5-1-2010

Focused multi-document summarization: Human summarization activity vs. automated systems techniques

Quinsulon L. Israel

Hyoil Han
Marshall University, hanh@marshall.edu

Il-Yeol Song

Recommended Citation

Israel, Q. L., Han, H., & Song, I. Y. (2010). Focused multi-document summarization: Human summarization activity vs. automated systems techniques. Journal of Computing Sciences in Colleges, 25(5), 10-20.

FOCUSED MULTI-DOCUMENT SUMMARIZATION: HUMAN SUMMARIZATION ACTIVITY VS. AUTOMATED SYSTEMS TECHNIQUES[*]

Quinsulon L. Israel
Drexel University
Philadelphia, PA 19104
(215) 397-4317
qisrael1906@acm.org

Hyoil Han
LeMoyne-Owen College
Memphis, TN 38126
(901) 435-1391
hyoil.han@acm.org

Il-Yeol Song
Drexel University
Philadelphia, PA 19103
(215) 895-2489
song@drexel.edu

ABSTRACT

Focused Multi-Document Summarization (MDS) is concerned with summarizing documents in a collection with a concentration toward a particular external request (i.e. query, question, topic, etc.), or focus. Although the current state-of-the-art provides somewhat decent performance for DUC/TAC-like evaluations (i.e. government and news concerns), other considerations need to be explored. This paper not only briefly explores the state-of-the-art in automatic systems techniques, but also a comparison with human summarization activity.

1. INTRODUCTION

Multi-document summarization aims to create a compressed summary of a collection of documents, while retaining the main characteristics and pertinent information within those documents. Adding a focus on this process creates a summary that also meets a specific request; hence, focused MDS. Extractive summarization uses whole sentences from the documents in the collection that “inform” the most regarding a specific request. However, abstractive summarization entails fusing bits and pieces of information from various sentences in the document collection into semantically similar information represented in a new way.

The process of summarization can be described in the following steps: 1) Process sentences using sentence boundary detection, optionally using stop word removal, tokenization, and other light syntactic parsing, 2) Score each sentence according to heuristically or automatically ascertained features: terms of the sentence, position of the terms and/or sentence, relevance with focus query, derived concepts, etc., 3) Rank and select the most salient and informative of those scored sentences according to some linear combination of features

Why do we need such an automated process? The burden of extensive amounts of information (i.e. information overload), so readily available in digital form, has placed a heavy cognitive burden (information overload) on society. It has become even more important for a user to be able to quickly read, understand and utilize digital information as easily as they can access it. For instance, fast generation of result snippets with search terms within context has been done in Google for search results [22]; however, a condensing of the actual text linked to each search result may distinguish better usefulness.

Focused MDS has a wide range of usefulness and is applicable across many domains. Considering the myriad of different types of texts publicly available, especially within an immensely large and ever growing corpus such as the World Wide Web, summarization representations and techniques need to be as robust and as efficient as possible, in order to be useful in the average laymen’s everyday life.

The remainder of this paper is organized as follows: Section 2 discusses the process of human summarization activity with some historical background, Section 3 discusses sentences processing in automatic systems and its use of linguistic techniques, Section 4 discusses the representations of documents and they facilitate MDS processing, Section 5 discusses how position information improvements sentence scoring, Section 6 discusses sentence scoring and ranking and using N-gram statistics versus language modeling and classification versus clustering, Section 7 discusses sentence selection and how the issues of redundancy, compression and coherence effect choosing the most salient sentences for a summary, and finally, Section 8 concludes this work.

2. THE INFLUENCE OF HUMAN SUMMARIZATION ACTIVITY

Human agreement studies have been performed since the 1960's [8], where there was little agreement between sentences extracted by machines using frequency expectation and those extracted by humans. For single document summarization, studies reported that 79% of the sentences in a human-generated abstract were a “direct match” to a sentence from a source document [10]. Therefore, it is natural that many current MDS (Multi-Document Summarization) systems have been made to be extractive, attempting to mimic this human behavior. However, for MDS, it has been shown that no more than 55% of the vocabulary contained in human-generated abstracts can be found within the source documents [5], signifying a move to a more abstractive means of summarizing for this task. Also, different people choose different content for their summaries [9, 12, 16]. This is also true for the human judges (usually four) creating model summaries in DUCs/TACs, and a correlation of automatic scores and these human model summaries is taken into account each year [6] to see how well automatic systems are improving, or not. Even multiple human summaries on the same collection of documents often do not have much agreement, with unigram overlap of only 40% in a study from Lin and Hovy [11]. In fact, in order for a consensus summary to be created from single document summaries from different humans, there must be as many as 30-40 summaries gathered before a consensus becomes stable [21].

