NUS at DUC 2005: Understanding Documents via Concept Links

Shiren Ye, Long Qiu, Tat-Seng Chua and Min-Yen Kan
School of Computing, National University of Singapore
{yesr|qiul|chuats|kanmy}@comp.nus.edu.sg

Abstract

The primary goal of our participation in DUC 2005 is two-fold. One is to benchmark the performance of a method of computing sentence semantic similarity. The other is to test the effectiveness of a new redundancy minimization formula inspired by Maximal Marginal Relevance (MMR). By using only these two features and eschewing other heuristics, our system performed competitively, achieving the top automated ROUGE scores among participants this year. This is a revised version of our notebook paper.

1 Introduction

This year's Document Understanding Conference (DUC) task is quite different from the previous ones. As suggested in [1], the task includes a set of relevant documents (a document cluster) and a *topic* in the form of one or more query sentences. In addition, a granularity is specified as being either "*specific*", if specific instances of events, people, locations, *etc.* are to be highlighted, or "*general*", if high-level generalization is preferred in the summary. Participating systems are supposed to generate a multi-document summary that best describes the topic at the right granularity level. Meanwhile, the summary has to cover as much of the important information as possible in the document cluster.

Sentence similarity has been widely investigated in the text summarization community. In single document summarization, it is used to calculate how representative the sentences are with respect to the whole document. In multi-document summarization, it also serves detect redundancy among candidate output sentences. Beyond the simple "bag-of-words" approach, recent approaches incorporate advanced word features and their relations to capture sentence similarity. For example, SIMFINDER[2] considers a richer set of text features, including proper noun overlap, verb overlap, WordNet collocation, *etc.,* to compute sentence similarity when it generates sentence clusters, the centroids of which form the summaries. LexPageRank[3] is a graph-based approach to calculating the centrality of sentences. To minimize redundancy in summaries, a Maximal Marginal Relevance (MMR) [4] component is used to extract key sentences. The latter approach requires sentence similarity to be explicitly computed, too.

As is stated in [1], the system task can be regarded as a "topic-oriented, informative multi-document summarization" and the goal is a "compressed version" of a document cluster. This suggests two attributes that a sentence should possess in order to make a good summary: it should be highly relevant to the topic and at the same time, the more document content it covers, the better. In our research, we first calculate an overall similarity score between each sentence and the remainder of the document cluster. This overall similarity score reflects the strength of *representative power* of the sentence in regard to the rest of the document cluster and is used as the primary sentence ranking metric while forming the summary. We also employ a module similar to MMR to build the summary incrementally, minimizing redundancy and maintaining the summary's relevance to the topic.

In the remainder of this report, we describe our implementation of concept link and its use in summarization. We conclude with a discussion on system performance and notes for future improvement.

2 System Overview

The input given to our summarization system is composed of a cluster of relevant documents and a topic. At the preprocessing phase, our system ignores the document boundaries in the document cluster. It takes all the documents as a single document which it delimits into sentences for further analysis. The topic is treated similarly: only its sentences boundaries, if any, are detected. No other features of the topic are collected. Figure 1 shows an overview of our system, including this preprocessing phase. The components in the figure are briefed below, in order of execution:

1. A *tokenizer* delimits numbers, words, and punctuations under the given format.
2. A *sentence delimiter* detects and annotates sentence boundaries.
3. Resulting sentences are then fed to the *concept link calculator*. This component, further described in Section 3, calculates the semantic

similarity among the sentences in the cluster, and between a sentence and the given topic.

4. A *sentence ranker* then iteratively sorts the sentences according to their relevance to the topic and representativeness of unselected sentences in the cluster. This is further described in Section 4. Only the top ranking sentence of each iteration is considered for inclusion in the summary.
5. For specific summaries, a *specific detector* is also executed. A specific summary is supposed to describe and name specific events, people, places, *etc*. In order to cover more such specifics in the summaries, we try to avoid unnecessary repetition of specific entities and reserve room for other distinct ones. Usually, these entities are represented as Named Entities (NEs) in text. The specific detector simply focuses on NEs and doubles the value of a penalty factor δ (refer to formula 5) if candidate sentences contain NEs that exist in sentences already selected for inclusion in the summary.
6. Finally, an *extractor* selects the top-ranked sentence in each iteration, concatenating it into the summary. Steps 4-6 are run repeatedly until the length reaches the pre-defined limit.

