Chinese Novelty Mining

Yi Zhang
Nanyang Technological University
50 Nanyang Avenue
Singapore 639798
yizhang@ntu.edu.sg

Flora S. Tsai
Nanyang Technological University
50 Nanyang Avenue
Singapore 639798
fst1@columbia.edu

Abstract

Automated mining of novel documents or sentences from chronologically ordered documents or sentences is an open challenge in text mining. In this paper, we describe the preprocessing techniques for detecting novel Chinese text and discuss the influence of different Part of Speech (POS) filtering rules on the detection performance. Experimental results on AP-WSJ and TREC 2004 Novelty Track data show that the Chinese novelty mining performance is quite different when choosing two dissimilar POS filtering rules. Thus, the selection of words to represent Chinese text is of vital importance to the success of the Chinese novelty mining. Moreover, we compare the Chinese novelty mining performance with that of English and investigate the impact of preprocessing steps on detecting novel Chinese text, which will be very helpful for developing a Chinese novelty mining system.

1 Introduction

The bloom of information nowadays brings us rich useful information as well as tons of redundant information in news articles, social networks (Tsai et al., 2009), and blogs (Chen et al., 2008). Novelty mining (NM), or novelty detection, aims at mining novel information from a chronologically ordered list of relevant documents/sentences. It can facilitate users to quickly get useful information without going through a lot of redundant information, which is usually a tedious and time-consuming task.

The process of detecting novel text contains three main steps, (i) preprocessing, (ii) categorization, and (iii) novelty mining. The first step preprocesses the text documents/sentences by removing stop words, performing word stemming, implementing POS tagging etc. Categorization classifies each incoming document/sentence into its relevant topic bin. Then, within each topic bin containing a group of relevant documents/sentences, novelty mining searches through the time sequence of documents/sentences and retrieves only those with "novel" information. This paper focuses on applying document/sentence-level novelty mining on Chinese. In this task, we need to identify all novel Chinese text given groups of relevant documents/sentences.

Novelty mining has been performed at three different levels: event level, sentence level and document level (Li and Croft, 2005). Works on novelty mining at the event level originated from research on Topic Detection and Tracking (TDT), which is concerned with online new event detection/first story detection (Allan et al., 1998; Yang et al., 2002; Stokes and Carthy, 2001; Franz et al., 2001; Brants et al., 2003). Research on document and sentence-level novelty mining aims to find relevant and novel documents/sentences given a stream of documents/sentences. Previous studies on document and sentence-level novelty mining tend to apply some promising content-oriented techniques (Li and Croft, 2005; Allan et al., 1998; Yang et al., 1998; Zhang and Tsai, 2009). Similarity metrics that can be used for detecting novel text are word overlap, cosine similarity (Yang et al., 1998), new word count (Brants et al., 2003), etc. Other works utilize ontological knowledge, especially taxonomy, such as WordNet (Zhang et al., 2002; Allan et al., 2003), synonym dictionary (Franz et al., 2001), HowNet (Eichmann and Srinivasan, 2002), etc.

Previous studies for novelty mining have been conducted on the English and Malay languages (Kwee et al., 2009; Tang et al., 2009; Tang and Tsai, 2009). Novelty mining studies on the Chinese language have been performed on topic de-

Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, pages 1561–1570,
Singapore, 6-7 August 2009.

tection and tracking, which identifies and collects relevant stories on certain topics from information stream (Zheng et al., 2008; Hong et al., 2008). Also many works have discussed the issues, such as word segmentation, POS tagging etc, between English and Chinese (Wang et al., 2006; Wu et al., 2003). However, to the best of our knowledge, no studies have been reported on discussing preprocessing techniques on Chinese document and sentence-level novelty mining, which is the focus of our paper.

The rest of this paper is organized as follows. Section 2 gives a brief overview of related work on detecting novel documents and sentences on English and Chinese. Section 3 introduces the details of preprocessing steps for English and Chinese. A general novelty mining algorithm is described in Section 4. Section 5 reports experimental results. Section 6 summarizes the research findings and discusses issues for further research.