3. SENTENCE PROCESSING

Processing a document can often involve several layers depending on the system’s task and strategy, as well as the representation, structure and syntactic variations of the document. Once processing is complete, the document is represented in a form in which the system can interface. Normally, “syntactic parsing” is performed at the most atomic level: each token of these sentences are “tagged” by machine-readable codes according to the grammatical rules and context of its respective language for part-of-speech (POS) [4].

The subjects, predicates and objects can also be parsed as terms or phrases into their semantic roles, such as pred(sub, obj). The functional relationships have started to gain more attention for summarization. This involves semantic role labeling and can be done with the aid of the well known and expensive PropBank [14], a collection of annotated propositions in predicates-argument form, or the freely available FrameNet [2], a collection of semantic elements of logical units centered on actions in frame form. By examining the roles of various elements in a sentence, a system can make better decisions on what exactly to compare. In fact, Nenkova et al. [12] used semantic annotation in their methodology to count occurrences of semantically equivalent content between summaries and the corpus, but only used shallow parsing to determine the respective important words. Consequently, they discovered factors that influence human judgments and used them to create their system, only to increase the accuracy of content counting during evaluation and distribution building. Wang et al. [24] used pair-wise sentence semantic similarity. For each sentence, a “frame” was created. Next, WordNet was used to discover the semantic relations of all terms in the first frame with those in the second to determine if there is a semantic relation (e.g. synonym, hypernym, etc.). However, this was solely used for their similarity metric. On the other hand, Ouyang et al. [13] used NER (Named Entity Recognition) and NER counts, informative words, semantic similarity using WordNet Lesk function, etc., but do not use deeper semantic parsing for comparisons between semantic roles of different sentences.

The order of linguistic processing and the depth at which it is performed is a touchy subject. For summarization tasks, it appears it is best to use “shallow” analysis techniques that do not require heavy computing resources and trained deep parsers. Mostly all research described in the previous sub-sections use shallow techniques such as POS tagging and NER. This has been the normal standard of operation; however, systems such as in Shi et al. [19] have started to use “deeper” syntactic analysis as a basis for nominal semantic role labeling.

4. DOCUMENT (TEXT) REPRESENTATIONS

What is meant by the representation of a document? Document representation is defined here as a tangible, either visual or physical, in our case digital, formalism of the concrete (text) and abstract (meaning) properties embodying the document. It is a finite way to automatically organize and store in some numeric form (index), and then later, retrieve the document for a system to interface with it (e.g. as in early information retrieval). Document representations must be applied to smaller bits of information such as user queries, questions, and more importantly, sentences in order to be useful toward MDS.

One of the fundamental and most efficient ways of representing a document was the Vector Space Model (VSM) [17]. In the VSM, a subset of terms appearing within the document at least once are placed in a weighted, linear combination of vectors of terms to represent the document; when all unique terms present in the document are used, it is also known as the “bag-of-words” model [18]. The term frequency and inverse document frequency of the term are multiplied together to determine the final weight for that term, and then, usually normalized. This allows for geometric calculations to be performed with the vectors of other documents or sentences for similarity. An information need such as a user query vector or a given topic vector can be compared against the document vector as well, for specific focus. The VSM became the de facto standard in text representation because of its simplicity and performance in large text collections, and is the basis of early summarization methods.

5. DOCUMENT POSITION

Until recently, the information provided by the structure of documents, or the hierarchy of its units of varying granularity, has mostly not been leveraged effectively. Research such as [13] for DUC 2005-2006 data has shown that sentence position can be of benefit when choosing the most informative sentences. The use of sentence position in an extractive method of summarization is intuitive since important information usually is placed in specific locations within a well-formed, but unstructured text based, formal document such as those found in DUC/TAC conferences. Since these corpora are normally composed of news articles, the importance of information is consistently related to how close it is to the beginning of the document, as reported by and found in the work of Yih et al. [26] involving word positions; news reporting seeks to provide the most salient information as quickly as possible. In fact, it is important to note that baseline systems in DUC/TAC and other government sponsored evaluations used the first *N* words in documents, the first (and sometimes last) sentences of every document, the first *N* sentences of the most recent document, or some other derivations of lead words/sentences to create baseline summaries. These automatic systems often outperformed peers. Hence, researchers have begun to use the same position information, concentrating on enhancing ROUGE scores on past DUC evaluation data.