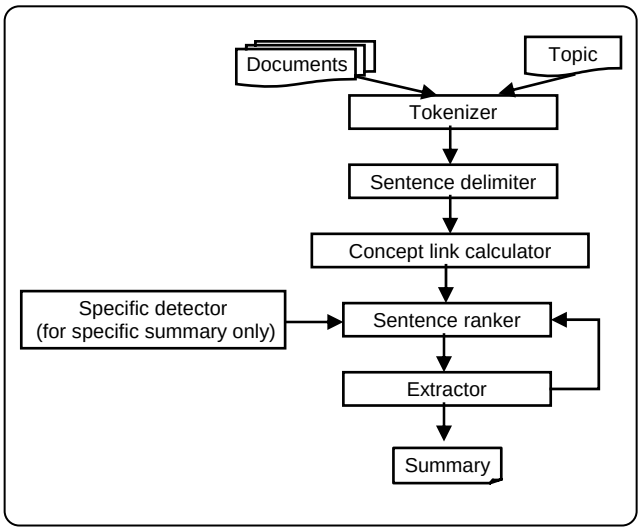


Figure 1. System Overview

3 Concept Link for Sentence Ranking

In DUC, multi-document summary extraction is considered the procedure of selecting key sentences that capture the main points of the documents in a cluster. Finding an appropriate similarity measure is pivotal since it will determine how representative the candidate sentences are of its document. Sentences should receive a higher score when they are more semantically linked to other sentences and the query topic. They represent not only themselves but also other semantically linked sentences which have not been selected for the summary.

Although there is no single, best way to compute the semantic similarity between sentences, we approximate this by accounting for similarity that exists due to:

1. Synonyms or hyponyms (such as war ~ battle);
2. Derivational morphological variations (such as decision ~ decide, Argentine ~ Argentina);
3. Inflectional morphological variations (such as relations ~ relation).

Although stemming algorithms can address some of the morphology-related problems, it does not help with synonymy and hyponymy which past studies have shown to be very common. Different authors may use a variety of expressions to describe similar or identical topics. However, the underlying entities and actions (whether realized as single words or phrases, NPs, VPs or ADJPs) are the same from author to author. Our goal is to detect and link these identical concepts, no matter how they are realized.

Usually, these concepts are represented by open class words, such as nouns, verbs, adjectives and adverbs rather than closed-class words. For example, in the three sentences from different DUC 2005 documents, as shown in Figure 2, the underscored terms are concepts. These concepts provide strong basis for measuring similarity between sentences. Instead of relying on stemming, POS tagging and sentence parsing, which can be problematic, we focus on developing a simple but effective technique to extract such concepts.

In our system, *concepts* are formally defined as the words that remain after removing closed-class words (such as the articles, prepositions and conjunctions, *etc*.) in sentences. We check WordNet for entries that match sequences of consecutive words as multi-word concepts (as denoted by double underscores in Figure 2). Remaining words are each treated as individual concepts. Named Entities not in WordNet such as "Document Understanding Conference" could have been treated as a single concept. But in our system they are currently broken into single-word concepts.

Argentine-British relations since the Falkland Islands War in 1982 have gradually improved.

Thirteen years after the war between Britain and Argentina over the Falkland Islands, Argentina still makes a ritual reference to Argentina's sovereignty over those islands.

Argentina was still obsessed with the Falkland Islands even in 1994, 12 years after its defeat in the 74-day war with Britain.

Figure 2. Sample sentences from DUC 2005

According to [5], word senses that are related are often defined using shared words. Based on this idea, the semantic similarity of two concepts can be measured by the ratio of word co-occurrence between their definitions. In our work, the definition of a sense of a concept consists of information from WordNet: its *synset* (a set of synonyms of the sense), *gloss* (the explanation of the sense with possibly specific examples), and direct *hypernyms* (is-a relation) and *meronyms* (has-a relation). For two concepts c_i and c_j with sense definitions $S_1^i, S_2^i, ..., S_m^i$ and $S_1^j, S_2^j, ..., S_n^j$ respectively, their semantic similarity is

$$sim(c_i, c_j) = \max_{1 \le x \le m, 1 \le y \le n} \frac{\sum_{z=0}^{k} | strOverlap^z(S_x^i, S_y^j) |^2}{| S_x^i | \times | S_y^j |} \quad (1)$$

