Experiments in TREC 2004 Novelty Track at CAS-ICT

Hua-Ping Zhang[+*] **Hong-Bo Xu**[+] **Shuo Bai**[+] **Bin Wang**[+] **Xue-Qi Cheng**[+]

[+]Institute of Computing Technology, Chinese Academy of Sciences, Beijing, 100080, CHINA

[*]Graduate School of the Chinese Academy of Sciences, Beijing, 100039 CHINA

zhanghp@software.ict.ac.cn

1 Introduction

The main task in Novelty Track is to retrieve relevant sentences and remove duplicates from a document set given a TREC topic. This track took place for the first time in TREC 2002 and it is refined to four tasks in TREC 2003. Besides 25 relevant documents, irrelevant ones are given in this year of Novelty track. In other words, a given document is either relevant or irrelevant to the topic. There are 1808 documents in 50 TREC topics. Average 11.18 documents are noise for each topic. In topic N75, the number of noise is 45. Once we mistook an irrelevant document as relevance, all results in the document are wrong. Except the document retrieval, more limited information could be applied in the last three tasks than ever. Among the first 5 given documents, average 3.14 documents are relevant and average 2.76 are new. Especially, 9 topics have no relevant sentence in the first 5 ones.

In TREC2004, ICT divided Novelty track into four sequential stages. It includes: customized language parsing on original dataset, document retrieval, sentence relevance and novelty detection. The architecture in novelty is given in Figure 1. In the first preprocessing stage, we applied sentence segmenter, tokenization, part-of-speech tagging, morphological analysis, stop word remover and query analyzer on topics and documents. As for query analysis, we categorized words in topics into description words and query words. Title, description and narrative parts are all merged into query with different weights. In the stage of document and sentence retrieval, we introduced vector space model (VSM) and its variance, probability model OKAPI and statistical language model. Based on VSM, we tried various query expansion strategies: pseu-feedback, term expansion with synset or synonym in WordNet[1] and expansion with highly local co-occurrence terms. With regard to the novelty stage, we defined three types of new degree: word overlapping and its extension, similarity comparison and information gain. In the last three tasks, we used the known results to adjust threshold, estimate the number of results, and turned to classifier, such as inductive and transductive SVM.

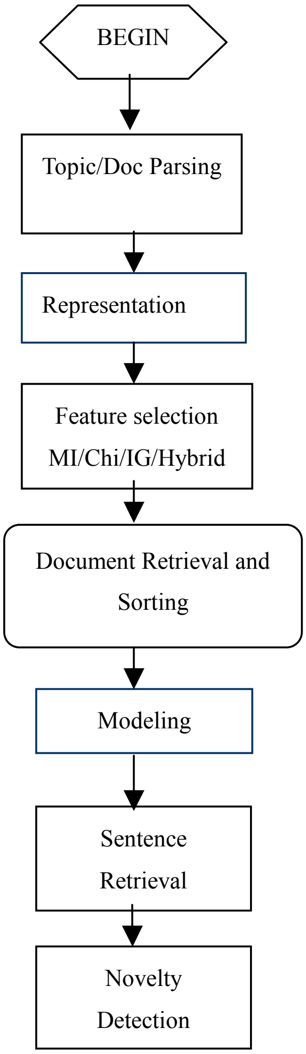

Figure 1. Novelty Architecture

2 Customized Language Parsing

Because of limitation of data size, the information contained in sentences is much less than documents. Therefore, we have to extend the single word through customized language parsing. In the general architecture, customized language parsing and feature selection is indispensable pre-processing for various novelty models. We applied customized language parsing on topics and documents. The process includes sentence segmenter (only on topic description and narrative) and query generation, tokenization, part-of-speech (POS) tagging, morphological analysis, stop words and POS removing, indexing and feature selection. We would give some brief introduction in the following sections.

2.1.1 Sentence segmenter and query generation

The queries should be extracted from the TREC topic fields: title, description and narrative. Such fields are written in natural language. Except stopping words, words in topic could be categorized into description words and query words. Query words expressed the user's intension, while description words tried to accomplish a sentence and had no relation with true query content. As demonstrated in Figure 2, description words are in italic font and other words are normal in topic N53. Query is also divided into positive and negative part. Negative query tends to be sentences contain "irrelevant" and "not", showed as the underlined sentence in the following figure. However, sentences in topic are not segmented as the given documents. We have to add a sentence segmenter. A sentence may end with '?' or '!' as well as a dot '.'. A blank or carriage return can also be a sentence boundary in special case. But not alldot '.' is a sentence boundary. It could be the dot in a URL, abbreviation and digit. Sentence segmentation is incorporated into tokenization process.

