Helping News Editors Write Better Headlines:
A Recommender to Improve the Keyword Contents & Shareability of News Headlines

Terrence Szymanski, Claudia Orellana-Rodriguez, Mark T. Keane
Insight Centre for Data Analytics &
School of Computer Science,
University College Dublin
{terrence.szymanski,claudia.orellana,mark.keane}@insight-centre.org

Abstract

We present a software tool that employs state-of-the-art natural language processing (NLP) and machine learning techniques to help newspaper editors compose effective headlines for online publication. The system identifies the most salient keywords in a news article and ranks them based on both their overall popularity and their direct relevance to the article. The system also uses a supervised regression model to identify headlines that are likely to be widely shared on social media. The user interface is designed to simplify and speed the editor's decision process on the composition of the headline. As such, the tool provides an efficient way to combine the benefits of automated predictors of engagement and search-engine optimization (SEO) with human judgments of overall headline quality.

1 Introduction

The headline is an extremely important component of every news article that performs multiple functions: summarizing the story, attracting attention, and signaling the voice and style of the newspaper [Conboy, 2007]. In the online realm, headlines are expected to meet several new functions; for instance, to convey the article's contents in different online contexts or to optimize the article for search engine queries (i.e, SEO). Indeed, arguably, the headline is now more important than ever, as it becomes the only visible part of the article in microblog posts, social media feeds and listings on news-aggregation sites. These multiple requirements on the news headline have complicated the composition task facing news editors, as they attempt to ensure that each headline is crafted as perfectly as possible.

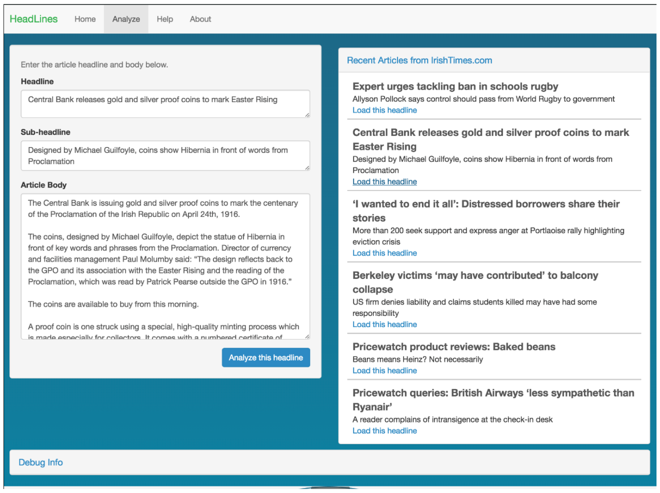
Prior NLP work in the area of news headlines has mostly focused on the task of automatic headline generation, cast as "very short summary generation" in the DUC tasks of the early 2000s; tasks that produced much of the research on the topic. The best-performing system in the 2004 DUC task worked by parsing the first sentence of the article and pruning it to the desired length [Zajic *et al.*, 2004], an approach that works by leveraging human intelligence: journalists generally compose news articles in the "inverted pyramid" style, which places the most important information in the lead paragraph [Conboy, 2007]. Other headline generation systems generally work by first using some metric to identify terms within the document that are likely to appear in the headline, and then constructing a headline containing these terms [Nenkova and McKeown, 2011].

This latter approach has much in common with the task of keyword selection for SEO, which first caught the attention of major newspapers at least ten years ago [Lohr, 2006], and continues to be a much-discussed issue today [Sullivan, 2015]. While even long-established, traditional news publications have begun to move away from classical forms of headlines towards more direct, keyword-laden headlines, many copy editors would still prefer to write clever, witty headlines [Wheeler, 2011], and readers of the news seem to value creativity in headlines over clarity or informativeness [Ifantidou, 2009]. Therefore, one of the key considerations in the design of our system was to balance the mechanical act of filling a headline with informative, relevant keywords, against the creative act of writing headlines that appeal to human interests and emotions.

