Using Signals of Human Interest to Enhance Single-document Summarization

Krysta M. Svore
Microsoft Research
1 Microsoft Way
Redmond, WA 98052
ksvore@microsoft.com

Lucy Vanderwende
Microsoft Research
1 Microsoft Way
Redmond, WA 98052

Christopher J.C. Burges
Microsoft Research
1 Microsoft Way
Redmond, WA 98052

Abstract

As the amount of information on the Web grows, the ability to retrieve relevant information quickly and easily is necessary. The combination of ample news sources on the Web, little time to browse news, and smaller mobile devices motivates the development of automatic highlight extraction from single news articles. Our system, NetSum, is the first system to produce highlights of an article *and* significantly outperform the baseline. Our approach uses novel information sources to exploit human interest for highlight extraction. In this paper, we briefly describe the novelties of NetSum, originally presented at EMNLP 2007, and embed our work in the AI context.

Introduction

With the ever-growing internet and increased information access, we believe single-document summarization is essential to improve quick access to large quantities of information. In particular, there are ample online news sources. News search engines allow querying an index of worldwide news articles, and in addition sometimes offer lists of top stories and related articles. However, there are no existing news services online that offer *automatic* creation of summaries or highlights of either a single news article or a cluster of related news articles. Highlights are an ideal solution for readers who want just an overview of the key points of an article, a summary of events on a small mobile device where screensize limits the amount of displayed content, or the ability to accumulate information about many topics quickly. Select online news sites have human-generated highlights for certain articles. Recently, CNN.com (CNN.com 2007a) added "Story Highlights" to many news articles on its site. These highlights briefly overview the article and appear as 3–4 bullet points rather than a summary paragraph, making them even easier to quickly scan.

Our work is motivated by both the addition of highlights to an extremely visible and reputable online news source, as well as the inability of past single-document summarization systems to outperform the extremely strong baseline of choosing the first n sentences of a newswire article as the summary (Nenkova 2005). Although some recent systems indicate an improvement over the baseline (Mihalcea 2005; Mihalcea & Tarau 2005), statistical significance has not been shown. We use a neural network ranking algorithm and exploit third-party datasets based on human interest to outperform the baseline with statistical significance. In this paper, we review the performance of our system, *NetSum*, on the task of matching (the content of) a set of human-generated highlights.

Highlight Extraction in the AI Context

Our goal is to extract three sentences from a single news document that best match the content and characteristics of human-generated highlights. Our task is related to several fields of AI.

Identifying content within a larger information source is related to the science of *information retrieval*. We must search for the important information in the document and retrieve it, by taking advantage of features and metadata of each sentence in the document. Although we employ our technology for single-document summarization, our system can also be adapted to retrieve information from, say, a cluster of news articles or even a cluster of general documents.

Most obviously, our work is related to the general field of *text summarization*. Automatic summarization was first studied almost 50 years ago by Luhn (Luhn 1958) and has continued to be a steady subject of research. Automatic summarization is the creation of a shortened version of a document by a machine (Mani 2001). Classic text summarization has been developed in the areas of abstraction and extraction. When creating an abstract summary, content can be drawn from both multiple sentences and information outside of the document(s) to generate new sentences used in the summary. An extract summary preserves the article content in its original form, i.e., sentences, and is the focus of our work. In addition to single-document summarization, summarization of multiple articles is also a focus of research. Summarization techniques have been applied to documents other than news, such as law journals, medical journals, books, and so on.

In 2001–02, the Document Understanding Conference (DUC 2001), issued the task of creating a 100-word summary of a single news article. The best performing systems (Hirao *et al.* 2002; Lal & Ruger 2002) used various learning and semantic-based methods, although no system could outperform the baseline with statistical significance (Nenkova 2005). After 2002, the single-document summarization task was dropped. NetSum is the first system to beat the baseline with statistical significance.

Producing highlights for a news article is also related to the larger field of *knowledge extraction*. In our work, we rank article sentences in order of human interest and importance to the article and topic. By extracting relevant sentences, we inherently extract knowledge from the article that is of highest interest to the reader.

