HUNUS at TAC 2009: with Better Performance in Update Summarization Task

Tingting He, Jinguang Chen, Zhuoming Gui, Fang Li
Huazhong Normal University, 430079, Wuhan, China
tthe@mail.ccnu.edu.cn,cjg2003@hutc.zj.cn, Wenzheng38@sina.com, fang__lf@163.com

Abstract:
In this paper, a new semantic-based extractive summary methodology is put forward. The approach makes use of WordNet synset to obtain sentence semantic similarity. The scoring function is expressed as a linear combination of two features: the query-independent feature and query-related feature. An efficient scoring function of considering historical information for the update summarization task is proposed in this paper, which considering historical information in the sentence scoring stage instead of considering them after the sentences are already scored. The system architecture as well as its linguistics processing parts are described. Finally, we present the results of our participation in TAC 2009 with possible perspectives.

1 Introduction

The Text Analysis Conference (TAC) is one of the most well-known series of workshops that provides the infrastructure necessary for large-scale evaluation of natural language processing methodologies and technologies. The update summarization task of TAC 2009 is similar to that in TAC 2008, which aims to generate two short and fluent summaries respectively for two chronologically ordered document sets to meet the topic-relevant information need. The summary of the second document set should be written under the assumption that the user has already read the earlier documents and should avoid repeating old information and inform the user of novel information about the specific topic.

Our summarization system, HZNUS (Tingting He et al., 2008), participated in TAC tasks since 2008. In TAC 2008, HZNUS selects sentences using a feature fusion based sentence scoring method, to identify the sentences with high query-relevant and high information density. First, we score each sentence by computing its similarity with query. Second, we re-score every candidate sentence's representative feature by the importance of candidate sentence in document set. At last, we adopt MMR (J. Carbonel&J. Goldstein, 1998) for sentence extracting. We reused the same framework with some enhancements for the TAC 2009 task, and these enhancements seem to be active. We submitted two results these year, HZNU1 (peerID 53) used the developed method and HZNU2 (peerID 14) used the same method as in TAC2008. Evaluated results show that HZNU1 performed far better than HZNU2, whether evaluated manually (Manual) or automatically (BE and ROUGE), indicates the effectiveness of the development made in HZNU2009. HZNUS's best system ranked around 15 in all 55 peer system in. Evaluate results also shows that HZNUS's best system performed far better in task B than in task A, indicates that our historical information removal module is efficient, comparing with decreased performance in task B than task A in TAC 2008.

The rest of the paper is organized as follows. In Section 2 we give a detail description of the HZNUS system. Section 3 presents the evaluation results. Finally, we conclude in Section 4.

2 The HZNUS System Architecture

HZNUS consists of the following five steps:

1. Content extracting and sentence splitting.
2. Syntactic-based anaphora resolution

3. Semantic sentence similarity obtaining.
4. Sentence scoring.
5. Removal of the historical information
6. Redundancy removal.

2.1 Content Extracting and Sentence Splitting

The document preparation step begins with content extracting. We develop a sentence splitter in HZNUS. The sentence splitter absorbed a majority of rules from a Perl sentence splitter with the following additional changes made to compensate for erroneous splits:

1. Remove all double quotation marks.
2. Wherever there is a colon, we eliminate the content lead it as well as itself.
3. Wherever there is a semicolon, we treat the content lead and follow it as complete sentences respectively.
4. We eliminate all question sentences and plaint sentences to avoiding their influences for fluency of summaries.

2.2 Anaphora Resolution

We used JavaRAP developed by Long Qiu et al. (2004) to build a syntactic-based anaphora resolution module in HZNUS. JavaRAP is an implementation of the classic Resolution of Anaphora Procedure (RAP) (Lappin&Leass, 1994). It resolves third person pronouns, lexical anaphors, and identifies pleonastic pronouns.

2.3 Semantic Similarity Obtaining

A log-linear part-of-speech tagger is used to extract effective words from sentences, including nouns, verbs and adjectives. After that Sentences are transferred into a vector of effective words.

Considering two sentences

$$S_i = \{w_{i1}, w_{i2}, \ldots, w_{im}\}$$

$$S_j = \{w_{j1}, w_{j2}, \ldots, w_{jn}\}$$

Score of effective words in S_i *and* S_j can be obtained as the following:

$$a_l = \max\{sim(w_{il}, w_{jk}) \mid k = 1, 2, \ldots, n\}, \quad l = 1, 2, \ldots, m$$

$$b_p = \max\{sim(w_{ip}, w_{jk}) \mid k = 1, 2, \ldots, m\}, \quad p = 1, 2, \ldots, n$$

Where $sim(w_i, w_j)$ is the similarity of two effective words, its obtained by using WodNet synset. If synset of word W_i in WordNet is U and synset of word W_j in WordNet is V, similarity of them can be obtained as the following:

$$sim(w_i, w_j) = \frac{U \cap V}{U \cup V}$$

If a word can not be found in WordNet, its synset is defined as itself.

Similarity of sentence S_i *and* S_j can be obtained as the following:

$$sim(S_i, S_j) = \left(\frac{1}{m} \bullet \sum_{l=1}^{m} a_l + \frac{1}{n} \bullet \sum_{p=1}^{n} b_p\right) / 2$$

And sentence scores are obtained from two sentence scoring model.

2.4 Sentence Scoring

To evaluate whether a sentence is appropriately included in the summary, two factors are considered. One is the association between a sentence and the query, and the other consideration is the information density of a sentence compared to other sentences in the topic set. Generally, more responsive a sentence is to the query and more information density the sentence contains, more possible the sentence is to be included in the final summary.

