WikiTopics: What is popular on Wikipedia and why.

Abstract

We establish a novel task in the spirit of news summarization and topic detection and tracking (TDT): daily determination of the topics newly popular with Wikipedia readers. Central to this effort is a new public dataset consisting of the hourly page view statistics of all Wikipedia articles over the last three years. We give baseline results for the tasks of: discovering individual pages of interest, clustering these pages into coherent topics, and extracting the most relevant summarizing sentence for the reader. When compared to human judgements, our system shows the viability of this task, and opens the door to a range of exciting future work.

Barack Obama
Joe Biden
White House
Inauguration
...
US Airways Flight 1549
Chesley Sullenberger
Hudson River
...
Super Bowl
Arizona Cardinals

Figure 1: Automatically selected articles for Jan 27, 2009.

1 Introduction

In this paper we analyze a novel dataset: we have collected the hourly page view statistics[1] for every Wikipedia page in every language for the last three years. We show how these page view statistics along with other features like inter-page links can be used to identify and explain popular trends, including popular films and music, sports championships, elections, natural disasters, etc.

Our approach is to select a set of articles whose daily pageviews for the last fifteen days increase above those of the previous fifteen days. Rather than simply selecting the most popular articles for a given day, this selects articles whose popularity is rapidly increasing. These popularity spikes tend to be due to significant current events in the real world. We examine 100 such articles for each of 5 randomly selected days in 2009 and attempt to group the articles into clusters such that the clusters coherently correspond to current events and extract a summarizing sentence that best explains the relevant event. Quantitative and qualitative analyses are provided along with the evaluation dataset.

We compare our automatically collected articles to those in the current events portal of Wikipedia where Wikipedia editors manually chronicle current events every day, which comprise armed conflicts, international relations, law and crime, natural disasters, social, political, sports events, etc. Each event is summarized into a simple phrase or sentence along with links to related articles. We view our work as an automatic mechanism that could potentially supplant this hand-curated method of selecting current events by editors.

Figure 1 illustrates examples of automatically selected articles for January 27, 2009. We would attempt to groups the articles into 3 clusters, { Barack Obama, Joe Biden, White House, Inauguration } which corresponds to the inauguration of Barack Obama, { US Airways Flight 1549, Chesley Sullenburger, Hudson River } which corresponds to the successful ditching of an airplane into the Hudson river without loss of life, and { Superbowl, Arizona Cardinals } which describes the then upcoming Superbowl XLIII.

We further try to explain the clusters by selecting sentences from the articles. For the first cluster, a good selection would be "the inauguration of Barack Obama as the 44th president ... took place on January 20, 2009". For the second cluster, it would be "Chesley Burnett 'Sully' Sullenberger III (born January 23, 1951) is an American commercial airline pilot, ..., who successfully carried out the emergency water landing of US Airways

[1] http://dammit.lt/wikistats

Flight 1549 on the Hudson River, offshore from Manhattan, New York City, on January 15, 2009, . . . " which provides links to the other articles in the same cluster. For the third cluster, "Superbowl XLIII will feature the American Football Conference champion Pittsburgh Steelers (14-4) and the National Football Conference champion Arizona Cardinals (12-7) ." would be a good choice which delineates the association with Arizona Cardinals.

Different clustering methods and sentence selection features are evaluated and results are compared. Topic models, such as K-means (Manning et al., 2008) clustering in vector space model and latent Dirichlet allocation (Blei et al., 2003) clustering, are compared to clustering using Wikipedia's hyperlink structure. To select sentences we make use of NLP technologies such as coreference resolution, named entity and date taggers. Note that the latest revision of each article as of the day on which the article is selected is used in clustering and textualization to simulate the situation in which article selection, clustering, and textualization are performed once every day.

Figure 2 illustrates the pipeline of our WikiTopics system: article selection, clustering, and textualization.