More importantly, our work is closely related to *natural language understanding* (NLU). Typically, in NLU, the emphasis is to derive the understanding from the text itself. If we consider that the task of generating summaries demonstrates the degree of understanding of the text, then improving on this task should correlate with increased understanding of the text. This was also argued for in (Vanderwende 2007), where the task of generating appropriate questions for a given text requires greater understanding than only the text itself.

In our work, we show that using sources of knowledge external to the text itself, that is, knowledge about what people are interested in (news search query logs) and knowledge about what people care about (Wikipedia) does improve our ability to generate summaries automatically. We can say, therefore, that understanding is a function of the specifics of the text, but also of what is important about the text and what is memorable about the text. Third-party sources have recently been used regularly to enhance infromation retrieval and extraction. Previously, third-party sources such as WordNet (Fellbaum 1998), the Web (Jagarlamudi, Pingali, & Varma 2006), or click-through data (Sun *et al.* 2005) have been leveraged to enhance performance of many types of information retrieval systems. However, the task of summarization more directly exhibits degrees of understanding, and so we feel it is the most compelling of tasks. Though we currently employ the news search query logs and Wikipedia, these are but a few of the indications people give of their levels of interest and shared knowledge. We encourage the reader to experiment with other sources of human interest to enhance existing summarization and knowledge extraction systems.

The NetSum System

Our system, NetSum, extracts three sentences from a news article that best match, or contain, the content of the human-generated highlights. We do not consider the order of the content. In (Svore, Vanderwende, & Burges 2007), we address how to produce highlights when the ordering of content matters. We call the three highlights a *block*. Throughout our paper, we refer to a human-generated highlight as simply a highlight. We evaluate our system's block against 1) the human-generated highlight block as well as 2) the baseline of creating a block from the first three sentences of the article. We assume the title has been seen by the reader and will be listed above the highlights.

One way to identify the best sentences is to rank the sentences using a machine learning approach, where each sentence is assigned a label indicating its importance and has a set of extracted features. From the labels and features for each sentence, we train a model that, when run on a test set of sentences, can infer the proper ranking of sentences in a document based on information gathered during training about sentence characteristics. To accomplish the ranking, we use RankNet (Burges *et al.* 2005), a ranking algorithm based on neural networks. The system is trained on pairs of sentences (S_i, S_j), such that S_i should be ranked higher or equal to S_j. Pairs are generated between sentences in a single document, not across documents. For details on training the model, see (Svore, Vanderwende, & Burges 2007).

