Headline Generation using Bayesian Phrase-Relation Topic Models

Karthik Anantha Padmanabhan
Department of Computer Science
The University of Texas at Austin
akarthik@cs.utexas.edu

Vineet Keshari
Department of Computer Science
The University of Texas at Austin
vkeshari@cs.utexas.edu

Abstract

In this paper, we present an unsupervised approach to abstractive headline generation for news articles. By extracting phrases and relations from the sentences in a document, and training topic models on the constituents so obtained, we identify a set of most likely phrases that can be part of the summary. We then combine the phrases to generate a syntactically coherent summary in the form of a single clause. We evaluate our model on the Reuters dataset and show that it performs better than a bag of words approach, based on the ROUGE evaluation metric.

1 Introduction

The last two decades have witnessed a lot of interest in Automatic Document Summarization. The large amount of text available freely on the internet has increased the feasibility of retrieving and processing documents to develop meaningful insights into their content. At the same time, due to an abundance of information on any given topic, largely overlapping and sometimes contradictory, there is an increasing need to summarize comprehensive documents from multiple sources into concise and meaningful summaries. One of the most popular examples of the latter is the aggregation of news articles by web-based agents such as *Google News*[1]. Multi-document summarization can utilize large amounts of data to train statistical models, which can then be used to generate summaries based on maximum likelihood estimates. This statistical approach to summarization has gained more popularity in the last decade, and has been used in a variety of ways by teams competing in the annual Document Understanding Conference (DUC).

There are two approaches to text summarization: Extractive and Abstractive. In extractive summarization, the sentences in the summary are picked from the document based on their relevance, calculated using some measure, to the overall content of the document. The approach is therefore suitable for generating summaries that consist of multiple sentences. Erkan & Radev in [2] and Wang et. al. in [3] have successfully used supervised and unsupervised techniques respectively to generate multi-sentence summaries from documents with reasonable accuracy. However, applications such as news headline generation require single-sentence summaries. The extractive approach fails to produce good results for this task as the meaning of a document is rarely contained in a single sentence in the text. This makes single-sentence summarization a challenging, and arguably harder, problem. In abstractive summarization, the sentences in the summary are generated, instead of picked, based on an understanding of both the document and the language. This makes the approach suitable for generating single-sentence summaries as well. In [4], Woodsend et. al. have used a quasi-synchronous grammar to generate abstractive summaries. Other techniques include use of singular value decomposition (Wan et. al. [5]) and hidden markov models (Zajic & Dorr [6]).

In this paper, we present an unsupervised approach to abstractive headline generation for news articles using topic models [7]. While topic models have been used before for summarization, the novelty in our approach lies in extracting phrases and relations from sentences during content selection and using them later as the constituents of the generated summary during surface realization. The paper is organized as follows: In Section 2, we discuss the theoretical foundations of topic models and introduce our algorithm, along with the various models we have implemented. Section 3 describes the experimental evaluation and results, followed by discussion. In Section 4 we mention related work in the field which includes two papers that act as

[1] http://news.google.com

baselines for measuring the performance of our approach. Finally, we identify future work and conclude our discussion in Sections 5 and 6 respectively.

2 Problem Definition and Algorithm

2.1 Problem Definition

The problem in this experiment is to generate single-sentence headlines for news articles. A headline is a concise summary of the contents of the article, usually composed of only one clause. Given a dataset containing several news articles, the task is to generate a headline for each article which best describes its contents in a single clause. The headlines obtained are evaluated against the human-generated headlines of the articles.

2.2 Algorithm Definition

2.2.1 Topic Modeling

Topic Models are a suite of statistical techniques that analyze words in text to annotate large document corpora with thematic information. A topic can formally be defined as a probability distribution over a fixed vocabulary. The most commonly used topic model is the Latent Dirchlet Allocation (LDA). It is based on the assumption that each document exhibits multiple topics in different proportions. Each document can be generated by the following generative process: First choose a distribution over topics, then for every word position choose a topic according to this distribution. Once the topic is chosen a word is emitted from this topic according to its probability distribution in the topic. The central computational problem in LDA is to reverse this generative process and infer the topic structure from the observed documents.