2 Article selection

We would like to identify an uptrend in popularity of articles. In an online encyclopedia such as Wikipedia, the pageviews for an article reflect its popularity. Following the Trending Topics software[2], WikiTopics's articles selection algorithm determines each articles' monthly trend value as increase in pageviews within last 30 days. The monthly trend value t^k of an article k is defined as below:

$$t^k = \sum_{i=1}^{15} d_i^k - \sum_{i=16}^{30} d_i^k$$

where

$$d_i^k = \text{daily pageviews } i-1 \text{ days ago of an article } k$$

We selected 100 articles of the highest trend value for each day in 2009. We call the articles WikiTopics articles. Many other methods to determine the trend value and choose articles are possible, which we leave as future work[3].

Wikipedia has a portal page called "current events", in which significant current events are listed manually by Wikipedia editors. Figure 3 illustrates spikes in pageviews of the hand-curated articles related to the inauguration of Barack Obama[4], which shows clear correlation between the spikes and the day on which the relevant event took place. It is natural to contrast WikiTopics articles to this set of hand-curated articles. We evaluated WikiTopics articles against hand-curated articles as gold standard and had negative results with precision of 0.13 and recall of 0.28.

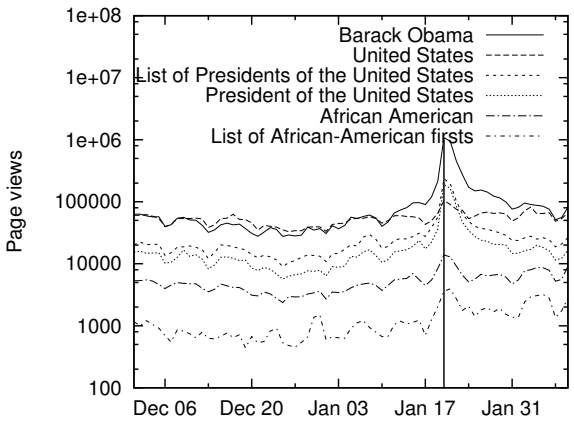

Figure 3: Pageviews for all the hand-curated articles related to the inauguration of Barack Obama. Pageviews spike on the same day as the event took place–January 20, 2009.

There are a few reasons for this. First, there are much fewer hand-curated articles than WikiTopics articles: 17,253 hand-selected articles vs 36,400[5] WikiTopics articles; so precision cannot be higher than 47%. Second, many of the hand-selected articles turned out to have very low pageviews: 6,294 articles (36.5%) have maximum daily pageviews less than 1,000 whereas WikiTopics articles have increase in pageviews of at least 10,000. It is extremely hard to predict the hand-curated articles based on pageviews. Figure 4 further illustrates hand-curated articles' lack of increase in pageviews as opposed to WikiTopics articles. On the contrary, nearly half of the hand-curated articles have decrease in pageviews. For them, spikes in pageviews are rather an exception than a commonality.

It is concluded that it is futile to predict hand-curated articles based on pageviews. The hand-curated articles suffer from low popularity and do not spike in pageviews. We recommend against this methodology for other researchers. Instead, we focus on WikiTopics articles. Figure 5 contrasts the WikiTopics articles and hand-curated articles. The WikiTopics articles shown here do not appear in hand-curated articles within fifteen days before or after, and vice versa. WikiTopics selected articles about people who played a minor role in the relevant event, recently released films, their protagonists, popular TV series, etc. Wikipedia editors selected articles about actions,

[2]http://www.trendingtopics.org

[3]For example, one might leverage additional signals of real world events, such as Twitter feeds, etc.

[4]We set up a website where you can see the sparkline graphs of pageviews for every hand-curated event in 2009: http://ANONYMIZED

[5]One day is missing from our 2009 pageviews statistics.