The final query is generated from topic fields with different weights. In our Novelty 2004 experiments, the ratio among title, description and narrative are 4:2:1 respectively.

2.1.2 POS tagging and tokenization

We modified rule-based POS tagger written by Eric Brill [2]. Besides some bugs in memory, Eric system cannot directly be employed on original lines but token sequences. For instance, "I'm a researcher." should be tokenized as "*I 'm a researcher .*". Various cases are considered in tokenization and sentence segmenter.

To be *relevant*, a *document contains* any *opinion* of the family, the public, the police, the judicial or even those of the news reporter as to the reason for the dragging. Also *relevant* is the ongoing investigation into the crime, the suspects, the juror selection, and the trial results regarding the dragging death of James Byrd, Jr. *Documents* that reflect only on the incident without elaboration are not relevant.

Figure 2. Segment in topic N53

2.1.3 Morphological Analysis

Document oriented application tends to make use of stemming on different word forms. However, it is far from requirement in sentence level. "computers" and "computing" would both be stemmed as "comput" using Porter stemming tool[3], which was very popular with information processing. However, the result eliminated the main difference between two words. It would affect sentence retrieval, and greatly reduce the performance of novelty detection. Hence, we apply morphological analysis instead of stemming. Based on WordNet codes, a powerful morphological analyzer was built with an English dictionary. It could get the base form of various word forms, such as noun plural, verb past tense and passive, adjective and adverb comparative.

2.1.4 Stop Remover

Three strategies are introduced in stop remover. Firstly, words whose POS are noun, verb, adjective, adverb are reserved, and others are filtered out. Digits in topic are also important. In topic N56, the digit in “Woodstock 99 music festival reunion in Rome, NY” emphasized the festival year. So, we also reserved digits in topic. Secondly, reserved words would be excluded if they were included in the stop word list. Finally, any word reserved in topic could not be eliminated from documents

After pre-processing, sentences in topics and documents are converted to sequences of term ID. It is easier to compute. The intermediate data is given in the following figure.

```
<SENTENCE>He/PRP will/MD also/RB seek/VB the/DT general/JJ 's/POS extradition/NN
to/TO Spain/NNP ./.    </SENTENCE>
<SENTENCE_Q>seek/VB general/JJ extradition/NN spain/NNP </SENTENCE_Q>
<SENTENCE_QID> 24358/NN 28581/JJ 63288/VB 66983/NNP</SENTENCE_QID>
```

Figure 3. Segment in the intermediate data set

2.2 Feature selection

Feature selection is performed with ^{2} (Chi) statistics, mutual information (MI) and hybrid. Given a topic, the corresponding documents are considered as relevance while documents in other 49 topics are irrelevant. Threshold in ^{2} statistics is 3.841 and MI threshold is 0.1. The results of hybrid are feature intersection between ^{2} statistics and MI.

3. Relevant Documents and Sentences Retrieval

Sentences are treated as similar as documents during the modeling. We have tried three approaches: vector space model (VSM) and its variations, probability approach used in OKAPI system [4] and statistical language model (SLM).

3.1 VSM and Variations

Query, documents and sentences are represented with vector in VSM. Each term is weighted with log(tf+1)* log(N/df+1.0), where tf is term frequency in a document/sentence and df is the document or sentence frequency. Traditionally, relevance or similarity is estimated with the following formula:

$$sim(q,d) = \sum_{t \in q \wedge d} w_{d,t} * w_{q,t}$$

In vector normalization, we have tried cosine, length, and pivoted document length normalization [5]. We took three query expansion (QE) strategies: pseudo-relevance feedback, WordNet and local co-occurrence. In pseudo-relevance feedback QE, sentences/documents, whose similarity with query ranked top x (5%, 10% and so on), were treated as “relevance” and were used to modify query as feedback. Experiments, which were conducted on Novelty 2003, indicated that such strategy could improve F-measure by 0.19. WordNet is used to extend a term with words both in its synset/synonym and given documents. We have tried similarity computation using WordNet in TREC 2003 [6]. It is interesting but not very useful in novelty detection. Further experiments showed that the synset expansion could be nearly ignored. And we also extend terms with highly co-occurrence other terms (over 3 times in our experiments) in relevant documents. Experiments on last track show that VSM with local co-occurrence expansion could achieve 0.643 in relevance F-measure.