LDA can be more formally defined with the following notation. The topics are $T_{1:k}$ where each T_t is a distribution over the vocabulary. The topic proportions for the d^{th} document are θ_d, where θ_d^t is the topic proportion for topic t in document d. The topic assignments for the d^{th} document are z_d. Finally the observed words for document d are $w_{d,n}$ for the n^{th} word in the document. These dependencies define the LDA. As mentioned earlier the central computational problem in LDA is to compute the conditional distribution of the topic structure given the observed words in a document. The posterior distribution can thus be defined as

$$p(T_t, \theta_d^t, z_d \big| w_{1:l}) = \frac{p(T_t, \theta_d^t, z_d, w_{1:l})}{p(w_{1:l})} \quad (2.1)$$

Exact inference of the posterior distribution is not tractable, since the numerator of (2.1) has an exponential number of possibilities. Instead, we can compute the approximate posterior distribution using a class of approximate inference techniques called Sampling based methods. The most commonly used sampler for LDA is the Gibbs Sampler.

In this project instead of viewing topics as a probability distribution over a fixed vocabulary of words, we consider a topic to be a distribution over phrase constituents that are extracted from documents. We now describe Phrase-Relation Topic models.

2.2.2 Phrase Relation Topic Models

We define phrase relation topic models as a set of models trained using the LDA algorithm on phrases in a sentence. A phrase is a part of a sentence that describes an entity or a relation. Noun phrases (NPs) usually describe a person, object, place or idea, whereas verb phrases (VPs) describe the relation between its surrounding noun phrases. A clause is the smallest unit in a sentence that defines a set of entities and their relations. A clause (c_j) may therefore consist of a subject (s_j), a relation (r_j) and an object (o_j), where the subject and object are NPs and the relation is a VP. Other phrases such as prepositional phrases (PPs) may also appear in a clause, but for headlines we focus on concise clauses containing only one phrase each of subject, relation and object, in that order.

A document d in our dataset is defined as a sequence of sentences, such that $d = S_1 S_2 S_3 \,..\, S_n$. We begin our algorithm by splitting each sentence S_i into clauses c_j, followed by identifying the subject s_j, relation r_j and object phrases o_j in each clause.

$S_i = c_1 c_2 c_3 \,..\, c_m,\ i \in \{1\ \dots\ n\}$, and

$c_j = s_j r_j o_j,\ j \in \{1\ \dots\ m\}$, where each of s_j, r_j and o_j consist of words w_l defined as follows:

$s_j = w_1 w_2 w_3 \,..\, w_l,\ l = |s_j|$

$r_j = w_1 w_2 w_3 \,..\, w_l,\ l = |r_j|$

$o_j = w_1 w_2 w_3 \,..\, w_l,\ l = |o_j|$

We then apply multiple techniques to group the subjects, relations and objects together to form the bags of words for training the topic models. For example, in the simplest approach defined in Section 2.3.2, we group the words w_l in each phrase into the same bag. We remove stop words from the bags before we train the topic models. The trained topic models give us parameters θ_d^t which represent the probability of assigning topic t to document d:

$$\theta_d^t = P(T = t|D = d), \text{ where } \sum_t \theta_d^t = 1 \quad (2.2)$$

We define the probability of a clause given the topic t as follows:

$$P(c_j|T = t) = P(s_j|T = t) * P(r_j|T = t) * P(o_j|T = t), \quad (2.3)$$

where the probability of each phrase is defined by the words in the phrase:

$$P(s_j \mid T = t) = \prod P(w_l \mid T = t), w_l \in s_j$$
$$P(r_j \mid T = t) = \prod P(w_l \mid T = t), w_l \in r_j$$
$$P(o_j \mid T = t) = \prod P(w_l \mid T = t), w_l \in o_j,$$

and the probabilities $P(w_l|T = t)$ are obtained from the topic models.