We expect that the most interesting and emotional stories are likely to be more popular with readers than the "average" story. Analysis of reader behavior has shown that there is no correlation between how much an article is shared on social media and how much of the article is read by an average user [Haile, 2014]; a fact that could be taken as evidence supporting the widely-held view that people share articles online that they have not fully read themselves [Manjoo, 2013]. In this case, the headline—which people presumably read even if they don't read the full text—may be an important factor in determining the "shareability" of a news article; an idea that is another key motivation behind the design of our system.

The tool presented here is designed to facilitate the decision-making process facing a news editor in composing a headline. The software employs state-of-the-art NLP and machine learning techniques to make its recommendations, but it is not designed to automatically generate headlines or to make decisions about a headline's goodness on its own.

In the sections below, we present the design and behavior of the tool before discussing the internal workings of the system. We conclude with an assessment of the current state of the project, including some preliminary evaluation results and a discussion of areas for improvement.



Enter the article headline and body below.

Headline

Central Bank releases gold and silver proof coins to mark Easter Rising

Sub-headline

Designed by Michael Guilfoyle, coins show Hibernia in front of words from Proclamation

Article Body

The Central Bank is issuing gold and silver proof coins to mark the centenary of the Proclamation of the Irish Republic on April 24th, 1916.

The coins, designed by Michael Guilfoyle, depict the statue of Hibernia in front of key words and phrases from the Proclamation. Director of currency and facilities management Paul Molumby said: "The design reflects back to the GPO and its association with the Easter Rising and the reading of the Proclamation, which was read by Patrick Pearse outside the GPO in 1916."

The coins are available to buy from this morning.

A proof coin is one struck using a special, high-quality minting process which is made especially for collectors. It comes with a numbered certificate of

Recent Articles from IrishTimes.com

Expert urges tackling ban in schools rugby
Allyson Pollock says control should pass from World Rugby to government
Load this headline

Central Bank releases gold and silver proof coins to mark Easter Rising
Designed by Michael Guilfoyle, coins show Hibernia in front of words from Proclamation
Load this headline

'I wanted to end it all': Distressed borrowers share their stories
More than 200 seek support and express anger at Portlaoise rally highlighting eviction crisis
Load this headline

Berkeley victims 'may have contributed' to balcony collapse
US firm denies liability and claims students killed may have had some responsibility
Load this headline

Pricewatch product reviews: Baked beans
Beans means Heinz? Not necessarily
Load this headline

Pricewatch queries: British Airways 'less sympathetic than Ryanair'
A reader complains of intransigence at the check-in desk
Load this headline

Figure 1: Screenshot of the tool in input mode, with input text areas on the left and a live feed of articles on the right.

2 Design and Behavior

From a user-interface perspective, the software has two modes of operation: input mode and analysis mode. The input mode (illustrated in Figure 1) facilitates the entry of a news article and its corresponding headline and sub-headline, which may either be entered manually or selected from a feed of recent articles. In practice, this feed would be integrated into the newspaper's workflow so that an editor could review all new articles with the software prior to publication.

After the editor-user selects an article, the system switches to the analysis mode, showing the results of the automated analysis (illustrated in Figure 2). This mode is designed to allow the user to quickly assess the strengths and weaknesses of the headline and decide whether any changes should be made to improve it. The five most highly-ranked keywords from the article are listed on the right side of the screen sorted by *weight*, a metric combining the keyword's *frequency* in the article and its *SEO score*, which respectively capture the keyword's local relevance to the article itself as well as its global prominence among news stories in general. (See section 3.1 below for details on these measures.)

The keywords are color-coded to distinguish keywords which already appear in the headline (green) from those which do not appear in the headline (red), and size-coded according to their weight. Thus, any large, red keywords are those which an editor should consider adding to the headline. In the example in Figure 2, the top three recommended keywords are already present in the headline; the two remaining recommendations, "Irish Republic" and "GPO", are both sensible suggestions for the article.