2.4.1 Query-related Sentence Scoring

In HZNUS, the association between a sentence and the query is obtained with a modified relevance-based language modeling using HAL spaces.

The query-related score can be obtained just the same as in TAC 2008:

$$QR(S_i) == \frac{\sum_{W_i \in S} P(w_i) \sum_{q_j} pHAL(q_j \mid w_i)}{L}$$

Details can be found in (Tingting He et al., 2008).

2.4.2 Query- independent Sentence Scoring

HZNUS obtained the query-Independent score with the semantic sentence similarity as described in section 2.3.

In a document set which including *k* sentences, the query-independent similarity can be obtained as the following:

$$QI(S_i) = \frac{1}{k} \sum_{j=1}^{k} sim(S_i, S_j)$$

The final score of sentence S_i is

$$\text{Score}(S_i) = \beta \bullet \frac{QI(S_i)}{\sum_{i=1}^{k} QI(S_i)} + (1-\beta) \bullet \frac{QR(S_i)}{\sum_{i=1}^{k} QR(S_i)}$$

Where k is the number of sentences in the whole document set, β is the factor of mixing the query-related and query-independent score. We trained β in the TAC 2008 data.

2.5 Removal of the historical information

In order to remove historical information form topic set B, we used a novel sentence scoring method in HZNUS. The basic idea of this method is considering historical information in the sentence scoring stage instead of considering them after the sentences are already scored.

Our sentence scoring method in task B can be defined as the following:

$$S_i = \alpha \bullet \beta \bullet \frac{QI(S_i)}{\sum_{i=1}^{k} QI(S_i)} + (1-\beta) \bullet \frac{QR(S_i)}{\sum_{i=1}^{k} QR(S_i)}$$

Where α is the historical factor, it can be obtained as the following:

$$\alpha = 1 - \underset{S_j \in A}{Sim}(S_i, S_j)$$

Where A is collection of all sentences in topic set A.

Notice that we only considering the historical information in the query-independent part, since for the query is just the same for both set A and set B.

2.6 Redundancy removal.

To avoid including redundancy information in the summary, we used an evolved Maximal Marginal Relevance method (J. Carbonel&J. Goldstein, 1998) in HZNUS, with slightly enhancement from TAC 2008.

Our MMR method can be defined as the following:

$$S_i = \arg\max_{S_i \in R-F}[Score(S_i) * \max_{S_j \in F} sim(S_i, S_j)]$$

Where F is collection of the selected sentences, R is collection of the sorted sentences in the topic set.

3 Results

HZNUS submitted 3 results, peerID 53, 14 respectively. To compare the effectiveness of different method, HZNU1 (peerID 53) used the developed method and HZNU2 (peerID 14) used the same method as in TAC2008.

Table 1 shows ranks of HZNUS in manual evaluation.

Task	peerID	average modified (pyramid) score	average numSCUs	average numrepetitions	macroaverage modified score with 3 models	average linguistic quality	average overall responsiveness
A	53	21	22	7	21	22	16
	14	28	32	33	14	48	34
B	53	18	17	15	18	25	22
	14	47	47	48	47	48	45

Table 1: Performance of HZNU in TAC 2009 Update Summarization Task (Manual Evaluation)

Table 2 shows ranks of HZNUS in automatic evaluation.

Task	PeerId	Evaluation Method		
		ROUGE-2	ROUGE-SU4	BE
A	53	21	22	24
	14	27	31	33
B	53	14	15	14
	14	41	42	44

Table 2: Performance of HZNU in TAC 2009 Update Summarization Task (Automatic Evaluation)

Evaluated results show that HZNU1 performed far better than HZNU2, whether evaluated manually (Manual) or automatically (BE and ROUGE), indicates the effectiveness of the development made in HZNU2009. Evaluate results also shows that HZNUS's best system performed far better in task B than in task A, indicates that our historical information removal module is efficient, comparing with decreased performance in task B than task A in TAC 2008.

4 Conclusions

This paper presented our participation in the update summarization task of TAC 2009 summarization track. In our HZNU system for topic-focused summarization task, we present a semantic based strategy to select the topic relevant sentences and adopt a novel algorithm to remove historical information. This approach got encouraging performance according to the official evaluation results. For the future works, we plan to improve it by using sophisticated natural language processing techniques.

Acknowledgements

This work was supported by the Major Research Plan of National Natural Science Foundation of China (NO. 90920005), the National Natural Science Foundation of China (NO.60773167), the Project in the National Science & Technology Pillar Program in the Eleventh Five-year Plan Period (NO. 2006BAK11B03), the 973 National Basic Research Program (NO. 2007CB310804) and the Programme of Introducing Talents of Discipline to Universities (NO. B07042).

References

[1] Tingting He, Jinguang Chen, Zhuomin Gui and Fang Li, CCNU at TAC 2008：Proceeding on Using Semantic Method for Automated Summarization Yield, http://www.nist.gov/tac/publications/2008/papers.html

[2] 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.

[3] Long Qiu, Min-Yen Kan and Tat-Seng Chua. (2004). A Public Reference Implementation of the RAP Anaphora Resolution Algorithm. In proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC 2004). Vol. I, pp. 291-294.

[4] Lappin, S. Leass, H. 1994. An algorithm for pronominal anaphora resolution, Computational Linguistics, 20(4), 535-561.

[5] Kristina Toutanova and Christopher D. Manning. 2000. Enriching the Knowledge Sources Used in a Maximum Entropy Part-of-Speech Tagger. In Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP/VLC-2000), pp. 63-70.

[6] Kristina Toutanova, Dan Klein, Christopher Manning, and Yoram Singer. 2003. Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network. In Proceedings of HLT-NAACL 2003, pp. 252-259.