The inference task, then, is to pick the clause with the highest probability c^* as the headline $H(d)$ given the document d where:

$$c^* = \underset{c_j}{\operatorname{argmax}} \sum_t P(c_j|T = t) * \theta_d^t \quad (2.4)$$

We now discuss other approaches to defining the bags, training the topic models and drawing our inference in the next subsection. The underlying idea in each of them is the same as described above.

2.3 Models

In this subsection, we define the various techniques we implemented for defining the inputs to train topic models on, as well as the methods of inference, if different.

2.3.1 Baseline Model

In this model, we implement the technique described by Wang et. al. in [3] to train topic models on sentences. Here, the sentences $S_i = w_1 w_2 .. w_l$ form the atomic units for inference instead of phrases. The inference problem requires determining S^* given the document d as the headline of the document. We note that this technique was defined by Wang et. al. for multi-sentence summarization, where the top k sentences are picked to form a summary. However, since our approach involves breaking the sentences down into phrases, and uses the same inference technique thereafter, we will use it as a baseline for single-sentence summarization.

2.3.2 Simple SVO Model

This model is the same as defined in Section 2.2. The difference from the Baseline model is that here we identify the Subject-Verb-Object clauses (SVOs) from the sentences and remove all other words. We then train the models on the SVO clauses instead of entire sentences and select the clause with the highest probability for the given document using equations 2.3 and 2.4

2.3.3 SVO Phrase Model

In this model, we group the subjects, relations and objects into different bags, and train one set of topic models for each. We then infer s^*, r^* and o^* independently and combine them to form the headline:

$$H(d) = s^* r^* o^*, \text{ where}$$
$$s^* = \underset{s_j}{\operatorname{argmax}} \sum_t P(s_j|T = t) * \theta_d^t$$
$$r^* = \underset{r_j}{\operatorname{argmax}} \sum_t P(r_j|T = t) * \theta_d^t$$
$$o^* = \underset{o_j}{\operatorname{argmax}} \sum_t P(o_j|T = t) * \theta_d^t$$

Figure 2.1 shows the setup for this model.

2.3.4 Condensed Phrase Model

The phrases we use in the SVO Phrase Model may contain many words such as descriptive adjectives, adverbs and determiners which we do not wish to train the topic models on. We therefore remove these words in this model to form a "condensed" phrase containing only the nouns in the subject and object phrases and only the verbs in the relation phrases.

$$s'_j = w_l, s.t. \; w_l \in s_j \wedge POSTag(w_l) \in \{\mathrm{NN, NNP, NNS, NNPS}\}$$
$$r'_j = w_l, s.t. \; w_l \in r_j \wedge POSTag(w_l) \in \{\mathrm{VB, VBD, VBG, VBN, VBP, VBZ}\}$$
$$o'_j = w_l, s.t. \; w_l \in o_j \wedge POSTag(w_l) \in \{\mathrm{NN, NNP, NNS, NNPS}\}$$

