Semantic Argument Frequency-Based Multi-Document Summarization

Cem Aksoy, Ahmet Bugdayci, Tunay Gur, Ibrahim Uysal, Fazli Can
Bilkent Information Retrieval Group,
Bilkent University, 06800, Ankara, Turkey

{caksoy, canf}@cs.bilkent.edu.tr, abugdayc@cs.purdue.edu, {tgur, uysal}@ug.bilkent.edu.tr

Abstract— **Semantic Role Labeling (SRL) aims to identify the constituents of a sentence, together with their roles with respect to the sentence predicates. In this paper, we introduce and assess the idea of using SRL on generic Multi-Document Summarization (MDS). We score sentences according to their inclusion of frequent semantic phrases and form the summary using the top-scored sentences. We compare this method with a term-based sentence scoring approach to investigate the effects of using semantic units instead of single words for sentence scoring. We also integrate our scoring metric as an auxiliary feature to a cutting edge summarizer with the intention of examining its effects on the performance. The experiments using datasets from the Document Understanding Conference (DUC) 2004 show that the SRL-based summarization outperforms the term-based approach as well as most of the DUC participants.**

Keywords- Frequency, Semantic role labeling, Summarization

I. Introduction

Generic multi-document summarization (MDS) has been a great interest to Information Retrieval (IR) and Natural Language Processing (NLP) societies in the recent years. The proposed systems use sentence extraction [12, 20] as well as more sophisticated linguistic methods [7, 10]. Some applications also explore a hybrid of these two approaches, increasing the presentation quality and coherence of extracted summaries by eliminating, simplifying and reformulating the sentences [3, 8]. Either using pure extraction or hybrid methods that go beyond the extraction, the main component of the summarization is the sentence scoring metric. The sentences to be included in the summary are selected according to the sentence score calculated by this metric. Proven summarization systems address this problem by using statistical approaches [23], machine learning techniques [11], graph-based methods [4, 19] or directly assigning salience scores to sentences based on a suite of features [20].

Although the potential use of semantic roles in IR and NLP tasks has been suggested [5], its application to summarization has largely remained undiscovered. In particular, there are systems using SRL for topic theme representation [6] and sentence similarity calculation to cluster the sentences [22]. All of these studies about SRL show that semantic analysis of the documents improves the summarization quality.

In this paper, we propose a sentence scoring approach based on semantic arguments for generic MDS. We also consider a similar approach using term-frequency for sentence scoring to explore the effects of using semantic arguments instead of individual terms. Our experiments show that the semantic argument frequency-based sentence scoring approach produces better results than the term-based approach.

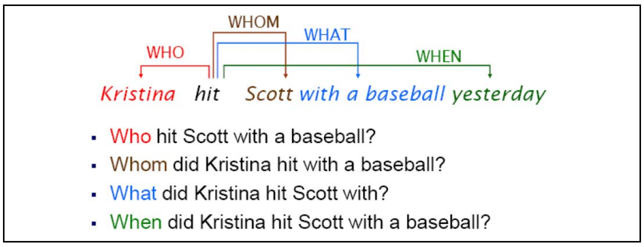
The major contribution of this study is using semantic arguments in generic MDS in a simple way. Our first attempt for utilizing semantic phrases to score sentences ended up with promising results. In addition to that, integrating SRL-based sentence scores to a state-of-the-art summarization system - MEAD- as an auxiliary feature improved its performance, which supports the idea of using SRL in MDS.

The remainder of this paper is organized as follows. We first introduce SRL and the motivation behind using SRL as a feature in summarization. In Section III, we explain the components of SRL-based MDS system, namely, *SrlSum*. We give a detailed description of our experimental setup in Section IV, so that the results presented in this paper can be reproduced. Finally, we discuss our results and conclude our findings in Section V and VI, respectively.

II. Semantic Role Labeling

A. Semantic Role Labeling

Semantic roles are defined as the relationships between syntactic constituents and the predicates. Most sentence components have semantic connections with the predicate, carrying answers to the questions such as *who*, *what*, *when*, *where*, *why*, and *how* as shown in the Fig. 1, adopted from [24]. The task of semantic role labeling is to identify these roles for each predicate in the sentence [5, 16].