2 Related Work

In the pioneering work for detecting novel documents (Zhang et al., 2002), document novelty was predicted based on the distance between the new document and the previously delivered documents in history. The detected document which is very similar to any of its history documents is regarded as a redundant document. To serve users better, it could be more helpful to further highlight novel information at the sentence level. Therefore, later studies focused on detecting novel sentences, such as those reported in TREC 2002-2004 Novelty Tracks (Harman, 2002; Soboroff and Harman, 2003; Soboroff, 2004), which compared various novelty metrics (Allan et al., 2003), and integrated different natural language techniques (Ng et al., 2007; Li and Croft, 2008).

Although novelty mining studies have mainly been conducted on the English language, studies on the Chinese language have been performed on topic detection and tracking. A prior study (Zheng et al., 2008) proposed an improved relevance model to detect the novelty information in topic tracking feedback and modified the topic model based on this information. Experimental results on Chinese datasets TDT4 and TDT2003 proved the effectiveness in topic tracking. Another study proposed a method of applying semantic domain language model to link detection, based on the structure relation among contents and the semantic distribution in a story (Hong et al., 2008).

3 Preprocessing for English and Chinese

3.1 English

Since the focus of this paper is on novelty mining, we begin from a list of relevant documents or sentences that have already undergone the categorization process.

The first step for English preprocessing is to remove all stop words from documents or sentences, such as conjunctions, prepositions, and articles. Stop words are words that are too common to be informative. These words should be removed, otherwise it will influence the novelty prediction of documents or sentences. After stop words removal, the remaining words are then stemmed. The inflected (or sometimes derived) words are reduced to their root forms. This paper used Porter stemming algorithm (Porter, 1997) for English word stemming. This algorithm removes the commoner morphological and inflexional endings from the words in English. The entire preprocessing steps in English novelty mining can be seen in Figure 1.

3.2 Chinese

In Chinese, the word is the smallest independent meaningful element. There is no obvious boundary between words so that Chinese lexical analysis, such as Chinese word segmentation, is the prerequisite for novelty mining.

Unlike English, Chinese word segmentation is a very challenging problem because of the difficulties in defining what constitutes a word (Gao et al., 2005). While each criteria provides valuable insights into "word-hood" in Chinese, they do not consistently lead us to the same conclusions. Moreover, there is no white space between Chinese words or expressions and there are many ambiguities in the Chinese language, such as: '主板和服务器' (means 'mainboard and server' in English) might be '主板/和/服务器' (means 'mainboard/and/server' in English) or '主板/和服/务/器' (means 'mainboard/kimono/task/utensil' in English). This ambiguity is a great challenge for Chinese word segmentation. In addition, there is no obvious inflected or derived words in Chinese so that word stemming is not applicable.

Therefore, in order to reduce the noise brought by Chinese word segmentation and get a better

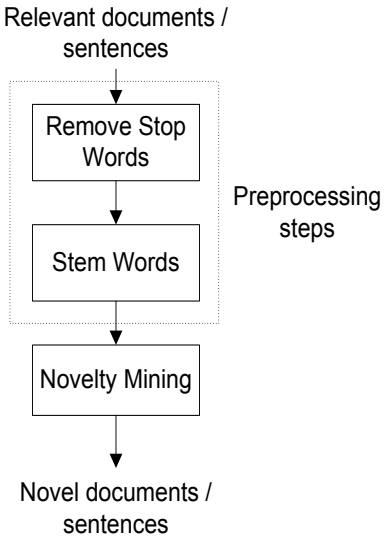

Figure 1: Preprocessing steps on English.

word list for one document or sentence, we firstly apply word segmentation on the Chinese text and then utilize Part-of-Speech (POS) tagging to select the meaningful candidate words. Figure 2 shows the preprocessing steps on the Chinese text for novelty mining. POS tagging is a process of marking up the word in a text as corresponding to a particular part of speech. It is learnt that the idea of a text mainly relies on some meaningful words, such as nouns and verbs, so that we can get the main content by extracting these meaningful words. Moreover, it will decrease the impact of the errors in Chinese word segmentation on novelty mining because only meaningful words are considered and other words (including stop words) such as '虽然' (means 'although' in English) will not appear in the word list for the following similarity computation in novelty mining. Losee also mentioned that POS tagging shows a great potential to avoid lexical ambiguity and it can help to improve the performance of information retrieval (Losee, 2001).