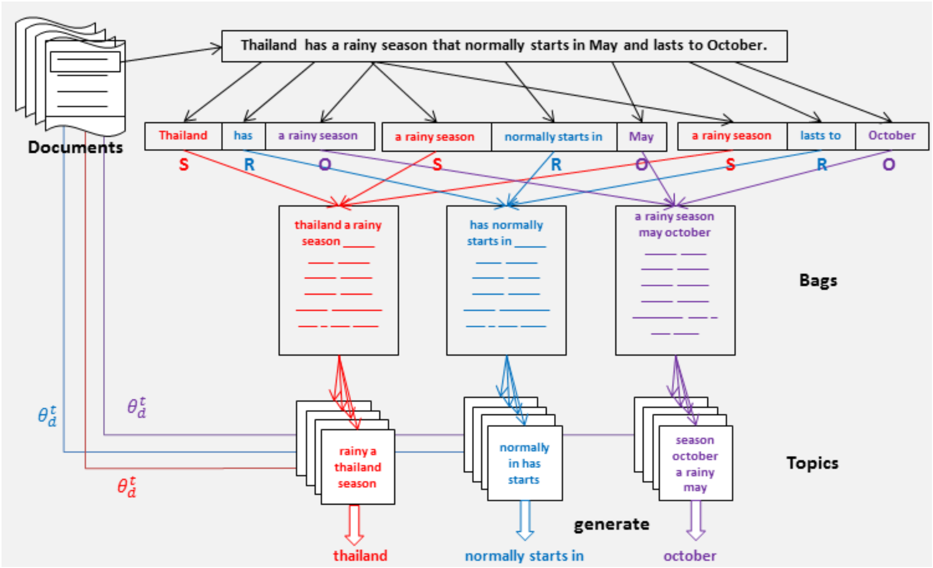

Figure 2.1: Processing a sentence with the SVO Phrase model defined in Section 2.3.3. This particular sentence from the Reuters dataset is broken into three clauses, with the subject-relation-object identified as shown above. A sample headline that may be generated for the document is also shown.

We then train the topic models similar to that in the SVO Phrase Model but with s_j', r_j' and o_j' as inputs instead. However, when generating the headline we use the original phrases corresponding to the condensed phrases chosen by the model. This is because we want to generate a grammatically coherent headline, a property which is lost when forming the condensed phrases.

2.3.5 Trimmed Phrase Model

We follow an approach similar to the condensed phrase model here, but instead of choosing only the words that are nouns or verbs, we choose the words in the smallest subphrase that contains all the nouns or verbs in the phrase, as applicable. For example, for subjects we define:

$$s_j' = w_l, s.t. w_l \in s_j \wedge POSTag(w_z) \notin \{NN, NNP, NNS, NNPS\} \; \forall z \notin range(l).$$

2.3.6 Concatenated Phrase Model

In this model, instead of defining the phrases in the Condensed Phrase Model as a bag of words, we concatenate them to form one word that represents the nouns or verbs in the phrase in order. For example, the sentence phrases are now defined as follows:

$$s_j' = w_1\_w_2\_ \ldots w_l$$

The other steps are similar to that in the Condensed Phrase Model.

2.3.7 Subject-Object Model

Since the subject and object phrases in the Concatenated Phrase Model both consist of only nouns, and noting that the subject for the headline of one document may be the object for another, we combine the words in the subject and object phrases to form one set of nouns:

$$so_j' = w_s \cup w_o, \quad s.t.\ w_s \in s_j, w_o \in o_j$$

We then train only two models, one for nouns and one for verbs and choose the subject, verb and object phrases from the appropriate models as before.

2.3.8 Subject Priority Model

In this model, we first choose the subject phrase from its topic model. We then reduce the list of verb and object phrases to correspond only to the top k subjects, and choose the most likely ones.

$$r^* = \underset{r_{j'}}{\text{argmax}} \sum_t P(r_{j'}|T = t) * \theta_d^t, and$$

$$o^* = \underset{o_{j'}}{\text{argmax}} \sum_t P(o_{j'}|T = t) * \theta_d^t, where$$

$$j' \in \arg max_j^k \sum_t P(s_j|T = t) * \theta_d^t$$

The idea here is that news articles can mostly be defined for a subject, and the relation and object are associated with that subject.

2.3.9 Subject-Object Priority Model

In addition to the top k subjects identified in the Subject Priority Model, we determine the top k objects in the document as well. We then take the intersection of the clauses which contain these phrases and choose our relation from the reduced set of relations so obtained.

$$r^* = \underset{r_{j'}}{\text{argmax}} \sum_t P(r_{j'} | T = t) * \theta_d^t, where$$

$$j' \in \left\{ \arg\max_j^k \sum_t P(s_j | T = t) * \theta_d^t \right\} \cap \left\{ \arg\max_j^k \sum_t P(o_j | T = t) * \theta_d^t \right\}$$

If the intersection is the null set, we fall back to the Subject Priority Trainer.