In addition to the keyword recommendations, the system scores each headline for its "shareability" on two social media platforms: Twitter and Facebook; if the shareability score on either platform exceeds a threshold value, then an alert is displayed to the user. In the example in Figure 2, the article has exceeded the Facebook threshold but not the Twitter threshold, so only one of the two alerts is displayed. The newspaper's editor in charge of social media can use this information when deciding which stories should be posted and promoted on social media sites. The threshold is set to a relatively conservative value, so that most articles will not produce alerts, and only the most promising headlines will come to the editor's attention.

Ultimately, it is up to the editor to decide what action, if any, to take based on the information presented by the software. The editor has the leeway to add keywords in the headline in creative ways that fit the style of the story and the news organization, and she can also flexibly deal with any errors that may be produced by the keyword recommendation system, rather than blindly following its advice.

3 Implementation

The system consists of three major components: a user-interface front-end, a text analysis back-end, and a web server that mediates communication between the two. The user interface is implemented with HTML and Javascript and accessed via a web browser; its behavior is described and illustrated in the previous section. The web server is implemented in Python (based on the *Flask* framework), which allows easy integration with the text analytic back-end, which is also mainly implemented in Python. We use the *sklearn* module for regression and Stanford's *CoreNLP* Java suite for NLP [Manning *et al.*, 2014]. The entire system is deployed on a web server and accessed by the client's web browser.

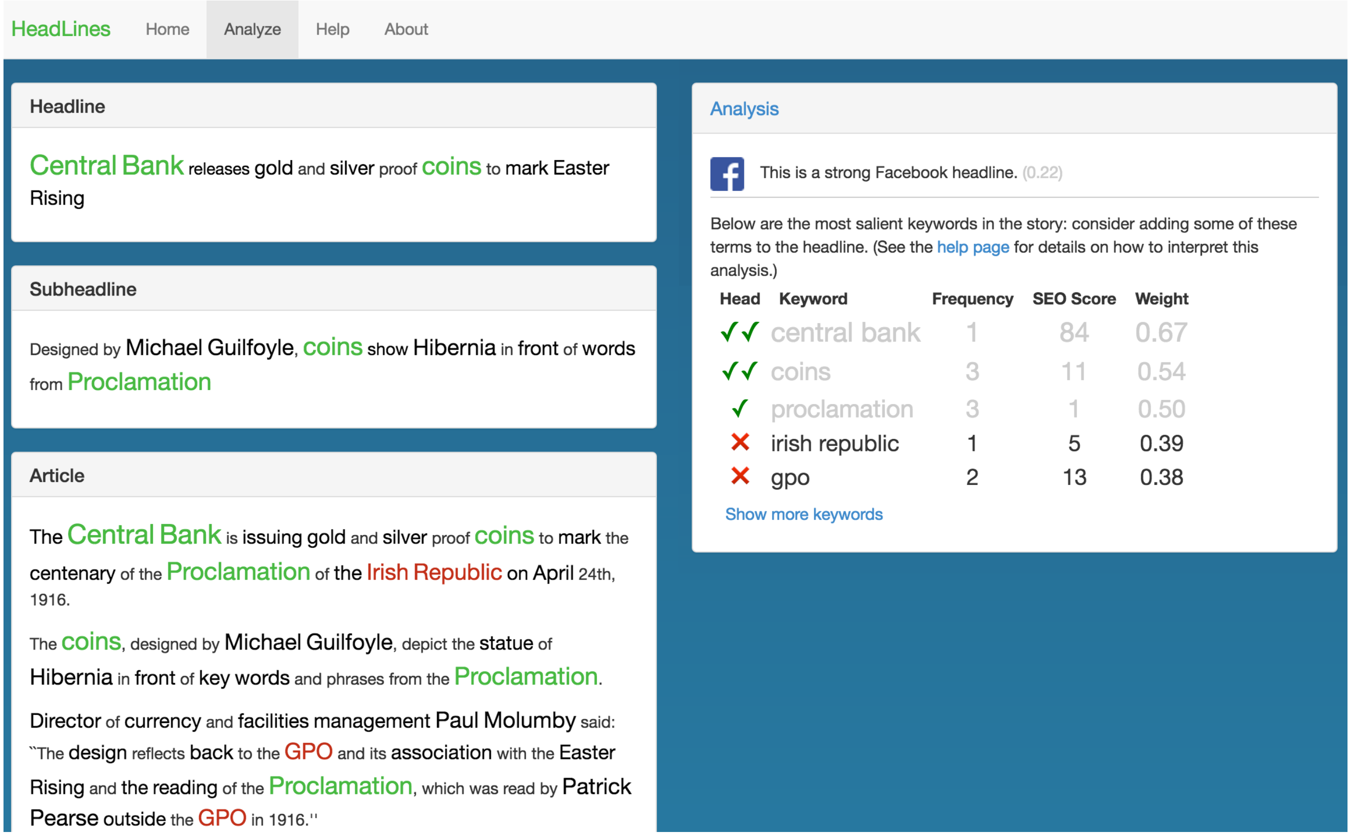
The back-end consists of two components—keyword analysis and shareability analysis—which operate independently of one another and are discussed in detail below.