There have also been studies of using surface level linguistic cues such as transition (connecting or "cue") terms [20] that identify changes in topic on both a document and a segment level. These transition terms can be weighted in a meaningful way to help improve focused judgments towards informative sentences. For instance, in work from Sun et al. [20], it was shown that "cue" words could identify topic shift in a single document as well as help identify the reversal of sub-topic order in a document compared to another, when the focus was on globally shared topics among documents.

6. SENTENCE SCORING AND RANKING

6.1 Simple Word (N-gram) Statistics vs. Statistical Language Modeling

Scoring for summarization involves calculating a numeric value for the significance of a sentence. However, of the utmost importance for focused summarization is scoring with an external factor besides sentence-to-sentence comparisons; not only do sentences need to be compared for similarity to each other, they also need to be skewed toward the user focus. After the score is calculated, the sentences are ordered from the highest score to the lowest. The most basic summarization, besides heuristic sentence choice based on position, is to count word-for-word matches of the terms of the focus input (query, question, or topic description) to the terms in sentences from the input collection. The simplest approach for frequency-based probabilities is in [12] where summation of within-document content word probabilities was empirically most performant.

Ouyang et al. [13] tested both frequency of occurrence and binary appearance on the DUC 2005 data in an indirect manner, using human reference summaries to learn linear feature combinations through Support Vector Regression, and on the DUC 2006 data for direct testing. Results show that frequency of occurrence was more performant for actual scoring estimation than binary appearance measures. However, the work of [4] showed that using simple summation of the binary appearances of topic terms along with so-called signature terms, to create an "approximate oracle score," can improve sentence scoring; signature terms for a topic under observation are believed to be those important terms derived from a sub-corpus of related documents but are used less in the rest of the corpus. More sophisticated than simply counting terms (or phrases) is the use of the cosine similarity measure on the vectors of these terms. Cosine similarity is a fast and efficient means to score sentences in terms of relatedness to each other; however, to add focus to such a process, it is necessary to also compare the focus input vector with those sentences. This is done in [25] by adding query-sensitive similarity to the centroid of two documents with a metric similar to cosine similarity but with the addition of a complex function on the query terms. In [3], the cardinality of overlap in topic terms is used in their linear feature combination function, along with cosine similarity between only the title of the topic in the DUC 2006 task and an individual sentence.

In fact, the work of Arora et al. [1] used the latent dirichlet allocation (LDA) model with estimated probabilities based on the corpus to create multiple summaries, and then, automatically chose the best. Their framework assumes that a complete sentence of a document belongs to only one topic. Performance was reportedly better than the two top systems in DUC 2002 for the ROUGE-1 (unigram) value.

6.2 Classification vs. Clustering

As stated by Ouyang et al. [13], classification based models are usually adopted to solve discrete problems; therefore, they are imprecise against continuous real-value functions like linear combination sentence scoring (using features). In any classification method, an item either belongs to a particular class, or it does not. However, more categorical delineations can be determined for the sentences if a real value can be calculated for each observation. As such, clustering has proven more performant; similar documents are gathered together according to the weight of their cosine similarity metrics, and the document that has the highest similarity in its cluster should contain the most worthy sentence for extraction into the summary about that topic. This process has been used on a sentential level as well, as in [23]. Wan and Yang [23] cluster the sentences of a document using various methods: k-means, agglomerative, and divisive. Using these clustering methods they were able to find the clusters within the document that truly represented sub-topics, and then extract the best similar sentence from each cluster centroid. This approach was similar to the MEAD system [15], which popularized the use of the document centroid for choosing the most salient sentence from a document. In this way, Wan and Yang [23] were able to choose the most salient yet diverse sentences. Also, terms can be clustered in order to help differentiate topics in a document as well [20]. The terms in each cluster can be associated by semantic similarity or by the entropy between different documents or different segments of documents.

Matrix representations of scoring equations add an efficient means of performing calculations due to the ability to use special manipulations such as in place normalization and finding eigenvectors. In fact, [24] use symmetric non-negative matrix factorization to calculate similarity to group similar sentences into clusters. In [3], a matrix was used to compute the hamming weights of terms in different sentences; the concern was to increase the significance of a sentence by the amount of pairs of significance words found. This hamming weight is then multiplied by the significant word count in segments and by word frequencies.