- Who hit Scott with a baseball?
- Whom did Kristina hit with a baseball?
- What did Kristina hit Scott with?
- When did Kristina hit Scott with a baseball?

Figure 1. Semantic arguments of a sentence

978-1-4244-5023-7/09/$25.00 ©2009 IEEE

As a result of this process, typical roles like *agent*, *patient*, and *instrument* are classified. The adjuncts, indicating locative, temporal, manner, etc. features are identified as well. Fig. 2, adopted from [24], demonstrates a sentence parsed into semantic arguments.

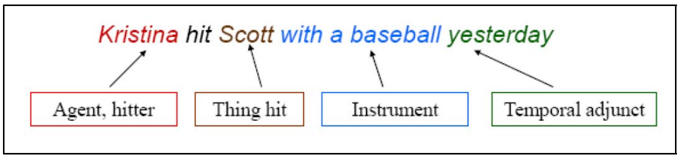


Figure 2. Semantic roles of the constitutents

After the Conference on Computational Natural Language Learning (CoNLL)[1] proposed semantic role labeling as the shared task in 2004 and 2005 [2], quite effective automatic parsers have been developed. ASSERT software[2] is one of these parser tools, which uses Support Vector Machines [18]. It is trained to tag PropBank arguments, thematic roles, and opinions. We use this publicly distributed tool throughout our study.

TABLE I. SEMANTIC ARGUMENTS OF A SENTENCE

Original Sentence	Hurricane Mitch cut through the Honduran coast like a ripsaw on Friday.
SRL-parsed Sentence	[*ARG0* Hurricane Mitch] [*TARGET* cut] [*ARG2* through the Honduran coast] [*ARGM-ADV* like a ripsaw] [*ARGM-TMP* on Friday].

Table I shows a sample sentence that illustrates how SRL process labels the constituents. As in this example, many arguments consist of more than one term. Thus, SRL-based summarization approach differs from classical term-based systems.

There is an issue related to the SRL parsing process that we should take into account. For each verb in a sentence, the SRL parser provides a different frame. It considers the verb as the predicate of the sentence and tries to label the remaining part of the sentence as proper arguments. However, if the selected verb is not the actual predicate, the parser fails to identify most of the words as part of an argument. Therefore, we consider the frame that leaves the least number of terms unlabeled as the correct parse of the sentence. In our calculations, we use just this correct frame.

Table II lists three different frames of a sentence produced by taking the following verbs as predicates: “say, accuse, engage”. Since *accuse* and *engage* are not the actual predicates, the corresponding frames cannot correctly identify the semantic arguments and leaves out a number of unlabeled terms. On the other hand, the correct frame with the predicate *say* manages to label each term in the sentence as part of some argument.

TABLE II. DIFFERENT FRAMES OF A SENTENCE

Correctly parsed	[ARG0 Qin] [TARGET said] [ARG1 a civil affairs official in the Hubei provincial capital of Wuhan accused him of engaging in illegal activities]
False parsed	Qin said [ARG0 a civil affairs official in the Hubei provincial capital of Wuhan] [TARGET accused] [ARG1 him] [ARG2 of engaging in illegal activities]
False parsed	Qin said a civil affairs official in the Hubei provincial capital of Wuhan accused [ARG0 him] of [TARGET engaging] [ARG2 in illegal activities]

B. *Motivation for using SRL*

Summarization task requires understanding the document and presenting the salient parts. For human annotators, this is a straightforward process. However, for automatic summarizers, figuring out which information is important becomes really challenging. In extractive summarization, this task is accomplished by determining the sentences to be included in the summary. The most common method to solve this problem is to rank the sentences according to their informativeness.

Since human annotators tend to include most frequent words in their summaries [15], word-based frequency calculations for sentence scoring is commonly used for MDS. However, this approach is semantically immature, because many words do not carry semantic information alone.

Our motivation for using SRL in sentence scoring for MDS originates from this concern. Instead of using individual terms for sentence scoring, we exploit semantic arguments, which hold more comprehensive information about the whole set of documents we are trying to summarize.