3.2 OKAPI approach

We used model in OKAPI system, which estimate the similarity between query *q* and document/sentence *d* with the following formulas:

$$sim(q,d) = \sum_{t \in q \wedge d} w_{d,t} * w_{q,t}$$

$$w_{d,t} = \frac{(k_1+1) * f_{d,t}}{k_1 * [(1-b) + b * \frac{W_d}{avr\_W_d}] + f_{d,t}} \qquad W_{q,t} = \frac{(k_2+1) * f_{q,t}}{k_2 + f_{q,t}} * \log \frac{N - f_t + 0.5}{0.5}$$

In this year of Novelty, k_1=1.2, k_2=1000, b=0.75, Avr_Wd=average document length, ft=number of documents which term t occurs, and fx,t=term t frequency in either q or d

3.3 Statistical Language Model

In statistical language model, similarity between a sentence *S* and topic *T* is computed by the logarithm of conditional probability within language distribution model. The formula is:

$$Sim(T,S) = \ln P(T \mid S) = \sum_{w \in T} \ln[\lambda p(w \mid S) + (1-\lambda) p(w \mid D)]$$

where *D* is the document what *S* belongs to and λ is the smoothing argument between sentence and document. Actually SLM is not applicable in sentence level in that the size is too small to get a reliable language probability. We eventually discard the approach in the official runs.

4. Novelty Detection

Three types of novelty degree are brought up in our work. They are: word overlap and its variation, similarity comparison and information gain.

4.1 Word Overlap and Variation

Word overlap is simple but and useful metric on novelty degree. The word overlap and sentence novelty degree is defined as:

$$Overlap(S_i \leftarrow S_j) = \frac{|S_i \cap S_j|}{|S_i|}$$

$$OverlapNov(S_i) = 1 - \max\{Overlap(S_i \leftarrow S_j)\} where, j < i$$

Here, a sentence is treated as a word set. However, different words in the sentence are equally weighted. It is far from practice. We brought a variation based on ^{2} weighting, that is:

$$Overlap^*(S_i \leftarrow S_j) = \frac{\sum_{w \in S_i \cap S_j} chi_w}{\sum_{w \in S_i} chi_w}$$

Where chi_w is the ^{2} weight of word *w*. Actually, the formula is a general form of the former.

4.2 Similarity Comparison

Given a sentence S_i, the novelty degree is estimated with the similarity comparison between topic and previous sentences. Using different strategy, we could get three variations as follows:

$$Max\Pr evNov(S_i) = \lambda Sim(S_i, T) - (1-\lambda)\max_{k<i \wedge S_k \in R} Sim(S_i, S_k)$$
$$Aver\Pr evNov(S_i) = \lambda Sim(S_i, T) - (1-\lambda) AVG_{k<i \wedge S_k \in R} Sim(S_i, S_k)$$

$$\Pr evAverNov(S_i) = \lambda Sim(S_i, T) - (1-\lambda) Sim(S_i, \overline{SP})$$
$$0 \le \lambda \le 1$$

4.3 Information Gain Degree

Word overlap and similarity comparison both evaluate the novelty degree on a given single sentence. However, a sentence is not as easy to compute as a sentence cluster. From the other view, novelty degree could be estimated with information gain degree after adding the given sentence.

Based on relevance degree, conditional probability and information entropy, we could get variations as follows:

$$\mathrm{IGN}ov(S_i) = \operatorname{Re} l(T, \Pr evSentences + S_i) - \operatorname{Re} l(T, \Pr evSentences)$$
$$\mathrm{IGN}ov'(S_i) = P(T \mid \Pr evSentences + S_i) - P(T \mid \Pr evSentences)$$
$$\begin{aligned}\mathrm{IGN}ov''(S_i) = {} & P(T \mid \Pr evSentences)\log P(T \mid \Pr evSentences) \\ & - P(T \mid \Pr evSentences + S_i)\log P(T \mid \Pr evSentences + S_i)\end{aligned}$$

5. Official Runs

Our group participate all tasks this year. We would report our 19 runs by tasks.

5.1 Task1

The official results are given in Table1. OKAPI approach achieved the best performance in all runs. According to previous works, OKAPI is very promising in document retrieval. In this task, relevant document results are the basis of sentence relevance and novelty. Lower precision in document retrieval would greatly decrease the final performance regardless of any further effort in sentence retrieval and novelty detection.

