Rieser, V. (2017). The E2E Dataset: New Challenges For End-to-End Generation. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue* (pp. 201–206).

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics* (pp. 311–318).

Pavlick, E., Post, M., Irvine, A., Kachaev, D., & Callison-Burch, C. (2014). The language demographics of amazon mechanical turk. *Transactions of the Association for Computational Linguistics*, *2*, 79–92.

Pipino, L. L., Lee, Y. W., & Wang, R. Y. (2002). Data quality assessment. *Communications of the ACM*, *45*(4), 211–218.

Portet, F., Reiter, E., Gatt, A., Hunter, J., Sripada, S., Freer, Y., & Sykes, C. (2009). Automatic generation of textual summaries from neonatal intensive care data. *Artificial Intelligence*, *173*(7-8), 789–816.

Post, M. (2018). A call for clarity in reporting BLEU scores. *arXiv preprint arXiv:1804.08771*.

Reiter, E. (2018). A Structured Review of the Validity of BLEU. *Computational Linguistics*, *44*(3), 393–401.

Reiter, E., Robertson, R., & Osman, L. M. (2003). Lessons from a failure: Generating tailored smoking cessation letters. *Artificial Intelligence*, *144*(1-2), 41–58.

Reiter, E., Sripada, S., Hunter, J., Yu, J., & Davy, I. (2005). Choosing words in computer-generated weather forecasts. *Artificial Intelligence*, *167*(1-2), 137–169.

Stavelin, E. (2014). *Computational Journalism. When journalism meets programming* (PhD thesis). University of Bergen.

STT. (n.d.). *Finnish News Agency Archive 1992-2018, source* [text corpus]. Kielipankki. Retrieved from `http://urn.fi/urn:nbn:fi:lb-2019041501`

Sundar, S. S. (1999). Exploring receivers' criteria for perception of print and online news. *Journalism & Mass Communication Quarterly*, *76*(2), 373–386.

van der Kaa, H., & Krahmer, E. (2014). Journalist versus news consumer: The perceived credibility of machine written news. In *Proceedings of the Computation+ Journalism Conference, Columbia University, New York* (Vol. 24, p. 25).

Vedantam, R., Lawrence Zitnick, C., & Parikh, D. (2015). Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4566–4575).

West, R., & Horvitz, E. (2019). Reverse-Engineering Satire, or "Paper on Computational Humor Accepted Despite Making Serious Advances". In *Proceedings of the 33rd AAAI Conference on Artificial Intelligence.*

Wiseman, S., Shieber, S., & Rush, A. (2017). Challenges in Data-to-Document Generation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing* (pp. 2253–2263).

Wölker, A., & Powell, T. E. (2018). Algorithms in the newsroom? News readers' perceived credibility and selection of automated journalism. *Journalism*. (Available online as a pre-print) doi: 10.1177/1464884918757072