3 Experimental Evaluation

3.1 Methodology

3.1.1 Dataset

We use news articles from the Reuters 21578 corpus to evaluate our model. The Reuters dataset consists of news articles and their corresponding headlines. We chose this dataset among other news article datasets as this was easily accessible, provided us with a large collection of documents and was similar to the dataset used in one of the baselines.

3.1.2 Implementation

The models are implemented in Java. For each news article, we detect sentences using the OpenNLP Sentence Detection tool. Each sentence is then chunked and passed to the ReVerb [8] Relation Extraction tool, which returns a set of binary relations for every sentence. We thus obtain a set of s, r, o triplets for every document.

These triplets form the input for the various Phrase Relation Topic Models described in Section 2.3. We use Mallet [10] to train the topic models, after preprocessing the text using the appropriate pipes for stop-word removal and conversion to lower case. We learn 100 topics for each input type, which is an approximation to the number of categories that the news articles belong to (the dataset defines 135 categories, but the articles are not uniformly distributed among them).

The generated headlines along with the true headlines are then wrapped in HTML documents so that they adhere to the input format of the ROUGE package. ROUGE then evaluates the results and returns the Precision, Recall and F1 scores based on various methods of evaluation.

3.1.3 Evaluation

To evaluate our model we use ROUGE [9] (Recall Oriented Understudy for Gisting Evaluation), which has been used in the Document Understanding Conference (DUC) 2004, a large-scale summarization evaluation sponsored by NIST. It includes several automatic evaluation metrics to compute the similarity between human-generated and system-generated summaries. The similarity between summaries is in terms of word overlap. For the generated summaries we compute ROUGE-1, ROUGE-L and ROUGE-SU scores after stemming the words and discarding stop words. In [9] Lin et. al show that ROUGE-1, ROUGE-L and ROUGE-SU have high correlation with human evaluation for short summaries. Hence in our evaluations we opt for these three metrics.

Method	**ROUGE-1**			**ROUGE-L**			**ROUGE-SU**		
	P	R	F1	P	R	F1	P	R	F1
Baseline Model	**0.39**	0.12	0.18	**0.35**	0.11	0.16	**0.24**	0.03	0.05
Simple SVO	0.21	0.22	0.20	0.20	0.21	0.15	0.09	0.10	0.08
SVO Phrase	0.28	0.20	0.23	0.27	0.20	0.22	0.14	0.08	0.09
Condensed Phrase	0.28	0.22	0.24	0.27	0.21	0.22	0.14	0.09	0.10
Trimmed Phrase	0.28	0.21	0.23	0.27	0.20	0.22	0.14	0.08	0.10
Concatenated Phrase	0.28	0.32	0.28	0.26	0.32	0.27	0.14	0.19	0.13
Subject-Object	0.31	**0.38**	**0.31**	0.30	**0.37**	**0.30**	0.16	**0.23**	**0.15**
Subject Priority	0.29	0.34	0.29	0.27	0.33	0.28	0.14	0.20	0.13
Subject-Object Priority	0.29	0.34	0.29	0.28	0.33	0.28	0.15	0.20	0.14

Table 3.1: Results for all models on Reuters dataset (P – Precision, R – Recall, F1 – F1Score)

No.	Original Headline	Generated Headline
H1	INDIA FOODGRAIN TARGET 160 MLN TONNES IN 1987/88	INDIA 'S NATIONAL FOODGRAIN TARGET HAS BEEN FIXED AT 160 MLN TONNES
H2	JAPAN TO RELEASE GNP FIGURES LATER TODAY	THE ECONOMIC PLANNING AGENCY WILL ANNOUNCE GROSS NATIONAL PRODUCT (GNP) FIGURES
H3	TAIWAN CABINET APPROVES REC-ORD DEFICIT BUDGET	THE TAIWAN CABINET APPROVED A RECORD DEFICIT BUDGET
H4	WORLD BANK TEAMS STUDYING PHILIPPINE LOAN REQUESTS	TWO WORLD BANK REVIEW TEAMS ARE IN THE PHILIPPINES
H5	PAKISTAN SAYS AFGHAN AIR RAID KILLED ALMOST 150	AFGHAN AIRCRAFT WERE KILLED IN PAKISTANI VILLAGES
H6	DAI NIPPON TO ISSUE THREE DOMESTIC CONVERTIBLES	IT SAID IT