Using the techniques described above gives good results in terms of either speed of computation (cosine similarity in a low-dimensional degree) or better accuracy (matrix calculations). Clustering methods have shown a two-fold purpose in that they not only group sentences according to similarity with each other and a topic, but also delineate multiple sub-topics. It is also important to note that matrix calculations, though complex, also cluster sentences but have not been used specifically to delineate topics/subtopics to the authors' knowledge. Also, more studies regarding the use of combining very simplistic n-gram (bigram, named entities, phrases) frequencies and probabilities with text unit clustering (or matrix use) are needed to be carried out in depth.

7. SENTENCE SELECTION: REDUNDANCY, COMPRESSION, AND COHERENCE

Sentence selection is the second most important step for processing sentences as this step actually adds sentences to the summary. Its consideration in MDS should not be based solely on having the highest score, but also on sentence length, redundancy removal, compression and coherence. Many sentences may be important; however, the best among those containing redundant information must be selected.

Clustering via a cosine similarity metric is normally the method that has been used to group similar sentences in order to choose the sentence closest to the centroid and make subsequent choices from other clusters, as in [13] where maximum marginal relevance (MMR) was used with a threshold of 0.60. Also, Wan and Yang [23] used a variant of MMR between sentences to eliminate redundancy, whereas [24] used their own semantic similarity score. However, [4] used a pivoted QR algorithm instead of MMR and reported better performance because of it. Pseudo relevance feedback, which uses sparse information to retrieve more information from a corpus, was also used by [4] to help reduce redundancy.

Sentence Compression may allow more sentences to be chosen into a summary. Yih et al. [26] eliminated syntactic units based on predefined heuristic templates and added these new, modified sentences to the pool along with their original counterparts. During sentence selection, the best sentence between the original sentence and its modifications are automatically chosen; this helps alleviate the problems of over-simplification. It is also important to note that simplification can either help or harm the coherence of system summaries depending on the techniques used: adverb removal, parenthetical phrase removal, phrases within commas, etc.

As mentioned in [7], coherence is a key factor that affects a judge’s perception of readability for automatic summaries. [19] used longest common subsequences (LCS) of the sentence to be chosen with the sentence previously chosen for the summary, to try to maintain coherence. However, the weighting toward LCS is a tradeoff with attaining the highest similarity to a particular search focus.

Although there are many considerations to be made for sentence selection, it appears from the literature that there have not been many studies concerned with coherence. Coherence has begun to improve slightly during the DUC/TAC evaluations, but it appears to be mostly a byproduct of improved sentence ranking. Different concerns that can affect coherence such as topic segmentation, document structure, developing a scheme for information ordering, using cue words, etc. need to be studied.

8. CONCLUSION

Along with an explanation of multi-document summarization and its benefits, we have explored briefly some of the most novel techniques for improvement of MDS from the most promising and recently published research. Although the state-of-the-art of focused multi-document summarization has seen quite some improvement within the last decade, it has been shown that there is still room for further experimentation, using knowledge gained from the previously outlined human summarization activity; sentence processing, scoring, ranking and selection; and document representations and term position information. Human summarization activity was explored from the angle of instability in sentence consensus to the stability of salient words consensus. The subject of sentence processing and related activities was explored from light syntactic processing and semantics to term scoring and modeling techniques to the use of complex matrix calculations and the use of position information. A few of the previously lesser used, but important techniques were discussed; sentence compression and coherence improvement techniques are described in the context of the most simplistic yet powerful use.

REFERENCES

[1] Arora, R. and B. Ravindran, Latent dirichlet allocation based multi-document summarization, *Proceedings of the second workshop on Analytics for noisy unstructured text data*, 303, 91-97, 2008.

[2] Baker, C.F., C.J. Fillmore, and J.B. Lowe, The Berkeley FrameNet Project, *Proceedings of the 17th international conference on Computational linguistics*, 1, 86-90, 1998.

[3] Boudin, F. and J.M.T. Moreno, *NEO-CORTEX: A Performant User-Oriented Multi-Document Summarization System*, in *Proceedings of the 8th International Conference on Computational Linguistics and Intelligent Text Processing*, 551-562, 2007.