TIMESTAMP: 1:59 p.m. EST, January 31, 2007

TITLE: Nigeria reports first human death from bird flu

HIGHLIGHT 1: Government boosts surveillance after woman dies

HIGHLIGHT 2: Egypt, Djibouti also have reported bird flu in humans

HIGHLIGHT 3: H5N1 bird flu virus has killed 164 worldwide since 2003

ARTICLE: **1. Health officials reported Nigeria's first cases of bird flu in humans on Wednesday, saying one woman had died and a family member had been infected but was responding to treatment. 2. The victim, a 22-year old woman in Lagos, died January 17, Information Minister Frank Nweke said in a statement. 3. He added that the government was boosting surveillance across Africa's most-populous nation after the infections in Lagos, Nigeria's biggest city.** 4. The World Health Organization had no immediate confirmation. 5. Nigerian health officials earlier said 14 human samples were being tested. 6. Nweke made no mention of those cases on Wednesday. 7. An outbreak of H5N1 bird flu hit Nigeria last year, but no human infections had been reported until Wednesday. **8. Until the Nigerian report, Egypt and Djibouti were the only African countries that had confirmed infections among people.** 9. Eleven people have died in Egypt. 10. The bird flu virus remains hard for humans to catch, but health experts fear H5N1 may mutate into a form that could spread easily among humans and possibly kill millions in a flu pandemic. 11. Amid a new H5N1 outbreak reported in recent weeks in Nigeria's north, hundreds of miles from Lagos, health workers have begun a cull of poultry. 12. Bird flu is generally not harmful to humans, but the H5N1 virus has claimed at least 164 lives worldwide since it began ravaging Asian poultry in late 2003, according to the WHO. 13. The H5N1 strain had been confirmed in 15 of Nigeria's 36 states. 14. By September, when the last known case of the virus was found in poultry in a farm near Nigeria's biggest city of Lagos, 915,650 birds had been slaughtered nationwide by government veterinary teams under a plan in which the owners were promised compensation. 15. However, many Nigerian farmers have yet to receive compensation in the north of the country, and health officials fear that chicken deaths may be covered up by owners reluctant to slaughter their animals. **16. Since bird flu cases were first discovered in Nigeria last year, Cameroon, Djibouti, Niger, Ivory Coast, Sudan and Burkina Faso have also reported the H5N1 strain of bird flu in birds.** 17. There are fears that it has spread even further than is known in Africa because monitoring is difficult on a poor continent with weak infrastructure. 18. With sub-Saharan Africa bearing the brunt of the AIDS epidemic, there is concern that millions of people with suppressed immune systems will be particularly vulnerable, especially in rural areas with little access to health facilities. 19. Many people keep chickens for food, even in densely populated urban areas.

Figure 1: Example document containing highlights and article text. Sentences are numbered by position. Bold sentences will be referred to in the paper. Article is from (CNN.com 2007b).

Our train and test data consists of 1365 news documents gathered from CNN.com (CNN.com 2007a). Each document was extracted by hand on consecutive days during February 2007, where a maximum of 50 documents per day were collected. Each document contains a title, timestamp, story highlights, and article text. The timestamp ranges between December 2006 to February 2007. There are 3–4 human-generated story highlights per article.

Sentence Labeling

To train and test our system, we annotate each sentence with a label. Our system selects three sentences that are most apt as highlights. Choosing three sentences most similar to the three highlights is very challenging. In other summarization data sets, it has been observed that 19% of human-generated summary sentences contain no matching article sentence (Jing 2002) and only 42% of summary sentences match the content, but not necessarily the syntax or semantics, of a single article sentence. Since each highlight is human generated and contains content gathered across sentences, and vocabulary not necessarily present in the text, we must identify how closely related a highlight is to a sentence. We use Recall-Oriented Understudy for Gisting Evaluation (ROUGE) (Lin 2004b) to measure the quality of a model-selected sentence against a "gold-standard", human-generated highlight. ROUGE is very effective for measuring both single-document summaries and single-document headlines (Lin 2004a). We label each sentence S_i by l_1, the maximum ROUGE-1 score between S_i and each highlight H_n, for $n = 1, 2, 3$, where ROUGE-1 measures recall over unigrams.

Features

RankNet takes as input a set of samples, where each sample contains a label and feature vector. We generate 10 features for each sentence S_i in each document. Each feature is chosen to identify characteristics of an article sentence that match those of a highlight. Some features such as position and N-gram frequencies are commonly used for scoring. We use variations on these features as well as a novel set of features based on third-party data. For a complete description of features, see (Svore, Vanderwende, & Burges 2007).

Since the first sentence of a news articles typicaly summarizes the article, we include a feature to indicate if S_i is the first sentence of the document. We also include a feature indicating sentence position; we found in empirical studies the sentence to best match highlight H_1 is on average 10% down the article, the sentence to best match H_2 is on average 20% down the article, and the sentence to best match H_3 is 31% down the article. We also calculate the SumBasic score (Nenkova, Vanderwende, & McKeown 2006) over unigrams and bigrams of a sentence to estimate the importance of a sentence based on word frequency. We also include the similarity between the sentence and the article title as a feature.

The remaining features are based on third-party data sources. News search queries and Wikipedia pages are generated by humans expressing an interest in a topic. By targeting concepts that people are showing their interest in, we can extract sentences likely to be of higher interest to a reader. In other words, we treat the news query logs and Wikipedia as strong signals of human interest. NetSum is the first summarization system to leverage signals of human interest to produce relevant highlights.