Table 3.2: Original and generated headlines for six stories from the Reuters dataset

3.2 Results

Table 3.1 shows the precision, recall and F1 scores for the ROUGE-1, ROUGE-L and ROUGE-SU measures on all models described in Section 2.3. The models have been trained on the entire Reuters dataset.

We can see that the Subject-Object model gives us the best results. We also note that while the precision for our models isn't as good as the baseline, the recall is significantly higher resulting in better F1 scores.

Table 3.2 contains examples of original and generated headlines for six documents, including correct and incorrect headline generation.

3.3 Discussion

Of the various models we train, the subject-object model outperforms the others on all ROUGE metrics. Its good results may be attributed to the fact that we group and concatenate all nouns together, which allows us to train only two topic models, while the subjects and objects are still generated independently.

It should be noted that our approach gives significantly better recall scores than a bag of words approach, indicating that we have much lesser false negatives. Therefore our results have a much higher measure of agreement with human-generated summaries. This can be attributed to our use of phrases instead of sentences for training the topic models.

Based on the results shown in Table 3.1, we can verify our claim that the phrase-relation based approach performs better than Wang et. al's sentence-based approach described in Section 4.1. The improved F1 scores show that the headlines generated are both more accurate and more consistent with the original headlines when compared to the baseline.

Our four best models also outperform Wan et. al's LSI based model described in Section 4.2. Their best unigram (same as ROUGE-1) recall score was 0.26 for a sentence length of 13.

We also note that all models that use concatenated words (models 6-9) perform better than those that use a bag of words model (models 1-5). We therefore also verify our claim that the bag of words approach is not suitable for headline generation.

Finally, looking at some interesting cases listed in Table 3.2, we can determine the strengths and weaknesses of our algorithm. In H2, for example, the algorithm generates an accurate headline but fails to provide context (in this case, the name of the country). H3 is in the past tense instead of present, which is expected since it is generated from the article without knowledge of the existing headline. H4, similar to H2, generates an accurate but semantically incoherent headline. H5 only gets the subject wrong, but we see that one wrong component can significantly impact the information contained in the headline. Lastly, H6 gets everything wrong, which is always a possibility in a probabilistic model.

4 Related Work

4.1 Multi-document summarization using Bayesian Sentence-based topic models

In [3], Wang et al. use an extractive approach to train topic models on the sentences in a document using Bayesian Sentence-Based Topic

Models (BSTMs). In this generative approach, each sentence is considered as a bag of words. The summarization task, therefore, is to determine the probability that sentence s describes the topic t, for the k most probable latent topics assigned to document d. Based on ROUGE evaluation metrics, their method performs as good as most of the teams in the DUC competitions on the same dataset.

We note that since their method selects entire sentences to form the summary, it is unlikely that it will be able to summarize the content of a document in a single sentence. At the same time, the sentences in the documents may be arbitrarily long, making the method unsuitable for generating a concise headline. We have attempted to address these issues in our approach by training topic models on phrases instead of sentences, enabling us to generate a summary as small as one SVO clause, as well as to incorporate information from the entire document into a single sentence.