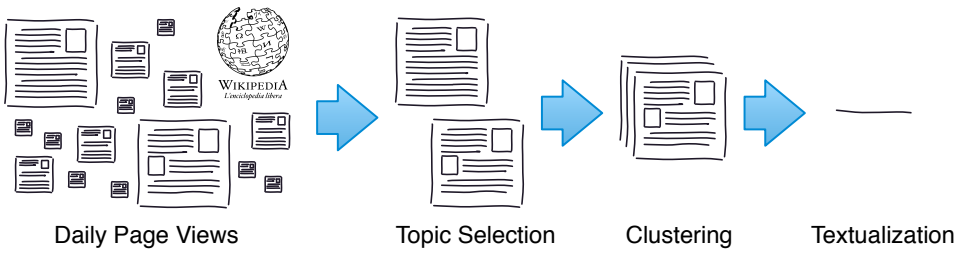

Figure 2: Process diagram: (a) Topic selection: select interesting articles based on increase in pageviews. (b) Clustering: cluster the articles according to relevant events using topic models or Wikipedia's hyperlink structure. (c) Textualization: select the sentence that best summarizes the relevant event.

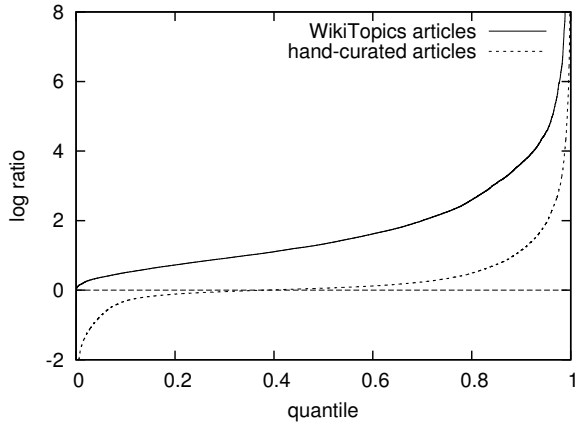

Figure 4: Log ratio of the increase in pageviews: $\log \sum_{i=1}^{15} d_i^k / \sum_{i=16}^{30}$. Zero means no change in pageviews. WikiTopics articles show pageviews increase in a few orders of magnitude as opposed to hand-curated articles.

WikiTopics articles
Joe Biden
Notorious (2009 film)
The Notorious B.I.G.
Lost (TV series)
...
hand-curated articles
Fraud
Florida
Hedge fund
Arthur Nadel
Federal Bureau of Investigation

Figure 5: Illustrative articles for January 27, 2009. WikiTopics articles here do not appear in hand-curated articles within fifteen days before or after, and vice versa. The hand-curated articles shown here are all linked from a single event "Florida hedge fund manager Arthur Nadel is arrested by the United States Federal Bureau of Investigation and charged with fraud."

things, geopolitical or organizational names and wrote event description mentioning all the related articles. We go on further to identify and describe relevant events from WikiTopics articles.

3 Clustering

Clustering plays a central role to identify current events; a group of coherently related articles corresponds to a current event. Clusters, in general, may have hierarchies and an element may be a member of multiple clusters. Whereas Wikipedia's current events are hierarchically compiled into different levels of events, we focus on flat clustering, leaving hierarchical clustering as future work, but allow multiple memberships. In addition to clustering using Wikipedia's inter-page hyperlink structure, we experimented with two families of clustering algorithms pertaining to topic models: K-means clustering in vector space model and latent Dirichlet allocation (LDA) probabilistic topic model. We used the Mallet software (McCallum, 2002) to run the topic models. The latest revision of each article as of the day on which WikiTopics selected the article was retrieved, with unnecessary HTML tags and Wiki templates stripped with mwlib[6] and sentences split with NLTK (Loper and Bird, 2002). Normalization, tokenization, and stop words removal were performed. The unigram (bag-of-words) model was used and the number of clusters/topics k was set to 50, which is half the number of articles. For K-means, the common settings were used: tf and tf-idf weighting and cosine similarity (Allan et al., 2000). For LDA, we chose the most probable topic for each article as the cluster ID. Two different clustering schemes make use of the link structure: ConnComp and OneHop. We treat inter-page links as undirected edges. ConnComp groups a set of articles into the same connected component. OneHop chooses an article and groups a set of articles within one hop away following links. The number of resulting clusters depends on the order in which you choose the next article to cluster. To find the minimum or maximum number of such