ICTCLAS is used when performing word segmentation and POS tagging in our experiments (ICTCLAS, 2008). It is an open source project and achieves a better precision in Chinese word segmentation and POS tagging than other Chinese POS tagging softwares (ICTCLAS, 2008). First, we apply word segmentation on the relevant Chinese documents/sentences. Chinese word segmentation includes atom segmentation, N-shortest path based rough segmentation and unknown words recognition (see Figure 3). Atom segmentation is an initial step of the Chinese language segmentation process, where atom is defined to be the minimal unit that cannot be split further. The atom can be a Chinese character, punctuation, symbol string, etc. Then, rough segmentation tries to discover the correct segmentation with as few candidates as possible. The N-Shortest Path (NSP) method (Zhang and Liu, 2002) is applied for rough segmentation. Next, we detect some unknown words such as person name, location name so as to optimize the segmentation result. Finally, we POS tag the words and keep some kinds of words in the word list according to the selective rule, which are used in novelty mining.

4 Novelty Mining

From the output of preprocessing, we can obtain a bag of words. The corresponding term-document matrix (TDM)/term-sentence matrix (TSM) can be constructed by counting the term frequency (TF) of each word. The novelty mining system predicts any incoming document/sentence by comparing it with its history documents/sentences in this vector space. Therefore, given a Chinese TDM/TSM, the novelty mining system designed for English can also be applied to Chinese.

In novelty mining, the novelty of a document/sentence can be quantitatively measured by a novelty metric and represented by a novelty score. The most popular novelty metric, i.e. cosine similarity (see (Allan et al., 2003)), is adopted. This metric first calculates the similarities between the current document/sentence d_t and each of its his-

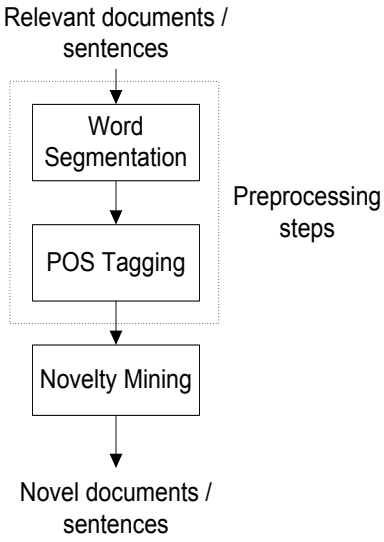

Figure 2: Preprocessing steps on Chinese.

tory documents/sentences d_i ($1 \leq i \leq t-1$). Then, the novelty score is simply one minus the maximum of these cosine similarities, as shown in Eq.(1).

$$Novelty\ Score(d_t) = 1 - \max_{1 \leq i \leq t-1} cos(d_t, d_i) \quad (1)$$

$$cos(d_t, d_i) = \frac{\sum_{k=1}^{n} w_k(d_t) \cdot w_k(d_i)}{\|d_t\| \cdot \|d_i\|}$$

where $N_{cos}(d)$ denotes the cosine similarity score of the document/sentence d and $w_k(d)$ is the weight of k^{th} element in the document/sentence weighted vector d. The term weighting function used in our work is TF(term frequency).

The final decision on whether a document/sentence is novel or not depends on whether the novelty score falls above or below a threshold. The document/sentence predicted as "novel" will be placed into the list of history documents/sentences.

5 Experiments and Results

5.1 Datasets

Two public datasets APWSJ (Zhang et al., 2002) and TREC Novelty Track 2004 (Soboroff, 2004) are selected as our experimental datasets for the document-level and the sentence-level novelty mining respectively. APWSJ data consists of news articles from Associated Press (AP) and Wall Street Journal (WSJ). There are 50 topics from Q101 to Q150 in APWSJ and 5 topics (Q131, Q142, Q145, Q147, Q150) are excluded from the experiments because they lack non-novel documents (Zhao et al., 2006). The assessors provide two degrees of judgements on non-novel documents, *absolute redundant* and *somewhat redundant*. In our experiments, we adopt the strict definition used in (Zhang et al., 2002) where only *absolute redundant* documents are regarded as non-novel. TREC 2004 Novelty Track data is developed from AQUAINT collection. Both relevant and novel sentences are selected by TREC's assessors. The statistics of these two datasets are summarized in Table 1.