4.2 Using thematic information in statistical headline generation

In [5], Wan et al use Latent Semantic Indexing (LSI) to guide the selection of words during content selection. LSI is a dimensionality reduction technique that projects documents into lower dimensional spaces. Documents with similar topical content are close to one another in the resulting space based on word co-occurrence patterns. The authors refer to documents in the reduced space as “themes”. They identify the themes associated with sentences, assuming that the highest scoring theme reflects the topical content of the entire document. The words selected for the headline are the terms from the highest scoring “theme”. The surface realization criterion is determined by the use of bigram statistics. The headline generation problem is then reduced to finding the set of words that maximize the content selection as well as the surface realization scores. They use unigram overlap (ROUGE-1) to evaluate the generated headlines.

One of the disadvantages that the authors report is the narrowness of the highest scoring theme which did not relate to the overall content of the document. Also, the surface realization in this method is not guaranteed to be syntactically correct as it considers only bigram probabilities. We have attempted to address this issue by using a Subject-Verb-Object structure for the headline.

5 Future Work

A major limitation of our approach is that it does not generalize to all languages, since we assume that the headline will have a subject-relation-object structure. While this assumption works well for English, it might not give good results for languages that follow a different grammar (e.g. Hindi). Future work can focus on making the model more general by defining a universal set of phrases and a method to combine them, or a technique to learn these from the data.

Another drawback of our approach is that completely unrelated phrases from the document can sometimes be combined to form a headline which, while syntactically correct, fails to be semantically comprehensible. We have attempted to address this issue in the model defined in Section 2.3.8, but more advanced techniques may be implemented to overcome this problem.

Lastly, while our results are better than other unsupervised techniques, it is debatable whether unsupervised methods are useful for headline generation, since a news article is always accompanied by an accurate headline written by its author. While this fact tends to undermine the importance of our work for news headline generation, we hope that the work presented in this paper will encourage further research into Phrase-Relation Topic Models.

6 Conclusion

In this paper, we have shown that better results can be obtained from a topic model based summarization technique by grouping or concatenating similar words instead of using them as a bag of words. We have also shown that it is possible to generate syntactically coherent summaries by extracting phrases and relations from a document’s sentences and treating them separately.

While our approach solves these two problems, it is limited to one class of languages. It also suffers from a high probability of generating semantically incoherent summaries by linking irrelevant or incorrect phrases together.

Further work on Bayesian phrase-relation topic models can help improve summarization results for both unsupervised and supervised models.

References

[1] Das, Dipanjan, and André FT Martins. "A survey on automatic text summarization." *Literature Survey for the Language and Statistics II course at CMU* 4 (2007): 192-195.

[2] Erkan, Günes, and Dragomir R. Radev. "LexRank: Graph-based lexical centrality as salience in text summarization." *J. Artif. Intell. Res. (JAIR)* 22 (2004): 457-479.

[3] Wang, Dingding, et al. "Multi-document summarization using sentence-based topic models." *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*. Association for Computational Linguistics, 2009.

[4] Woodsend, Kristian, Yansong Feng, and Mirella Lapata. "Generation with quasi-synchronous grammar." *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2010.

[5] Wan, Stephen, et al. "Using thematic information in statistical headline generation." *Proceedings of the ACL 2003 workshop on Multilingual summarization and question answering-Volume 12*. Association for Computational Linguistics, 2003.

[6] Zajic, David, Bonnie Dorr, and Richard Schwartz. "Automatic headline generation for newspaper stories." *Workshop on Automatic Summarization*. 2002.

[7] Blei, David M., Andrew Y. Ng, and Michael I. Jordan. "Latent dirichlet allocation." *the Journal of machine Learning research* 3 (2003): 993-1022.

[8] Fader, Anthony, Stephen Soderland, and Oren Etzioni. "Identifying relations for open information extraction." *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2011.

[9] Lin, Chin-Yew. "Rouge: A package for automatic evaluation of summaries." *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*. 2004.

[10] McCallum, Andrew Kachites. "MALLET: A Machine Learning for Language Toolkit." http://mallet.cs.umass.edu. 2002.

[11] Srivastava, Ashok, and Mehran Sahami, eds. *Text mining: Classification, clustering, and applications*. Chapman and Hall/CRC, 2010.