[6] http://code.pediapress.com/wiki/wiki/mwlib

Test set	# Clusters	B^3 F-score
Human-1	48.6	0.704 ± 0.083
Human-2	50.0	0.710 ± 0.108
Human-3	53.8	$\mathbf{0.741 \pm 0.103}$
ConnComp	31.8	0.424 ± 0.183
OneHop	45.2	0.580 ± 0.172
K-means tf	50	0.521 ± 0.042
K-means tf-idf	50	$\mathbf{0.584 \pm 0.089}$
LDA	44.8	0.426 ± 0.080

Table 1: Clustering evaluation: F-scores are averaged across gold standard datasets. ConnComp and OneHop are using the link structure. K-means clustering with tf-idf performs best.

clusters is NP-complete. Instead of attempting to find the optimal number of clusters, we iteratively create clusters that maximize the central node connectivity, stopping when all nodes are in at least one cluster.[7]

Three annotators manually clustered WikiTopics articles for five randomly selected days. The three manual clusters were evaluated against each other to measure inter-annotator agreement, using the multiplicity B^3 metric (Amigó et al., 2009). Table 1 shows the results. The B^3 metric is one of the extrinsic clustering evaluation metrics, which need a gold standard set of *categories* to evaluate against. The multiplicity B^3 works nicely for overlapping clusters: each item e has potentially multiple gold standard categories, and also potentially multiple clusters. Let $C(e)$ be the set of the clusters that e belongs to, and $L(e)$ be the set of e's categories. The multiplicity B^3 scores for a pair of e and e' are evaluated as follows:

$$\text{Prec}(e, e') = \frac{\min\left(|C(e) \cap C(e')|, |L(e) \cap L(e')|\right)}{|C(e) \cap C(e')|}$$

$$\text{Recall}(e, e') = \frac{\min\left(|C(e) \cap C(e')|, |L(e) \cap L(e')|\right)}{|L(e) \cap L(e')|}$$

The overall B^3 scores are evaluated as follows:

$$\text{Prec} = \text{Average}_{e \neq e'} \text{Prec}(e, e')$$

$$\text{Recall} = \text{Average}_{e \neq e'} \text{Recall}(e, e')$$

The inter-annotator agreement in the B^3 scores are in the range of 67%–74%. K-means clustering performs best, achieving 79% precision compared to manual clustering. OneHop clustering using the link structure achieved comparable performance. LDA performed significantly worse, comparable to ConnComp clustering.

Clustering the articles according to the relevance to recent popularity is not trivial even for humans. In WikiTopics articles for February 10, 2009, Journey (band) and Bruce Springsteen may seem to be relevant to Grammy Awards, but in fact they are relevant on this day because of the Super Bowl. K-means fails to recognize this and put them into the cluster of Grammy Awards, while ConnComp merged Grammy Awards and Super Bowl into the same cluster. OneHop kept the two clusters intact and benefited from putting Bruce Springsteen into both the clusters. LDA clustering might have suffered from our allowing only a single membership for an article. Clustering using the link structure performs comparably with other clustering algorithms without using topic models. It is worth noting that there are a few "octopus" articles that have links to many articles. The United States on January 27, 2009 was disastrous, with its links to 58 articles, causing ConnComp clustering to group 89 articles into a single cluster. OneHop clustering's condition that groups only articles that are one hop away alleviates the issue and it also benefited from putting an article into multiple clusters.

4 Textualization

We would like to generate textual descriptions for the clustered articles to explain why they are popular and what event is relevant. The date expressions and the reference to the topic of the article are used as features. Currently, our work is limited to select the best sentence that describes the relevant event for each article; future work will consider generating a summary that describes the relevant event and the relationships of the related articles in a cluster using sentence fusion or paraphrasing.