3.1 Keyword Analysis

The role of keyword analysis is to identify terms in the article body that are good candidates for inclusion in headline. We believe that headlines containing informative and popular keywords can be both more appealing to readers and more prominent in users' search results and on news aggregator websites.

Processing of an input article begins with tokenization and named-entity recognition using *CoreNLP*, which identifies all entities (e.g. people, organizations, locations) in the article. Next, any known keywords appearing in the text are identified, by using a database of 90k keywords and their frequencies from Irish news articles in recent years, which we populated with data provided to us by two other Irish news-related projects [Shi *et al.*, 2014; Bordea *et al.*, 2013]. This process results in a list of entities, which may be unique to the given article, and a list of keywords, which are known to have been encountered in previous news articles. These keywords and named entities are linked using a simple, rule-based approach that resolves pairs like "Enda Kenny" and "(Mr.) Kenny", yielding a single list of resolved keywords, along with a list of all positions in the text where each keyword appears.

Our keyword ranking system aims to capture the intuition that salient keywords should ideally be both *locally* prominent (i.e. appearing frequently in the given news article) and *globally* popular (i.e. appearing frequently in articles other then the current one). Thus, we calculate the weight w of



Headline

Central Bank releases gold and silver proof coins to mark Easter Rising

Subheadline

Designed by Michael Guilfoyle, coins show Hibernia in front of words from Proclamation

Article

The Central Bank is issuing gold and silver proof coins to mark the centenary of the Proclamation of the Irish Republic on April 24th, 1916.

The coins, designed by Michael Guilfoyle, depict the statue of Hibernia in front of key words and phrases from the Proclamation.

Director of currency and facilities management Paul Molumby said: "The design reflects back to the GPO and its association with the Easter Rising and the reading of the Proclamation, which was read by Patrick Pearse outside the GPO in 1916."

Analysis

This is a strong Facebook headline. (0.22)

Below are the most salient keywords in the story: consider adding some of these terms to the headline. (See the help page for details on how to interpret this analysis.)