Table 1: Statistics of experimental data

Dataset	Novel	Non-novel
APWSJ	10839(91.10%)	1057(8.90%)
TREC2004	3454(41.40%)	4889(58.60%)

5.2 Evaluation Measures

From many previous works, redundancy precision (RP), redundancy recall (RR) and redundancy F Score (RF) are used to evaluate the performance of document-level novelty mining (Zhang et al., 2002). Precision (P), recall (R) and F Score (F) are mainly used in evaluating the performance for sentence-level novelty mining (Allan et al., 2003). Therefore, we use RP, RR, RF and redundancy precision-recall (R-PR) curve to evaluate our experimental results on the document level. P, R, F and precision-recall (PR) curve are used to evaluate the performance on the sentence-level novelty mining. The larger the area under the R-PR

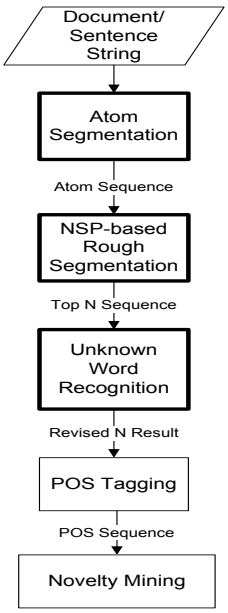

Figure 3: Word segmentation on Chinese.

curve/PR curve, the better the algorithm. Also we drew the standard redundancy F Score/F Score contours (Soboroff, 2004), which indicate the F Score values when setting precision and recall from 0 to 1 with a step of 0.1. These contours can facilitate us to compare redundancy F Scores/F Scores in R-PR curves/PR curves. Redundancy precision, redundancy recall, precision and recall on a certain topic are defined as:

$$Redundancy\ Precision = \frac{R^-}{R^- + N^-} \quad (2)$$

$$Redundancy\ Recall = \frac{R^-}{R^- + R^+} \quad (3)$$

$$Precision = \frac{N^+}{N^+ + R^+} \quad (4)$$

$$Recall = \frac{N^+}{N^+ + N^-} \quad (5)$$

where R^+,R^-,N^+,N^- correspond to the number of documents/sentences that fall into each category (see Table 2).

Based on all the topics' RP/P and RR/R, we could get the average RP/P and average RR/R by calculating the arithmetic mean of these scores on all topics. Then, the average redundancy F Score (RF)/F Score (F) is obtained by the harmonic average of the average RP/P and average RR/R.

Table 2: Categories for evaluation

	Non-novel	Novel
Delivered	R^+	N^+
Not Delivered	R^-	N^-

5.3 Experimental Results

In this experimental study, the focus was novelty mining rather than relevant documents/sentences categorization. Therefore, our experiments started with all given relevant Chinese text, from which the novel text should be identified.

Since the datasets that we used for document-level novelty mining and sentence-level novelty mining both were written in English, we first translated them into Chinese. During this process, we investigated issues on machine translation vs. manually corrected translation.

We compared the novelty mining performance on 107 text in TREC 2004 Novelty Track between automatically translated using Google Translate API[1] and the manually corrected translation. For example, here is an English sentence in Topic 51:

According to a Chilean government report, a total of 4,299 political opponents died or disappeared during Pinochet's term.

After machine translation using Google Translator, the above sentence is translated as:

根据智利政府的报告，共有4299政敌的死亡或失踪期间，皮诺切特的任期。

[1]http://code.google.com/p/google-api-translate-java

Then we manually corrected the machine translation and obtained the corrected translation:

根据智利政府的报告，在皮诺切特的任期期间，共有4299政敌死亡或失踪。

After novelty mining on the machine translation sentences and the humanly corrected translation sentences individually, we found that there is a slight difference (<2%) in precision and F Score. Thus, we used machine translation to translate the remaining documents/sentences to Chinese. This indicates that the noise in machine translation for Chinese had little impact on our actual results.

Then on English text, we applied the preprocessing steps discussed in Section 3.1, including stop word removing and word stemming. For Chinese datasets, we segmented the documents/sentences into words and then performed POS filtering to acquire the candidate words for the space vector.

Based on the vectors of Chinese text, we calculated the similarities between documents/sentences and predicted the novelty for each document/sentence in the Chinese and English datasets. An incoming Chinese/English document will be compared with all the system delivered 10 novel documents. If the novelty score is above the novelty score threshold, the document is considered to be novel. Thresholds used were between 0.05 and 0.65. We also performed Chinese/English sentence-level novelty mining. Whether an incoming Chinese/English sentence is novel is predicted by comparing with the most recent system-delivered 1000 novel sentences. Thresholds adopted were between 0.05 and 0.95 with an equal step of 0.10. Then, we evaluated the Chinese/English novel text detection performance by setting a series of novelty score thresholds.