We base several features on query terms frequently issued to Microsoft's news search engine http://search.live.com/news, and entities (titles of a Wikipedia page) found in the online open-source encyclopedia Wikipedia (Wikipedia.org 2007). Sentences containing query terms or Wikipedia entities should contain more important content. If a query term or Wikipedia entity appears frequently in an article, we assume highlights should include that term since it has been identified through outside interest in the topic.

System	Sent. #	ROUGE-1
Baseline	S_1, S_2, S_3	0.36
NetSum	S_1, S_8, S_{16}	**0.52**

Table 1: Block results for the block produced by NetSum and the baseline block for the example article. ROUGE-1 scores computed against the highlights as a block are listed.

We collected the daily top 200 most frequently queried terms in February 2007 for ten days. Our hypothesis is that a sentence with a higher number of news query terms is a better candidate highlight. We derive several features that express the frequency and relative importance of a query term in a sentence.

We perform term disambiguation on each document using an entity extractor (Cucerzan 2007). Terms are disambiguated to a Wikipedia entity only if they match a surface form in Wikipedia. For example, the surface forms "WHO" and "World Health Org." both refer to the World Health Organization and disambiguate to the entity "World Health Organization". We then extract features that express the frequency and importance of Wikipedia entities in the sentence.

Evaluation

NetSum produces a ranked list of sentences for each document. We create a block from the top 3 ranked sentences. We evaluate the performance of NetSum using ROUGE-1 and compare against the baseline of choosing the first three sentences as the highlight block. A similar baseline outperforms all previous systems for news article summarization (Nenkova 2005) and has been used in the DUC workshops (DUC 2001). Our task is novel in attempting to match highlights rather than a human-generated summary.

For each block produced by NetSum and the baseline, we compute the ROUGE-1 score of the block against the set of highlights as a block. For 73.26% of documents, NetSum produces a block equal to or better than the baseline block. The two systems produce blocks of equal ROUGE-1 score for 24.69% of documents. Thus, on average, NetSum produces a higher quality block under ROUGE-1.

Table 1 lists the sentences extracted by NetSum and the baseline, for the article shown in Figure 1. The NetSum summary achieves a ROUGE-1 score of 0.52, while the baseline summary scores only 0.36.

In feature ablation studies, we confirmed that the inclusion of news-based and Wikipedia-based features improves NetSum's peformance. We removed all news-based and Wikipedia-based features in NetSum. The resulting performance moderately declined. Although NetSum still outperforms the baseline without third-party features, leading us to conclude that RankNet and simple position and term frequency features contribute the maximum performance gains, the third-party features help target the content of the article that is most appealing to the reader. By bringing forward interesting content as opposed to simply content in the first three sentences, we are able to create more effective and informative highlights.

Conclusions and Future Directions

Our novel approach to automatic single-document summarization, NetSum, is the first to use both neural networks and third-party datasets for summarization. We evaluate our system on a novel task, highlight extraction, and a novel dataset, articles gathered from CNN.com. Our system is the first to show remarkable performance over the baseline of choosing the first n sentences of the document, where the performance difference is statistically significant.

An immediate future direction is to further explore feature selection. We found third-party features beneficial to the performance of NetSum and such sources can be mined further. We would also like to extract content across sentence boundaries. Most content in human-generated highlights is drawn from many sentences and sources. We hope to incorporate sentence simplification and sentence splicing and merging in a future version of NetSum.

Not only have we developed a state-of-the-art news story highlights system, we have also introduced novel techniques for leveraging sources of human interest to improve performance, supporting our hypothesis that text understanding is a function of both the text itself and what is noteworthy about the text. We believe our methods can extend beyond news article summarization to many other areas of research.

References

Burges, C.; Shaked, T.; Renshaw, E.; Lazier, A.; Deeds, M.; Hamilton, N.; and Hullender, G. 2005. Learning to Rank using Gradient Descent. In Raedt, L. D., and Wrobel, S., eds., *ICML*, 89–96. ACM.