III. PROPOSED METHOD (*SRLSUM*)

The summarization method we propose, namely *SrlSum*, works in the following way as illustrated in Fig. 3. First, the documents are given to the SRL parser. Then we extract the semantic arguments from each parsed sentence. Both these arguments and the original documents are stemmed [17] and their stop words are removed. Later, we calculate the frequencies of each semantic argument in the document set on a particular topic. Using these arguments, each sentence is scored. Subsequently, the top scoring sentences are selected one-by-one and put into the summary, while eliminating redundancy. Finally, the selected sentences are reordered to create a coherent summary. The lengths of the summaries are fixed to 665 bytes in order to make a fair comparison of the results from DUC '04 participants.

Each of these tasks is explained in the following subsections.

[1] http://www.cnts.ua.ac.be/conll2005/

[2] http://cemantix.org/assert

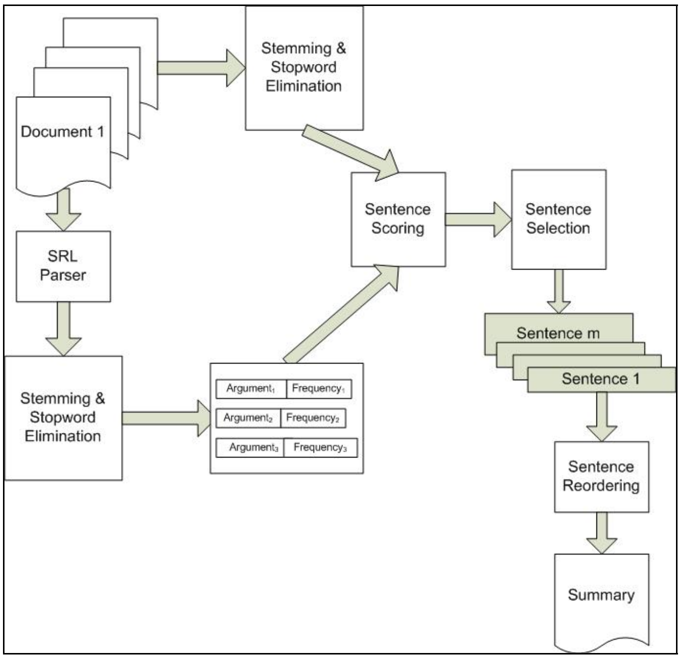


Figure 3. Overview of *SrlSum*

A. *SRL Parser*

SRL Parser takes each sentence in the document set and labels the semantic word phrases properly. We refer to these phrases as semantic arguments or shortly arguments.

After stemming and removing the stop words, we calculate the frequency of each argument in the document set.

B. *Sentence Scoring*

In order to calculate the sentence scores, which indicate how important the sentences are, we use each semantic phrase and its frequency. We calculate the score of a sentence by summing up its similarity to the phrases in the document set multiplied by their frequencies. Hence, the sentences which are likely to contain frequent phrases are considered to be more important. An implementation of this approach is presented in Algorithm 1.

S_i : Sentence i.
P_j : Semantic phrase j.
N_j : The frequency of P_j in the document set.
I_{S_i} : The importance score assigned to S_i so far.
Cos_Sim (S_i , P_j): The cosine similarity value of S_i and P_j .
n : Total number of sentences in the document set.
m : Total number of semantic phrases in the document set.

Algorithm 1 Sentence Scoring

```
for $i \leftarrow 1$ to $n$ do
    for $j \leftarrow 1$ to $m$ do
        $I_{S_i} \leftarrow I_{S_i} + (Cos\_Sim(S_i, P_j) * N_j)$
    end for
end for
```

While scoring a sentence, we do not check whether the sentence exactly contains a semantic phrase to receive some importance value from that phrase. The reason is that even one word difference in a long argument prevents the argument from being effective on the score. Instead, we use the cosine similarity value between the sentence and the arguments; so that each argument related to the sentence contributes to the score.

In the term-based sentence scoring approach, we use this algorithm with terms (instead of phrases) and their frequencies. Also, instead of the similarity measure between a sentence and the phrase, we look for an exact inclusion of the term.

C. *Sentence Selection*

One of the crucial problems in MDS is getting rid of the repeating information, which is referred as redundancy removal. Since the important sentences in a set of documents are likely to contain similar information, just taking the top scoring sentences to form the summary is prone to cause redundancy. Therefore, while selecting the top scored sentences we compare each one with the already selected ones to prevent having repetitive information in the summary. Therefore, if the candidate sentence's similarity to the already selected sentences does not exceed a threshold, then we take that sentence. In the experiments, a sentence pair is considered to be similar, if the cosine similarity ratio exceeds 0.80, an empirically chosen threshold.