Head	Keyword	Frequency	SEO Score	Weight
✓✓	central bank	1	84	0.67
✓✓	coins	3	11	0.54
✓	proclamation	3	1	0.50
✗	irish republic	1	5	0.39
✗	gpo	2	13	0.38

Show more keywords

Figure 2: Screenshot of the tool in analysis mode. In this example, three of the top five keywords (highlighted in green) are already in the headline; the remaining two (in red) are recommendations that the user may consider adding to the headline.

each keyword k in the document d as the weighted sum of its local weight w_{local} and its global weight w_{global}:

$$w(k, d) = \lambda\, w_{local}(k, d) + (1 - \lambda)\, w_{global}(k)$$

The local weight is calculated as the normalized within-document frequency of the keyword, so that the most frequent keyword in the document gets a w_{local} of 1. The global weight is calculated in a similar way, using the across-document frequencies from the keyword database and applying a nonlinear (log) transformation to compensate for the highly skewed distribution of these frequencies (note that it is possible for a keyword to have a zero global weight if it does not appear in our database; this is common for named entities in the article which have not been mentioned in the news before). The relative contributions of the local and global weights are balanced with the λ parameter.

This formula was chosen as the simplest method (a linear combination) of combining the two factors. It is similar to a *tf-idf* score in that it combines both term frequency and document frequency, but it is critically different in that it rewards, rather than penalizes, terms that occur in many documents. This is a good thing because we believe that terms which may be very common (e.g. the names of well-known politicians or celebrities) can be good headline terms, and also because our method of selecting terms (via a closed set of keywords and automatic named entity detection) generally avoids selecting words which may be high-frequency but low-quality (like stopwords).

We manually set the value of λ to achieve rankings which we subjectively deemed to be suitable.[1] This manual parameter setting allowed us to deploy our system quickly with acceptable performance, but a better option would be to learn these parameters automatically. To do so would require a dataset containing news articles, their headlines, and either some measure of the quality of the headline or an assurance that the headlines in the data set are "good", in order to ensure that the parameters are set based on "good" headline examples. This type of data was not available to us when the system was under development.

This method ultimately assigns a weight to each keyword between 0 and 1.0, which determines its ranking in the analysis output (Figure 2). In the user interface, the weight is displayed in a table alongside the keyword's "frequency" and "SEO Score", which we consider to be more user-friendly

[1] We found that a value of 0.6 (i.e. slightly favoring local frequency over global frequency) worked well for our data, but this value changed depending on which keyword list we used. Ultimately, we combined both keyword lists, which introduced a large number of noisy terms. To suppress these noisy terms, we added an additional term to boost the score of keywords which were identified as named entities in the article (up to 0.2 of the overall weight).

than w_{local} and w_{global} themselves (the frequency is exactly the number of times the term appears in the article, and the SEO score is just w_{global} scaled to the familiar scale of 0 to 100).

3.2 Shareability Analysis

The role of shareability analysis is to identify headlines that are likely to be shared on social media. With the rise of social media as dissemination channels for the news, headlines now need to be both informative and "shareable"; that is, the headline somehow needs to attract people to post, share, and engage with the article on social media, in order to reach a large online audience.

According to the Reuters Institute Digital News Report [Kirk *et al.*, 2015], Facebook and Twitter generate 52% of the visits to online news sites, suggesting that direct visits to the home pages of news providers are being supplanted by social media mediated access. However, Facebook and Twitter are known to have quite different audiences and engage users in different ways [Kirk *et al.*, 2015]. Users on Twitter generally actively search for news, whereas on Facebook, news tends to be just encountered by sharing amongst friends. Therefore, in our system, we model the two social networks separately.

Using the Twitter streaming API we collected over 700k tweets and retweets posted by each one of 200 media outlets and journalist accounts for two time periods in 2013 and 2014, for a duration of 71 and 50 days, respectively. From the collected tweets we extracted all the URLs and used the Facebook and Twitter APIs to collect the number of times each URL was shared on Facebook or posted on Twitter. Because these posts were made by journalists, the links in the tweets are mainly to news articles, from which we extracted headlines. This step yielded a data set of 55k headlines with corresponding counts of social shares for each one.

We used a regression analysis to estimate the relationship between features of the headlines and the target variable of number of shares. Each headline in our collection is represented as a vector consisting of eight features covering three main aspects of the headline's content: the sentiment polarity (as computed by the *TextBlob* Python package), the presence of named entities, and the length in words. The complete list of features is presented in Table 1.