Here $strOverlap^z(S_x^i, S_y^j)$ is defined as the *z-th* shared segment in S_x^i and S_y^j. In the formula, each such shared segment should not be subsumed in another longer shared segment. Consider the following two sense definitions (only synset, gloss and direct hypernyms are shown; meronyms are omitted for brevity):

> Canada: [Canada]$_{synset}$ [(a nation in northern North America; the French were the first Europeans to settle in mainland Canada) "the border between the United States and Canada is the longest unguarded border in the world"]$_{gloss}$; [North American country, North American nation]$_{hypernyms}$ and
>
> USA: [United States, United States of America, America, the States, US, U.S., USA, U.S.A.]$_{synset}$ [(North American republic containing 50 states - 48 conterminous states in North America plus Alaska in northwest North America and the Hawaiian Islands in the Pacific Ocean; achieved independence in 1776)]$_{gloss}$; [North American country, North American nation]$_{hypernyms}$.

There are ten shared segments that can be considered as *strOverlaps* : "North American country", "North American nation", "North America", "United States", "in the", "and", two "in"s and two "the"s. We do not consider for example "American country" as a valid *strOverlap* since it is subsumed in a longer shared segment "North American country". Neither do we count more than one *strOverlap* for "United States", because the second instance of it in the definition of "USA" does not have an unaccounted counterpart in that of "Canada" to overlap. Taking isolated closed-class words like "in" or "the" as *strOverlaps* may introduce noises, but they are not filtered out in our current system.

We can then explore the semantic similarity between the sentences based on the similarity between concepts. For sentence s_i containing concepts $c_{i,1}, ..., c_{i,m}$ and sentence s_j containing concepts $c_{j,1}, ..., c_{j,n}$, we define the set of concept links, $CL(s_i,s_j)$, as a set consisting of disjointing concept pairs $<c_{i,x}, c_{j,y}>$ whose $sim(c_{i,x}, c_{j,y})$ is greater than a predefined threshold θ (set to 0.2). Figure 3 illustrates how $CL(s_i,s_j)$ is constructed by a greedy algorithm. $CL(s_i,s_j)$ is used later to compute the semantic similarity between sentence s_i and s_j.

```
CONCEPTLINKS(s_i, s_j)
    CL ← φ;
    Ci ← CONCEPTDETECTOR(s_i);
    Cj ← CONCEPTDETECTOR(s_j);
    while(Ci ≠ φ and Cj ≠ φ)
        if(max_{c_{i,x}∈Ci, c_{j,y}∈Cj} sim(c_{i,x}, c_{j,y}) > θ)
            CL = CL ∪ < c_{i,x}, c_{j,y} >;
            Ci = Ci − c_{i,x}; // removal
            Cj = Cj − c_{j,y}; // removal
        else
            break;
        end
    end
return CL;
```

Figure 3. Function ConceptLinks(s_i, s_j)

Here, the computational cost must be addressed since computing CONCEPTLINKS(s_i,s_j) requires a scan through all possible pairs of senses for all pairs of concepts. To reduce the run-time computational cost, we pre-compute the semantic similarity between all possible pairs of WordNet entries (the concepts in our discussion) offline and store non-zero pairs into a hash table. At runtime, the semantic similarity of a pair of concepts can be found by an O(1) hash lookup. As the semantic similarity of most concept pairs is zero, the size of this hash table is acceptable (238,728 records in total) for keeping in main memory.

4 Sentence Ranking and Selection

As stated earlier, sentences in a document cluster are viewed as coming from a single concatenated document instead of from individual documents. They are initially ranked by their representative power – the weighted sum of their similarity to all other sentences. The similarity between two sentences, $sim(s_i,s_j)$, is calculated as the weighted sum of the strength of each concept link in $CL(s_i,s_j)$. We believe our idea of concept links can

outperform word co-occurrence as it highlights not only identical words, but also words that are semantically related. Thus for each sentence, we have its representative power $Rep(s_i)$:

$$\text{Rep}(s_i) = \sum_{s_j \in D - s_i} \omega_{len}^{s_i} \omega_{len}^{s_j} sim(s_i, s_j) \quad (2)$$

where

$$sim(s_i, s_j) = \sum_{<c_x, c_y> \in CL(s_i, s_j)} \omega_{freq}^{c_x} \omega_{freq}^{c_y} sim(c_x, c_y) \quad (3)$$

Here, D is the set of sentence in the document cluster, ω^c_{freq} is TF*IDF[1] of the concept c and ω^s_{len} is the weight of the corresponding sentence s. In order to alleviate the bias towards longer sentences (which have more concepts), ω^s_{len} is set to $log_2(|\#concept| + 1)$ in our experiments.