CNN.com. 2007a. Cable news network. http://www.cnn.com/.

CNN.com. 2007b. Nigeria reports first human death from bird flu. http://edition.cnn.com/2007/WORLD/africa/01/31/ nigeria.bird.flu.ap/index.html?eref=edition_world.

Cucerzan, S. 2007. Large scale named entity disambiguation based on wikipedia data. In *EMNLP 2007: Empirical Methods in Natural Language Processing, June 28-30, 2007, Prague, Czech Republic*.

DUC. 2001. Document understanding conferences. http://www-nlpir.nist.gov/projects/duc/index.html.

Fellbaum, C., ed. 1998. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA.

Hirao, T.; Sasaki, Y.; Isozaki, H.; and Maeda, E. 2002. Ntt's text summarization system for DUC–2002. In *DUC 2002: Workshop on Text Summarization, July 11–12, 2002, Philadelphia, PA, USA*.

Jagarlamudi, J.; Pingali, P.; and Varma, V. 2006. Query independent sentence scoring approach to DUC 2006. In *DUC 2006: Document Understanding Conference, June 8–9, 2006, Brooklyn, NY, USA*.

Jing, H. 2002. Using hidden markov modeling to decompose human-written summaries. *Computational Linguistics* 4(28):527–543.

Lal, P., and Ruger, S. 2002. Extract-based summarization with simplification. In *DUC 2002: Workshop on Text Summarization, July 11–12, 2002, Philadelphia, PA, USA*.

Lin, C. 2004a. Looking for a few good metrics: Automatic summarization evaluation — how many samples are enough? In *Proceedings of the NTCIR Workshop 4, June 2–4, 2004, Tokyo, Japan*.

Lin, C. 2004b. Rouge: A package for automatic evaluation of summaries. In *WAS 2004: Proceedings of the Workshop on Text Summarization Branches Out, July 25–26, 2004, Barcelona, Spain*.

Luhn, H. 1958. The automatic creation of literature abstracts. *IBM Journal of Research and Development* 2(2):159–165.

Mani, I. 2001. *Automatic Summarization*. John Benjamins Pub. Co.

Mihalcea, R., and Tarau, P. 2005. An algorithm for language independent single and multiple document summarization. In *Proceedings of the International Joint Conference on Natural Language Processing (IJCNLP), October, 2005, Korea*.

Mihalcea, R. 2005. Language independent extractive summarization. In *ACL 2005: Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics, June, 2005, Ann Arbor, MI, USA*.

Nenkova, A.; Vanderwende, L.; and McKeown, K. 2006. A compositional context sensitive multi-document summarizer: exploring the factors that influence summarization. In Efthimiadis, E. N.; Dumais, S. T.; Hawking, D.; and Järvelin, K., eds., *SIGIR*, 573–580. ACM.

Nenkova, A. 2005. Automatic text summarization of newswire: Lessons learned from the document understanding conference. In *Proceedings of the 20th National Conference on Artificial Intelligence (AAAI 2005), Pittsburgh, PA*.

Sun, J.; Shen, D.; Zeng, H.; Yang, Q.; Lu, Y.; and Chen, Z. 2005. Web-page summarization using click-through data. In Baeza-Yates, R. A.; Ziviani, N.; Marchionini, G.; Moffat, A.; and Tait, J., eds., *SIGIR*. ACM.

Svore, K.; Vanderwende, L.; and Burges, C. 2007. Enhancing single-document summarization by combining ranknet and third-party sources. In *EMNLP 2007: Empirical Methods in Natural Language Processing, June 28-30, 2007, Prague, Czech Republic*.

Vanderwende, L. 2007. Answering and questioning for machine reading. In *AAAI 2007 Spring Symposium on Machine Reading, March 26–28, 2007, Stanford, CA, USA*.

Wikipedia.org. 2007. Wikipedia org. http://www.wikipedia.org.