[4] Conroy, J.M., J.D. Schlesinger, and D.P. O'Leary, Topic-focused multi-document summarization using an approximate oracle score, *Proceedings of the COLING/ACL on Main conference poster sessions*, 152-159, 2006.

[5] Copeck, T. and S. Szpakowicz, Vocabulary Agreement Among Model Summaries and Source Documents, *ACL Text Summarization Workshop*, 2004.

[6] Dang, H.T. Overview of the DUC 2005, *Document Understanding Conference*, 2005.

[7] Dang, H.T. Overview of the DUC 2006, *Document Understanding Conference*, 2006.

[8] G. J. Rath, A. Resnick, and T.R. Savage, The formation of abstracts by the selection of sentences. Part I. Sentence selection by men and machines. *American Documentation*, 12(2), 139-141, 1961.

[9] Halteren, H.v., New Feature Sets for Summarization by Sentence Extraction, *IEEE Intelligent Systems*, 18(4), 34-42, 2003.

[10] Kupiec, J., J. Pedersen, and F. Chen, A trainable document summarizer, *Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval*, 68-73,1995.

[11] Lin, C.-Y. and E. Hovy. Manual and Automatic Evaluation of Summaries, *Document Understanding Conference*, 2002.

[12] Nenkova, A., L. Vanderwende, and K. McKeown, A compositional context sensitive multi-document summarizer: exploring the factors that influence summarization, *Proceedings of the 29th Annual International ACM SIGIR*

Conference on Research and Development in Information Retrieval, 573-580, 2006.

[13] Ouyang, Y., S. Li, and W. Li, Developing learning strategies for topic-based summarization, *Proceedings of the sixteenth ACM Conference on information and knowledge management*, 79-86, 2007.

[14] Palmer, M., D. Gildea, and P. Kingsbury, The Proposition Bank: An Annotated Corpus of Semantic Roles, *Computational Linguistics*, 31(1), 71-106, 2005.

[15] Radev, D.R., H. Jing, and M. Budzikowska, Centroid-based summarization of multiple documents: sentence extraction, utility-based evaluation, and user studies, *NAACL-ANLP 2000 Workshop on Automatic summarization,* 4, 21-30, 2000.

[16] Radev, D.R., Teufel, S., Saggion, H., Lam, W., Blitzer, J., Qi, H., Çelebi, A., Liu, D., Drabek, E., Evaluation challenges in large-scale document summarization, *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics,* 1, 375-382, 2003.

[17] Salton, G., A. Wong, and C.S. Yang, A Vector Space Model for Automatic Indexing, *Communications of the ACM*, 18(11), 613-620, 1974.

[18] Salton, G., Singhal, A., Mitra, M., Buckley, C., Automatic text structuring and summarization, *Information Processing & Management*, 33(2), 193-207, 1997.

[19] Shi, Z., Shi, Z., Melli, G., Wang, Y., Liu, Y., Gu, B., Kashani, M., Sarkar, A., Popowich, F., Question Answering Summarization of Multiple Biomedical Documents, *Proceedings of the 20th conference of the Canadian Society for Computational Studies of Intelligence on Advances in Artificial Intelligence*, 284-295, 2007.

[20] Sun, B., Prasenjit, M., Hongyuan, Z., Lee, G., John, Y., Topic Segmentation with Shared Topic Detection and Alignment of Multiple Documents, *Special Interest Group on Information Retrieval*, 2007.

[21] Teufel, S. and H.v. Halteren, Evaluating Information Content by Factoid Analysis: Human Annotation and Stability, *Empirical Methods in Natural Language Processing*, 2004.

[22] Turpin, A., Turpin, A., Tsegay, Y., Hawking, D., Williams, H. E., Fast generation of result snippets in web search, *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, 127-134, 2007.

[23] Wan, X. and J. Yang, Multi-document summarization using cluster-based link analysis, *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, 299-306, 2008.

[24] Wang, D., Wang, D., Li, T., Zhu, S., Ding, C., Multi-document summarization via sentence-level semantic analysis and symmetric matrix factorization, *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, 307-314, 2008.

[25] Wei, F., Li, W., Lu, Q., & He, Y., Query-sensitive mutual reinforcement chain and its application in query-oriented multi-document summarization, *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, 283-290, 2008.

[26] Yih, W.-T., Goodman, J., Vanderwende, L., Suzuki, H., Multi-document summarization by maximizing informative content-words, *International Joint Conferences on Artificial Intelligence*, 1776-1782, 2007.