In [4], it is proposed that a document has MMR if it is relevant to the query and contains minimal similarity to previously selected documents. It is defined as,

$$MMR \stackrel{def}{=} \arg\max_{D_i \in R-S} [\lambda sim(D_i, Q) - (1-\lambda) \max_{Dj \in S} sim(D_i, D_j)] \quad (4)$$

where S is the set of already selected documents from a document collection, Q is the query and R is the ranked list of documents that an information retrieval system suggested.

There are two points that make the original MMR unsuitable for our summarization task. First, the original MMR relies positively only on $sim(D_i, Q)$. However, in a summarization task, the unit is sentence, which is generally much shorter than a document or a passage targeted in the IR task. Consequently, the number of open class words in sentences would be relatively small. This causes many sentences in R to have a $sim(s_i, Q)$ score of 0. In some cases, these sentences could have strong links to other sentences in the document cluster (We observed that there is little positive correlation between $Rep(s_i)$ and $sim(s_i, Q)$), and are good candidates for a general summary. For a topic-oriented summary that according to [1] should be a "compressed version" of the document cluster, these sentences should still be considered despite their dissimilarity to the topic.

Second, the original MMR definition considers only the maximal similarity $\max_{s_j \in S} sim(s_i, s_j)$, which is not always optimal. As the mock-up scenario in Figure 4 shows, MMR passes on sentence s_a (which has concepts unseen in S, the set of selected sentences, and overlaps with individual sentence in S up to n concepts) to choose sentence s_b (whose concepts are all distributed among the three sentences in S, but at most $m<n$ in each of them). We think the better choice is to take all concepts seen in S into account and rank s_a higher than s_b, thus achieving high content coverage by minimizing redundancy.

s_1:[Argentina and British]$_1$ [fought over Falkland islands]$_2$ [in 1982]$_3$.

s_2:[Commercial relations have continued to improve]$_4$ [between UK and Argentina]$_1$.

s_3:[Mr Douglas Hurd, Britain foreign secretary,]$_5$ [is to visit Argentina]$_6$ [early next year]$_7$.

s_a:[Britain lifted military protection zones around the Falklands]$_{unseen1}$ [in 1990]$_{unseen2}$, [8 years after]$_3$ [the Argentina-British]$_1$ [war over the area]$_2$.

s_b:[Mr Hurd's]$_5$ [visit to Argentina is the first by a cabinet minister]$_6$ [since the Falklands conflict]$_2$, [indicating improved diplomatic relations between]$_4$ [UK and Argentina]$_1$.

Figure 4. Motivation Behind MMR Modification. For illustration's purpose, each sentence is segmented into indexed content units, which can be regarded as approximations of "summarization content units" [6], instead of concepts.

To address these two problems, we employ the following modified version of MMR. It considers how similar a candidate sentence is to the whole S. Counting SCU_1 twice as it is in both s_1 and s_2, it turns out s_b overlaps S by 6 SCUs in total while s_a overlaps S by 4 SCUs. The modified MMR will thus favor s_a in spite of the great overlap between s_a and s_1. Note that since the system task of DUC 2005 is topic-oriented, the influence of similarity between the candidate sentence and the topic is important and therefore also included. We keep denoting topic as Q in the formula:

$$MMR_{\text{mod}} \stackrel{def}{=} \arg\max_{s_i \in R-S} [\lambda \cdot \text{Rep}(s_i) + (1-\lambda) \cdot sim(s_i, Q) - \delta \cdot \sum_{s_k \in S} sim(s_i, s_k)] \quad (5)$$