D. *Sentence Reordering*

So far, we have selected the sentences to form the summary. The question remains is in which order we should present these sentences. According to our observations on the different documents about same topic, although the expressions differ, they usually follow the same story flow. Taking this into account, we decided to re-order the selected sentences considering their positions in the original documents relative to document length. We believe this approach makes the summary more coherent since we try to preserve the flow of the original documents at some degree.

The calculation of the positions is done as described in the following equation.

$$Position(S_i) = \frac{\sum_{j=1}^{i}(length(S_j))}{\sum_{j=1}^{n}(length(S_j))}$$

The position value of a sentence is the ratio of the total length from the beginning till the end of the sentence to the document length. Our intuition is that since the story is told in similar order in every related document, the sentences in close positions express similar events. Hence, ordering selected sentences according to their positions will resemble the story summarized by human annotators.

IV. EXPERIMENTAL SETUP

A. *Dataset and Evaluation Environment*

National Institute of Standards and Technology (NIST) organized DUC between 2001 and 2007 whose focus was on summarization. In 2004, DUC proposed a task for generic MDS, Task 2, in which the participants are asked to produce a short summary[3] of the given set of documents. For this task, DUC 2004 dataset includes newswire/paper documents from TDT collections. It has 50 clusters each of which contains approximately 10 documents. In addition to these news articles, the dataset includes annotations prepared by professionals in order to evaluate system summaries. Since this dataset is a commonly used benchmark, we use it in our evaluations as well.

As the evaluation tool, we utilize ROUGE[4] [13], which has been used in DUC since 2004.

B. *Integration into a State-of-the-art Summarizer*

We also integrate *SrlSum*'s sentence scoring feature into another summarizer. We intend to show that our scoring feature can be used to improve the performance of known summarization systems. For this purpose, we choose to use MEAD[5] which allows using combination of different features for extractive summarization.

MEAD is an extractive multi-document summarizer. It comes with a default feature configuration as a complete summarizer. Centroid[19], position, and length features are included in this default configuration. LexRank [4] is also provided by MEAD as a scoring feature. The system basically calculates scores for each sentence with the features in the configuration. Then, the scores from different features are combined with the weights provided by the configuration and summaries are formed from the extracted sentences according to the sentence scores. Sentences are finally reorganized for the presentation order.

During the integration process, we calculate sentence scores using *SrlSum*. These extracted scores are then formatted into MEAD feature files by scaling the values into [0, 1] interval. Using these files, we experiment with two different configurations of MEAD. In both of the configurations, we use length and position metrics as supporting features. Length metric is given 9 as a threshold value to make MEAD ignore sentences shorter than 9 words and position feature is given a fixed weight of 1.0 which basically gives sentences a score between 0 and 1 according to their position in the document. These values are used as suggested in [4]. Table III gives the exact weights for $\text{MEAD}_{\text{Centroid}}$ and $\text{MEAD}_{\text{LexRank}}$ configurations and *SrlSum* integrated versions of them. The summaries extracted according to these configurations are then evaluated using ROUGE. The results are presented and discussed in Section V.

TABLE III. FEATURE WEIGHT CONFIGURATION USED IN SRLSUM-MEAD INTEGRATION

	Features		
Systems	**SrlSum**	**Centroid**	**LexRank**
MEAD$_{\textbf{Centroid}}$	0.0	1.0	0.0
MEAD$_{\textbf{Centroid\_SrlSum}}$	1.0	1.0	0.0
MEAD$_{\textbf{LexRank}}$	0.0	0.0	1.0
MEAD$_{\textbf{LexRank\_SrlSum}}$	1.0	0.0	1.0

V. RESULTS AND DISCUSSION

A. *SrlSum Results*

We compare the results of *SrlSum* and the term frequency-based sentence-scoring approach in Table IV. The results indicate that when semantic units are used for sentence scoring the summarization system produces better performance with respect to term-based system.