Feature	**Description**
neutral	# of neutral sentiment words
positive	# of positive sentiment words
negative	# of negative sentiment words
organizations	# of ORGANIZATION entities
persons	# of PERSON entities
places	# of LOCATION entities
day	(T/F) headline contains the name of a day
length	total # of words in the headline

Table 1: Headline features used for regression. The first six features are normalized by the length of the headline.

We used Regularized Linear Regression (RLR), Random Forest (RF) and Gradient Boosting Trees (GBT) as our methods for regression and used the metric Mean Squared Error (MSE) to assess their performance. We split our headlines set into 44k (80%) for training and the remaining 11k (20%) for testing. We train two different regression models, one for Facebook and one for Twitter. RF and GBT performed better than the RLR models. Between RF and GBT models, GBT performed slightly better than RF, although no significant difference was observed. On the basis of these results we use GBT as our method for regression. GBT have shown to outperform other models in classification and regression tasks and have been used successfully for audience engagement prediction [Diaz-Aviles *et al.*, 2014]. We observe that the models for Twitter and Facebook behave differently: comparing the values of the MSE for both models, predictions for shareable headlines on Facebook present an MSE of 41.8, while for Twitter the error is slightly smaller, 37.6.

Once the GBT models are trained, we store them and incorporate them into the system's pipeline. Every inputted headline receives two shareability scores, one for each social media site; however, in order to avoid triggering too many notifications to the journalist or news editor, the system only shows a result if the score is equal or larger than a manually-defined threshold of 3.7 and 1.7 for Facebook and Twitter, respectively, which correspond to the median number of shares (on each platform) received by the headlines in our collection.

4 Evaluation

The tool was developed in collaboration with The Irish Times, and several professional editors have tested its usability. The feedback from these sessions has been positive and has informed several design features. In particular, the color-coding and font-size features of the interface have been noted for their usability. On the basis of this success, we are now looking at integration into editors' daily workflow, to allow more usability data to be gathered.

Current tests of the system have identified some potential areas for improvement. The keyword system commonly fails to recognize when pairs of equivalent but non-identical keywords have the same referent; for example *Taioseach* and *Enda Kenny*, or *GPO* and *General Post Office*. While editors easily recognize this duplication, this error affects frequency counts, which in turn affect keyword rankings. This type of co-reference resolution is an open question in NLP research, with typical solutions relying on a rule-based or gazette-based approach to fix commonly-occurring cases.

The system could also be improved by moving from a static keyword database to a dynamic, real-time database. We were fortunate to be able to bootstrap our system with the keyword sources discussed in section 3.1; however neither of these sources were created with this specific use-case in mind, and the static nature of these lists means that the keyword database will become outdated over time. Updating the keyword frequency counts on a rolling basis is an easy first step; but a more sophisticated approach is probably required, where new entities are added to the database over time, and more recent articles are given a greater weight than older articles. Because our system already identifies named entities in news articles, these entities could be added as new keywords in our database as they are encountered.

We are also evaluating the impact of using the tool on SEO, based on determining whether it improves article rankings in news aggregators and search engines. While the lack of click-through from GoogleNews has led some to question its effectiveness at driving traffic to news sites [Wauters, 2010], for *The Irish Times'* website, GoogleNews is a major source of referrals. An analysis of 30k *Irish Times* articles (from 1/10/15 to 31/3/2016) has shown that articles listed on the GoogleNews (Irish edition) front pages received significantly ($p < 0.01$) more page views than unlisted articles; with GoogleNews listed articles receiving almost twice as many views ($n = 11,125$, $\mu = 1665.5$ views per article) as unlisted articles ($n = 19,339$, $\mu = 892.4$ views per article). GoogleNews' ranking algorithm is not publicly known, so the exact factors leading to this correlation are opaque; however, for practical purposes, if our keyword recommender leads to greater visibility on GoogleNews, then we know it should increase readership.