Table 1. Official runs in Task 1

RunID	Rel P/R/F	New P/R/F	Description
ICTOKAPIOVLP	0.32/0.73/0.415	0.17/0.57/0.239	OKAPI+OverlapNov
ICTVSMLCE	0.29/0.73/0.392	0.12/0.71/0.199	VSM, LCE+ AverPrevNov
ICTVSMFDBKL	0.28/0.77/0.385	0.13/0.71/0.202	VSM, light feedback +AverPrevNov
ICTVSMFDBKH	0.28/0.77/0.389	0.12/0.77/0.198	VSM, heavy feedback+ AverPrevNov
ICTVSMCOSAP	0.31/0.68/0.397	0.08/0.47/0.130	VSM Cosine, AverPrevNov

(Rel P/R/F refers to be average relevance precision, recall and F-measure, respectively. New P/R/F refers to be average novelty precision, recall and F-measure, respectively. Similar in other tables)

5.2 Task2

Given all relevant sentences, word overlap novelty could get 0.599 F-measure. The simple approach performed very well. From table 2, we could also conclude that information gain is competitive.

Table 2. Official runs in Task 2

RunID	New P/R/F	Description
ICT2VSMLCE	0.43/0.89/0.559	VSM,LCE+IGNov
ICT2VSMIG95	0.42/0.77/0.523	VSM+IGNov (top 95%)

ICT2OKAPIAP	0.42/0.81/0.534	OKAPI+AverPrevNov
ICT2OKALCEAP	0.42/0.83/0.539	OKAPI+AverPrevNov
ICT2VSMOLP	0.47/0.89/0.599	VSM, Cosine+OverlapNov

5.3 Task3

Given relevant and new sentences in the first 5 documents, we dynamically adjust parameters and threshold according to the known relevant/all and relevant/new proportion. And further feature selection was made in the first 5 documents. The last two official runs in table 3 have achieved first rank in 11 topics and second rank in 6 topics. OKAPI is also proved very well in relevance retrieval both in documents and sentences.

Table 3. Official runs in Task 3

RunID	Rel P/R/F	New P/R/F	Description
ICT3OKAPFDBK	0.38/0.68/0.441	0.12/0.46/0.212	OKAPI, FDBK+ IGNov
ICT3OKAPIIG	0.37/0.72/0.459	0.15/0.50/0.216	OKAPI +IGNov
ICT3OKAPIOLP	0.37/0.76/0.464	0.15/0.52/0.217	OKAPI+ OverlapNov
ICT3VSMOLP	0.370.76/0.464	0.14/0.63/0.213	VSM+ OverlapNov

5.4 Task4

In this task, two runs with information gain degree achieve better than other approaches. Information gain degree is promising in novelty modeling.

Table 4. Official runs in Task 4

RunID	New P/R/F	Description
ICT4OVLPCHI	0.41/0.63/0.469	OverlapNov, Chi feature selection
ICT4OVERLAP	0.41/0.65/0.476	OverlapNov considering weight
ICT4OKAPIIG	0.37/0.85/0.496	OKAPI+IGNov
ICT4OKAAP	0.37/0.73/0.464	OKAPI +AverPrevNov
ICT4IG	0.37/0.88/0.504	OKAPI+IGNov

References

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

[2] E. Brill. Some advances in rule-based part of speech tagging. In *Proceedings of the Twelfth National Conference on Artificial Intelligence (AAAI-94)*, Seattle, Wa., 1994.

[3] Porter, M.F., *An Algorithm For Suffix Stripping*, Program 14 (3), July 1980, pp. 130-137.

[4] Robertson, S.E., Walker, S., Beaulieu, M.M., Gatford, M., Payne, *A. Okapi at TREC-4, The Fourth Text REtrieval Conference (TREC-4)*, NIST Special Publication 500-236, National Institute of Standards and Technology, Gaithersburg, MD, pp. 73-86, October 1996.

[5] Singhal A., C. Buckley, and M. Mitra. *Pivoted Document Length Normalization*. In H. Frei, D. Harman, P. Schauble, and R. Wilkinson, editors, Proceedings of the Nineteenth Annual International A CM SIGIR Conference on Research and Development in Information Retrieval, 1996.

[6] Hua-Ping ZHANG, Jian SUN, Bing WANG, Shuo BAI. *Computation on Sentence Semantic Distance for Novelty Detection,* Journal of Computer and Science Technology, No.6 2004 (to be published).