TABLE IV. DUC '04 ROUGE SCORES FOR TOP 5 PARTICIPANTS, TERM-BASED APPROACH, BASELINE, AND *SRLSUM*

Systems	ROUGE-1	ROUGE-2	ROUGE-L
MEAD$_{\textbf{Centroid\_SrlSum}}$	**0.3847**	**0.0990**	**0.3376**
Peer 65	0.3825	0.0922	0.3306
MEAD$_{\textbf{LexRank\_SrlSum}}$	**0.3811**	**0.0946**	**0.3322**
MEAD$_{\textbf{LexRank}}$	0.3807	0.0928	0.3297
SrlSum	**0.3803**	**0.0911**	**0.3329**
MEAD$_{\textbf{Centroid}}$	0.3769	0.0938	0.3299
Peer 104	0.3747	0.0856	0.3259
Peer 35	0.3746	0.0834	0.3326
Peer 19	0.3744	0.0805	0.3238
Term-based	**0.3736**	**0.0886**	**0.3209**
Peer 124	0.3712	0.0831	0.3226
Baseline$_{\textbf{LeadBased}}$	0.3594	0.0800	0.3139

Additionally, we listed the ROUGE scores of the top 5 systems at the DUC '04 and the results of a baseline system, which creates the summaries using the leading sentences in the documents. Even though in *SrlSum* we did not consider sentence positions or other summary quality improvement techniques such as sentence reduction, its overall performance is promising. The use of semantic roles in summarization can make considerable improvements to the existing systems even though the results presented here do not report a significant difference.

The results show that integrating our scoring metric into MEAD improves the summarization performance. Table IV compares MEAD's Centroid-based summarization results and *SrlSum* scoring metric integrated version of it. It shows that *SrlSum* provides an improvement when used with Centroid-based feature, according to ROUGE-1 and ROUGE-L metrics. For LexRank feature, *SrlSum* integrated version scores better but the difference between the scores is negligible.

[3] 665 characters

[4] We use ROUGE-1.5.5 with these exact parameters: -c 95 -b 665 -m -n 4 -w 1.2 -a. We re-evaluated 2004 DUC participant summaries according to indicate that when semantic units are used for sentence scoring the summarization system produces better performance.

[5] http://www.summarization.com/mead.

VI. Conclusion and Future Work

In this paper, we examined the usage of semantic argument frequencies in sentence scoring for generic MDS problem. We ranked sentences according to the importance of semantic phrases they contain, where the importance of a phrase is determined by its frequency. We also present the results of a term-based method for sentence score calculation. The results we obtained using DUC 2004 dataset showed that SRL-based sentence scoring approach outperforms both all the participants of DUC 2004 workshop except one and the term-based approach. Besides, when SRL-based scoring is integrated to MEAD as a supplementary feature, its performance increased, which supports the idea that it is rational to utilize semantic arguments in MDS.

The performance of our system supports the claim that sophisticated feature calculations may not necessarily perform better than simpler approaches [15].

In the future, we are planning to accomplish the following goals:

- Explore the effects of SRL-based scoring on query-oriented summarization.
- Experiment on different datasets to examine the significance of the proposed method.

Acknowledgment

We would like to thank the distributors of ASSERT software and ROUGE evaluation package.

References

[1] Carbonell, J. and Goldstein, J. 1998. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st Annual International ACM SIGIR Conference*, Melbourne, Australia. SIGIR '98. ACM, pp. 335-336.

[2] Carreras, X. and Marquez, L. 2004. Introduction to the CoNLL-2004 shared task: Semantic role labeling. In *Proceedings of HLT-NAACL 2004 Workshop: Eighth Conference on Computational Natural Language Learning (CoNLL-2004)*, Boston, Massachusetts, USA. 2004. pp. 89–97.

[3] Conroy, J., Schlesinger, J., Goldstein, J. and O'Leary, D. 2004. Left-brain/right-brain multi-document summarization. In *Proceedings of 4th Document Understanding Conference (DUC-2004)*.

[4] Erkan, G. and Radev, D. R. 2004. Lexrank: Graph-based lexical centrality as salience in text summarization. *Journal of Artificial Intelligence Research (JAIR)* 22, 457–479.

[5] Gildea, D. and Jurafsky, D. 2002. Automatic labeling of semantic roles. *Computational Linguistics*, 28, 3, 245–288.