Finally, the quality of our keyword recommendations can, in part, be assessed by noting whether the system's top-recommended keywords are already present in the original headline written for the article, as it shows that the system corresponds to human judgments (n.b., the keyword analysis only uses the article body, not the headline, as input). We processed a sample of roughly 3,000 Irish Times headlines with our system, and found that a majority of these (64%) contained two or more of the top five keywords recommended by our system (in either the headline or the sub-headline), and a large majority (88%) contained at least one of the recommended keywords. We take this as evidence that the keywords recommended by our system generally correspond with the types of keywords that a human editor would normally include in the headline.

5 Conclusion

In this paper, we have presented a system for recommending keywords for inclusion in newspaper headlines and for identifying headlines with high potential shareability on social media. The system identifies plausible keywords that are both relevant to the given news article and popular overall in past news articles, in an effort to maximize both the reader interest and the SEO aspect of the headline. In addition, the system identifies headlines that are likely to receive above-average engagement on social media, allowing editors to effectively target their social media strategy. We believe that this tool can be a helpful component in modern, online-oriented newsrooms.

6 Acknowledgments

The authors would like to thank *The Irish Times* for their funding and help on this project. This work is supported by Science Foundation Ireland through the Insight Centre for Data Analytics under grant number SFI/12/RC/2289.

References

[Bordea *et al.*, 2013] G. Bordea, P. Buitelaar, and T. Polajnar. Domain-independent term extraction through domain modelling. In *Proceedings of TIA*, 2013.

[Conboy, 2007] M. Conboy. *The Language of the News*. Routledge, 2007.

[Diaz-Aviles *et al.*, 2014] E. Diaz-Aviles, H. T. Lam, F. Pinelli, S. Braghin, Y. Gkoufas, M. Berlingerio, and F. Calabrese. Predicting user engagement in Twitter with collaborative ranking. In *Proceedings of the 2014 RecSys Challenge*, 2014.

[Haile, 2014] T. Haile. What you think you know about the web is wrong. *Time Magazine*, 2014.

[Ifantidou, 2009] E. Ifantidou. Newspaper headlines and relevance: Ad hoc concepts in ad hoc contexts. *Journal of Pragmatics*, 41(4):699–720, 2009.

[Kirk *et al.*, 2015] N. Kirk, J. Suiter, and P. McNamara. *Reuters Institute Digital News Report (Ireland).* 2015.

[Lohr, 2006] S. Lohr. This boring headline is written for Google. *The New York Times*, April 9 2006.

[Manjoo, 2013] F. Manjoo. You won't finish this article. *Slate Magazine*, June 6 2013.

[Manning *et al.*, 2014] C. D. Manning, M. Surdeanu, et al. The Stanford CoreNLP natural language processing toolkit. In *ACL System Demonstrations*, 2014.

[Nenkova and McKeown, 2011] A. Nenkova and K. McKeown. Automatic summarization. *Foundations and Trends in Information Retrieval*, 5(2–3), 2011.

[Shi *et al.*, 2014] B. Shi, G. Ifrim, and N. Hurley. Be in the know: Connecting news articles to relevant Twitter conversations. In *Proceedings of ECML*, 2014.

[Sullivan, 2015] M. Sullivan. Hey, Google! Check out this column on headlines. *The New York Times*, April 18 2015.

[Wauters, 2010] R. Wauters. Report: 44% of Google News visitors scan headlines, don't click through. *TechCrunch*, January 19 2010.

[Wheeler, 2011] D. R. Wheeler. 'Google doesn't laugh': Saving witty headlines in the age of SEO. *The Atlantic*, May 11 2011.

[Zajic *et al.*, 2004] D. Zajic, B. J. Dorr, and R. Schwartz. BBN/UMD at DUC-2004: Topiary. In *Proceedings of DUC*, 2004.