We preprocess the Wikipedia articles using the Serif system (Boschee et al., 2005) for date tagging and coreference resolution. The identified temporal expressions are in various formats such as exact date ("February 12, 1809"), a season ("spring"), a month ("December 1808"), a date without a specific year ("November 19"), and even relative time ("now", "later that year", "The following year"). Some examples are shown in Figure 6. The entities mentioned in a given article are compiled into a list and the mentions of each entity, including pronouns, are linked to the entity as a co-reference chain.

In our initial scheme, we picked the first sentence of each article because the first sentence is usually an overview of the topic of the article and often relevant to the current event. For example, a person's article often has the first line with one's recent achievement or death. An article about an album or a film often begins with the release date. We call this **First**.

We also picked the sentence with the most recent date to the day on which the article was selected. Dates in the near future are considered in the same way as the recent dates. Dates may vary in their formats, so we put precedence over the formats so that a more specific date i.e. "February 20, 2009" is selected over vaguer dates such

[7] This allows for singleton clusters.

February 12, 1809	September
1860	Later that year
now	November 19
the 17th century	that same month
some time	The following winter
December 1808	The following year
34 years old	April 1865
spring	late 1863

Figure 6: Selected examples of temporal expressions identified by Serif from 247 such date and time expressions extracted from the article Abraham Lincoln.

2009-01-27: Barack Obama

Before: He was inaugurated as President on January 20, 2009.
After: **Obama** was inaugurated as President on January 20, 2009.

2009-05-12: Eminem

Before: He is planning on releasing his first album since 2004, Relapse, on May 15, 2009.
After: **Eminem** is planning on releasing his first album since 2004, Relapse, on May 15, 2009.

Figure 7: Selected examples of sentence selection and pronoun replacement: from each article the best sentence is selected based on the most recent date expression and reference of the topic of the article, and personal pronouns are substituted with their proper names, which are in **bold**.

as "February 2009" or "2009". We call this scheme **Recent**.

As the third scheme, we picked the sentence with the most recent date among those with a reference to the article's topic. The reasoning behind this is if the sentence refers to the topic of the article, it is more likely to be relevant to the current event. We call this scheme **Self**.

After selecting a sentence for each cluster, we substitute personal pronouns in the sentence with their proper names. This step enhances readability of the sentence, which often refers to people by a pronoun such as "he", "his", "she", or "her'. The examples of substituted proper names appear in Figure 7 in bold face. The Serif system classifies which entity mention are proper names, but choosing the best reference for a person is not a trivial task: proper names may vary from *John* to *John Kennedy* to *John Fitzgerald "Jack" Kennedy*. We choose the most frequent proper name.

For ten randomly chosen articles of the five selected days, two annotators selected the sentence that best describes why an article gained popularity recently, among 289 sentences per each article on average. For each article, annotators picked a single best sentence, and possibly multiple second best sentences. If there is no such single sentence that best describes a relevant event, annotators marked none as the best sentence and listed sentences that partially explain the relevant event as second best sentences. The evaluation results for all the selection schemes are shown in Table 2. To see inter-annotator agreement, two annotators' selection was evaluated against each other. The other selection schemes are evaluated against both the two annotators' selection and their scores in the table are averaged across the two. The precision and recall score for best sentences are determined by evaluating a scheme's selection of the best sentences against a gold standard's selection. To evaluate second-best sentences, precision is measured as the fraction of articles where the test and gold standard selections overlap (share at least one sentence), compared to the total number of articles that have at least one sentence selected according to the test set. Recall is defined by instead dividing by the number of articles that have at least one sentence selected in the gold standard.

	Best		Second best	
Scheme	Precision	Recall	Precision	Recall
Human	0.50	0.55	0.85	0.75
First	0.14	0.20	0.33	0.40
Recent	0.31	0.44	0.51	0.60
Self	0.31	0.36	0.49	0.48
Self fallback	**0.33**	**0.46**	**0.52**	**0.62**

Table 2: Textualization: evaluation results of sentence selection schemes. Self fallback scheme first tries to select the best sentence as the Self scheme, and if it fails to select one it falls back to the Recent scheme.