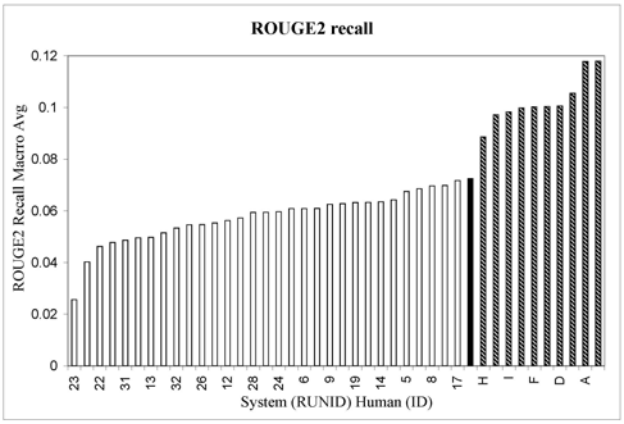
Here, δ is the penalty factor which is used to decrease the rank of such sentences that are similar to the already selected summary sentences. It is manually set to 1.5 and will be doubled when appropriate during specific summary generation. The factor λ represents a tuning factor between a sentence's representative power and its relevance to the topic, and is set empirically to 0.8 in our experiments.

[1] We use the data available at http://elib.cs.berkeley.edu/docfreq to get the term document frequency.

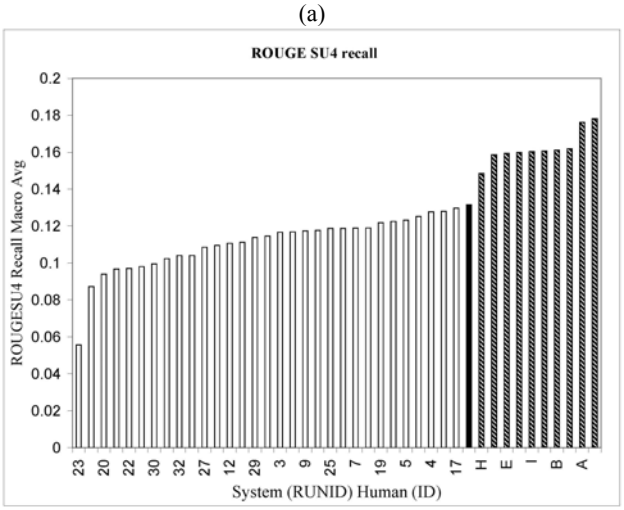
5 Experimental Results

This is the first time for our team to participate in DUC, and we have been concentrated on developing a general purpose text summarization tool, building on our experience in TREC on the tasks of question answering, information fusion and information extraction. We then developed and tested our system based on our existing NLP and IR infrastructure. We have just concentrated on investigating proper use of semantic similarity and MMR.

Unlike many other existing approaches, we have not used other heuristics such as sentence position, length, centroid, and title overlap. We also have not attempted to fine-tune our system to the domain characteristics of DUC corpus, although we plan to do this in future participation. We feel these features are definitely desirable and may improve system performance.



(a)



(b)

Figure 5. ROUGE Scores of system and human summaries. Blank, solid and striped bars denote peer systems, our system and human annotators, respectively.

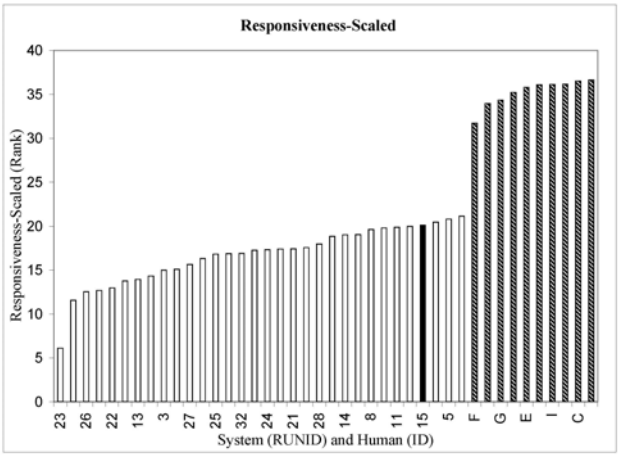


Figure 6. Scaled Responsiveness Scores. Blank, solid and striped bars denote peer systems. our system and human annotators, respectively.