[6] Harabagiu, S. and Lacatusu. F. 2005. Topic themes for multi-document summarization. In *Proceedings of the 28th Annual International ACM SIGIR Conference*, Salvador, Brazil. SIGIR '05. ACM, pp. 202-209.

[7] Jing, H. 2002. Using hidden Markov modeling to decompose human-written summaries. *Computational Linguistics*, 28, 4, 527–543.

[8] Jing, H. and McKeown, K. 2000. Cut and paste based text summarization. In *Proceedings of the 1st North American Chapter of the Association for Computational Linguistics Conference*, Seattle, Washington, USA. pp. 178–185.

[9] Jones, K. S. 2007. Automatic summarising: The state of the art. *Information Processing and Management* , 43, 6, 1449-1481.

[10] Knight, K. and Marcu, D. 2000. Statistics-based summarization - step one: Sentence compression. In *Proceedings of AAAI/IAAI*, pp. 703–710.

[11] Kupiec, J., Pedersen, J. and Chen, F. 1995. A trainable document summarizer. In *Proceedings of the 18th Annual International ACM SIGIR Conference*, Seattle, Washington, USA. SIGIR '95. ACM, pp. 68–73.

[12] Li, W., Wu, M., Lu, Q., Xu, W. and Yuan, C. 2006. Extractive summarization using inter- and intra- event relevance. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association For Computational Linguistics*, Sydney, Australia. pp. 369–376.

[13] Lin, C.-Y. 2004. Rouge: A package for automatic evaluation of summaries. In *Proceedings of ACL 2004 workshop on Text Summarization Branches Out*, Barcelona, Spain, pp. 74-81.

[14] Lin, C.-Y. and Hovy, E. 2002. Automated multi-document summarization in neats. In *Proceedings of the Second International Conference on Human Language Technology Research*, San Diego, California, USA. pp. 59-62s.

[15] Nenkova, A., Vanderwende, L. and McKeown, K.. 2006. A compositional context sensitive multi-document summarizer: exploring the factors that influence summarization. . In *Proceedings of the 29th Annual International ACM SIGIR Conference*, Seattle, Washington, USA. SIGIR '06. ACM, pp. 573-580.

[16] Palmer, M., Gildea, D. and Kingsbury, P. 2005. The proposition bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31 ,1, 71–106.

[17] Porter, M. 1980. An algorithm for suffix stripping. *Program*, 14 ,3, 130-137.

[18] Pradhan, S., Ward, W., Hacioglu, K., Martin, J. and Jurafsky, D. 2004. Shallow semantic parsing using support vector machines. In *Proceedings of the Human Language Technology Conference/North American chapter of the Association of Computational Linguistics (HLT/NAACL)*, Boston, MA, USA.

[19] Radev, D. R., Jing, H., & Budzikowska, M. 2000. Centroid-based summarization of multiple documents: sentence extraction, utility-based evaluation, and user studies. In *ANLP/NAACL Workshop on Summarization*, Seattle, WA, USA. pp. 21-29.

[20] Ribeiro, R. and Matos, D. 2007. Extractive summarization of broadcast news: Comparing strategies for european portuguese. In *Lecture Notes in Computer Science*, Springer, 4629, 115–122.

[21] Schiffman, B., Nenkova, A. and McKeown, K. 2002. Experiments in multi-document summarization. Experiments in multi-document summarization. In *Proceedings of the Second International Conference on Human Language Technology Research*, San Diego, California, USA. pp. 52-58

[22] Wang, D., Li, T., Zhu, S. and Ding, C. 2008. Multi-document summarization via sentence-level semantic analysis and symmetric matrix factorization. In *Proceedings of the 31st Annual International ACM SIGIR Conference*, Singapore, Singapore. pp. 307-314.

[23] Witbrock, M. J. and Mittal, V. O. 1999. Ultra-summarization (poster abstract): A statistical approach to generating highly condensed non-extractive summaries. In *Proceedings of the 22nd Annual International ACM SIGIR Conference*, Berkeley, California, USA. pp. 315–316.

[24] Yih, W. T. and Toutanova, K. 2006. Automatic semantic role labeling. In *Proceedings of the Human Language Technology Conference of the NAACL*, New York, USA. pp. 309-310.