The low inter-annotator agreement for the best sentence selection shows the difficulty of the problem. However, when a sentence selection scheme is evaluated against the two gold standards, its resulting scores are not that different. It seems that there are a set of articles in which it is easy to pick the best sentence that two annotators and automatic selection schemes easily agree on, and the other set of articles in which it is difficult to find such a sentence. In the *easier* articles, the best sentence often includes a recent date expression, which is easily picked up by the Recent scheme. Figure 7 illustrates such cases. In the more difficult articles, there are no such sentences with recent dates. X2 (film) is such an example; it was released in 2003. The release of the prequel X-Men Origins: Wolverine in 2009 renewed its popularity and the X2 (film) article still does not have any recent dates. There is a more subtle case: the article Farrah Fawcett includes many sentences with recent dates in a section, among which it is hard to pinpoint the best one.

Sentence selection heavily depends on other NLP components, so errors in them could result in the error in sentence selection. Serena Williams is an example that the

error in sentence splitting propagates to sentence selection. The best sentence manually selected was the first sentence in the article "Serena Jameka Williams . . . , as of February 2, 2009, is ranked World No. 1 by the Women's Tennis Association . . . ." The sentence was disastrously divided into two sentences right after "No." by NLTK during preprocessing. In other words, the gold standard sentence could not be selected no matter how well selection performs.[8] Another source of error propagation is co-reference resolution. The Self scheme limits sentence selection to the sentences with a reference to the articles' topic, and it failed to improve over Recent. In qualitative analysis, 3 out of 4 cases that made a worse choice resulted from failing to recognize a reference to the topic of the article. By having it fall back to Recent's selection when it failed to find any best sentence, its performance marginally improved. Improvements of the components would result in better performance of sentence selection.

WikiTopics's current sentence selection succeeded in generating the best or second best sentence that summarizes the relevant current event for more than half of the articles, in enhanced readability through co-reference resolution. For the other difficult cases, it needs to take different strategies rather than looking for the most recent date expressions. Sentence compression, fusion or paraphrasing may be helpful to make a short summary of a current event from possibly multiple and possibly long sentences.

5 Related work

WikiTopics's pipeline architecture has a resemblance to that of news summarization systems such as Columbia Newsblaster (McKeown et al., 2002). Newsblaster's pipeline is comprised of components for performing web crawls, article text extraction, clustering, classification, summarization, and web page generation. The system processes a constant stream of newswire documents. In contrast, WikiTopics analyzes a static set of articles. Hierarchical clustering like three-level clustering of Newsblaster (Hatzivassiloglou et al., 2000) could be applied to WikiTopics to organize current events hierarchically. Summarizing multiple sentences that are extracted from the articles in the same cluster would provide a comprehensive description about the current event. Integer linear programming-based models (Woodsend and Lapata, 2010; Woodsend et al., 2010) have proven to be useful to generate summaries while global constraints like length, grammar, and coverage are met.

The problem of Topic Detection and Tracking (TDT) is to identify and follow new events in newswire, and to detect the first story about a new event (Allan et al., 1998). Allan et al. (2000) evaluated a variety of clustering schemes in the vector space model, where the best settings from those experiments were then used in our work. This was followed recently by Petrović et al. (2010), who took an approximate approach to first story detection, as applied to Twitter in an on-line streaming setting. Such a system might provide additional information to the WikiTopics pipeline by helping to identify and describe current events that have yet to be explicitly described in a Wikipedia article, but where page view logs indicate something new is of interest.

Wikipedia inter-article links have been utilized to construct a topic ontology (Syed et al., 2008), word segmentation corpora (Gabay et al., 2008), or to compute semantic relatedness (Milne and Witten, 2008). In our work, we found the link structure to be as useful to cluster topically related articles as topic models using article text.