As shown in Figures 5 and 6, our system achieves relatively good results with respect to ROUGE measures [7] and the scaled responsiveness score. In particular, the scores rank us among the top systems with respect to the suggested ROUGE-2 and ROUGE-SU4 measures. The average recall under ROUGE-2 and ROUGE-SU4 are 7.25% and 13.16%, respectively. Furthermore, we also observed that for our system, the difference between recall and precision for each document cluster is small. The values of recall, precision and F_1 are close to one another.

Table 1 shows results of additional experiments we did after this paper's first notebook version was submitted. They assess the performance impact of our two key components with respect to the automatic ROUGE assessment measures. A combination of both (Concept Link + MMR_mod) outperforms settings by around 2% where either the original MMR (Concept Link + MMR_org) or a word co-occurrence model (Word Co-occurrence + MMR_mod) is used instead to summarize the 50 clusters.

Setting	ROUGE 2	ROUGE SU4
Concept Link +MMR_mod	**7.25%**	**13.16%**
Concept Link + MMR_org	5.87%	11.27%
Word Co-occurrence +MMR_mod	5.06%	11.03%
Concept Link-based ***Rep(s)*** only	6.73%	12.45%
Concept Link-based ***Sim(Topic, s)*** only	6.21%	12.32%

Table 1. Contributions of Key Components

Furthermore, the last two rows in the table also show that when only sentence representative power (Concept Link-based *Rep(s)* only) or sentence similarity with the topic (Concept Link-based *Sim(Topic,s)* only) is considered, the system still achieves reasonable performance compared to peer

systems. However, these settings are suboptimal as the two factors of document coverage and topic relevance are not used in sentence selection simultaneously.

Last but not least, it is interesting to note that the system based on simple word co-occurrence uses no heuristics. Its ROUGE scores (recall 5.06% and 11.03% respectively) lie in the middle of all peer systems. We thus conclude that our modified MMR is a suitable component for multi-document summarization, given a reasonable sentence similarity measure.

6 Conclusion

In DUC 2005, we have investigated a model of sentence similarity feasibility as modeled by concept links. These links consider all the senses of phrases and words found in WordNet. These concept links serve as basis to compute the similarity of sentences. We also propose a modified version of MMR, which we feel is more suited for summarization. This version overcomes sparse data problems caused by the short length of sentences encountered in summarization. Experiments in the DUC competition validate our system as one of the top performing sets. Our additional experiments indicate that our modified MMR is largely responsible for the improvement over other peer systems. We plan to experiment further on how to fully automate our system's manual parameter settings.

7 References

[1] Enrique Amigo, Julio Gonzalo, Victor Peinado, Anselmo Peñas and Felisa Verdejo. *An Empirical Study of Information Synthesis Task*. In Proceedings of ACL 2004.

[2] Vasileios Hatzivassiloglou, Judith L. Klavans, Melissa L. Holcombe, Regina Barzilay, Min-Yen Kan, and Kathleen R. McKeown. *SIMFINDER: A Flexible Clustering Tool for Summarization*. In Proceedings of the Workshop on Summarization in NAACL `01. Pittsburg, Pennsylvania, USA, June 2001.

[3] G. Erkan, D. R. Radev, *The University of Michigan at DUC 2004*, DUC 2004.

[4] J. Carbonell and J. Goldstein. *The Use of MMR, Diversity-Based Reranking for Reordering Documents and Producing Summaries*. In Proceedings of SIGIR'98. Melbourne, Australia, 1998.

[5] Michael Lesk. *Automatic Sense Disambiguation Using Machine Readable Dictionaries: How to tell a Pine Cone from an Ice Cream Cone*. In Proceedings of SIGDOC'86, Toronto, Ontario, June, 1986.

[6] Rebecca Passonneau and Ani Nenkova. *Evaluating content selection in summarization: The pyramid method lexical information*. In Proceedings of the Human Language Technology Research Conference/North American Chapter of the Association of Computational Linguistics, Boston, Massachusetts, USA, 2004.

[7] Lin, Chin-Yew and Franz Josef Och, *Automatic Evaluation of Machine Translation Quality Using Longest Common Subsequence and Skip-Bigram Statistics*, In Proceedings of ACL 2004, system available at http://www.isi.edu/~cyl/ROUGE/.