5.3.1 POS Filtering Rule

We adopted two different rules to select the candidate words to represent one document/sentence and investigated the POS filtering influence on detecting the novel Chinese text.

- Rule1: only some non-meaningful words, including pronouns ('r' in Peking University/Chinese Academy of Sciences Chinese POS tagging criterions (PKU and CAS, 1999)), auxiliary words ('u'), tone words ('y'), conjunctions ('c'), prepositions ('p') and punctuation words ('w') are removed.
- Rule2: fewer kinds of words are selected to represent a document/sentence. Only nouns (including 'n' short for common nouns, 'nr' short for person name, 'ns' short for location name, 'nt' short for organization name, 'nz' short for other proper nouns), verbs ('v'), adjectives ('a') and adverbs ('d') are kept.

For example, here is a simple Chinese sentence: "墙上挂着一幅画。" (There is a picture on the wall). After POS filtering using Rule1, the words we keep are: "墙('n'), 上('v'), 挂('v'), 一('m'), 幅('q'), 画('n')". After POS filtering using Rule2, the remaining words are: "墙('n'), 上('v'), 挂('v'), 画('n')". It is noticed that by using Rule2, we can remove more non-important words.

Figure 4 and Figure 5 show the performances on the document and sentence-level novelty mining when choosing the stricter rule (Rule2) and the less strict rule (Rule1) in POS filtering. The grey dashed lines show contours at intervals of 0.1 points of F Score.

From Figure 4 and Figure 5, we learn that the Chinese novelty mining performance varies when choosing the stricter rule (Rule2) and the less strict rule (Rule1) in POS filtering. We can obtain a better performance when choosing a stricter rule (Rule2). Therefore, it is necessary to perform POS filtering in the preprocessing steps on Chinese and just removing some non-meaningful words (like stop words) may not be enough. POS filtering can help to remove the less meaningful words so that each vector is represented better. Compared to choosing more kinds of words (Rule1), only keeping nouns, verbs, adjectives and adverbs (Rule2) will be a better choice for novelty mining. We also noticed that the selection of words to represent Chinese text is of vital importance to the success of Chinese novelty mining.

5.3.2 Comparison with English

We compared the novelty mining performance on the English and Chinese documents/sentences datasets. For Chinese, we chose Rule2 to select the candidate words. Figure 6 and Figure 7 show the R-PR and PR curves of document/sentence-level novelty mining in English and Chinese when given a series of novelty score thresholds.

From Figure 6 and Figure 7, we observe that the performance on detecting novel Chinese documents is slightly lower than that on English. This may be due to the different linguistical characteris-

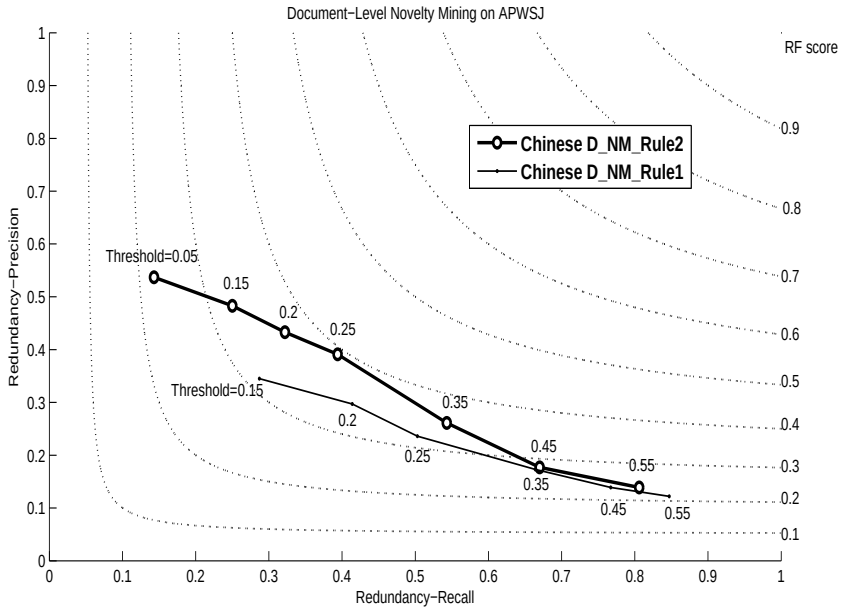

Figure 4: R-PR curves for document-level novelty mining on Chinese when choosing different rules on APWSJ. The grey dashed lines show contours at intervals of 0.1 points of RF.