6 Conclusions

We have described a pipeline for article selection, clustering, and sentence selection in order to identify and describe significant current events as according to Wikipedia content, and metadata. Similarly to Wikipedia editors maintaining that site's "current events" pages, we are concerned with neatly collecting articles of daily relevance, only automatically, and more in line with expressed user interest (through the use of regularly updated page view logs). We have suggested against the use of Wikipedia's hand-curated articles as an appropriate objective to predict based on pageviews. Clustering methods based on topic models and inter-article link structure are shown to be useful to group a set of articles that are coherently related to a current event. Clustering based on only link structure achieved comparable performance with clustering based on topic models. In a third of cases, the sentence that best described a current event could be extracted from the article text based on temporal expressions within an article. We employed a co-reference resolution system assist in text generation, for improved readability. As future work, ensemble clustering and text classification could be used to improve clustering performance. Sentence compression, fusion, and paraphrasing could be applied to selected sentences with various strategies to more succinctly summarize the current events. Finally, we plan to leverage social media such as Twitter as an additional signal, especially in cases where essential descriptive information has yet to be added to a Wikipedia article of interest.

[8] In a small followup experiment, we found splitta (Gillick, 2009) to give more accurate segmentations, such as not reproducing this particular error.

References

James Allan, Jaime Carbonell, George Doddington, Jonathan Yamron, and Yiming Yang. 1998. Topic Detection and

Tracking Pilot Study Final Report. In *Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop*.

James Allan, Victor Lavrenko, Daniella Malin, and Russell Swan. 2000. Detections, bounds, and timelines: UMass and TDT-3. In *Proceedings of Topic Detection and Tracking Workshop*.

Enrique Amigó, Julio Gonzalo, Javier Artiles, and Felisa Verdejo. 2009. A comparison of extrinsic clustering evaluation metrics based on formal constraints. *Inf. Retr.*, 12(4):461–486.

D.M. Blei, A.Y. Ng, and M.I. Jordan. 2003. Latent Dirichlet allocation. *Journal of Machine Learning Research*.

Elizabeth Boschee, Ralph Weischedel, and Alex Zamanian. 2005. Automatic information extraction. In *Proceedings of IA*.

David Gabay, Ziv Ben-Eliahu, and Michael Elhadad. 2008. Using wikipedia links to construct word segmentation corpora. In *Proceedings of AAAI Workshops*.

Dan Gillick. 2009. Sentence boundary detection and the problem with the U.S. In *Proceedings of HLT/NAACL*.

Vasileios Hatzivassiloglou, Luis Gravano, and Ankineedu Maganti. 2000. An investigation of linguistic features and clustering algorithms for topical document clustering. In *Proceedings of SIGIR*.

Edward Loper and Steven Bird. 2002. NLTK: the Natural Language Toolkit. In *Proceedings of ACL*.

C. Manning, P. Raghavan, and H. Schütze. 2008. *Introduction to information retrieval*. Cambridge University Press.

Andrew Kachites McCallum. 2002. MALLET: A Machine Learning for Language Toolkit. http://mallet.cs.umass.edu.

Kathleen R. McKeown, Regina Barzilay, David Evans, Vasileios Hatzivassiloglou, Judith L. Klavans, Ani Nenkova, Carl Sable, Barry Schiffman, and Sergey Sigelman. 2002. Tracking and summarizing news on a daily basis with Columbia's Newsblaster. In *Proceedings of HLT*.

David Milne and Ian H. Witten. 2008. An effective, low-cost measure of semantic relatedness obtained from Wikipedia links. In *Proceedings of AAAI Workshops*.

Saša Petrović, Miles Osborne, and Victor Lavrenko. 2010. Streaming first story dectection with application to Twitter. In *Proceedings of NAACL*.

Zareen Saba Syed, Tim Finin, and Anupam Joshi. 2008. Wikipedia as an ontology for describing documents. In *Proceedings of ICWSM*.

Kristian Woodsend and Mirella Lapata. 2010. Automatic generation of story highlights. In *Proceedings of ACL*.

Kristian Woodsend, Yansong Feng, and Mirella Lapata. 2010. Title generation with quasi-synchronous grammar. In *Proceedings of EMNLP*.