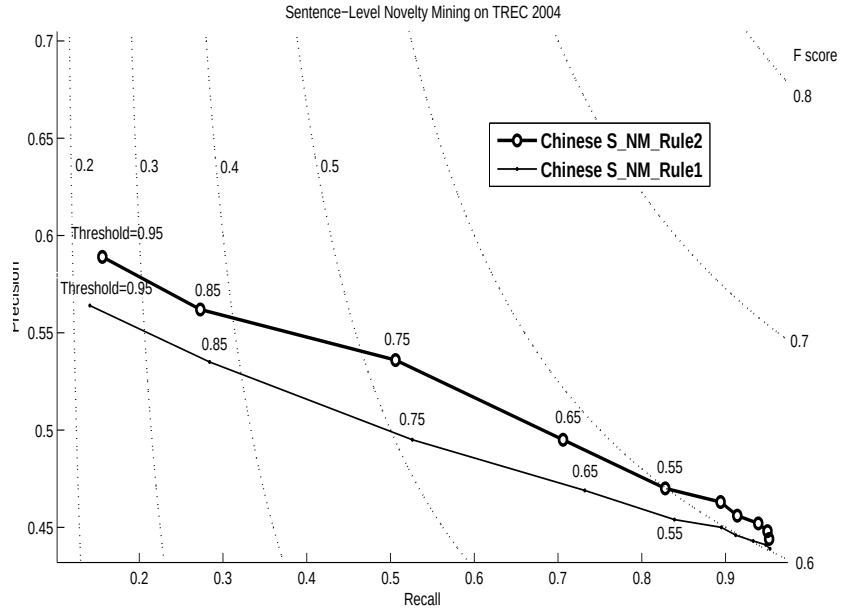

Figure 5: PR curves for sentence-level novelty mining on Chinese when choosing different rules on TREC 2004. The grey dashed lines show contours at intervals of 0.1 points of F.

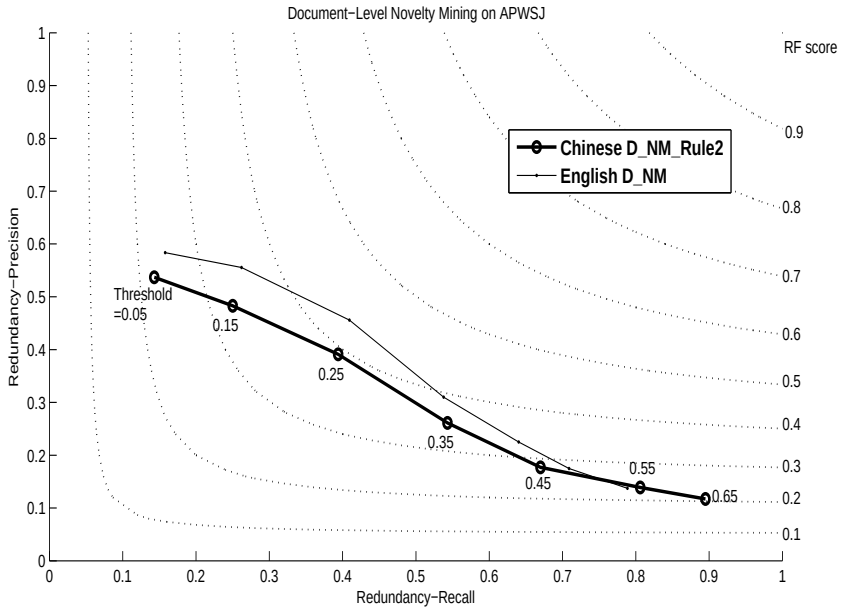

Figure 6: R-PR curves for document-level novelty mining on Chinese and English on APWSJ. The grey dashed lines show contours at intervals of 0.1 points of RF.

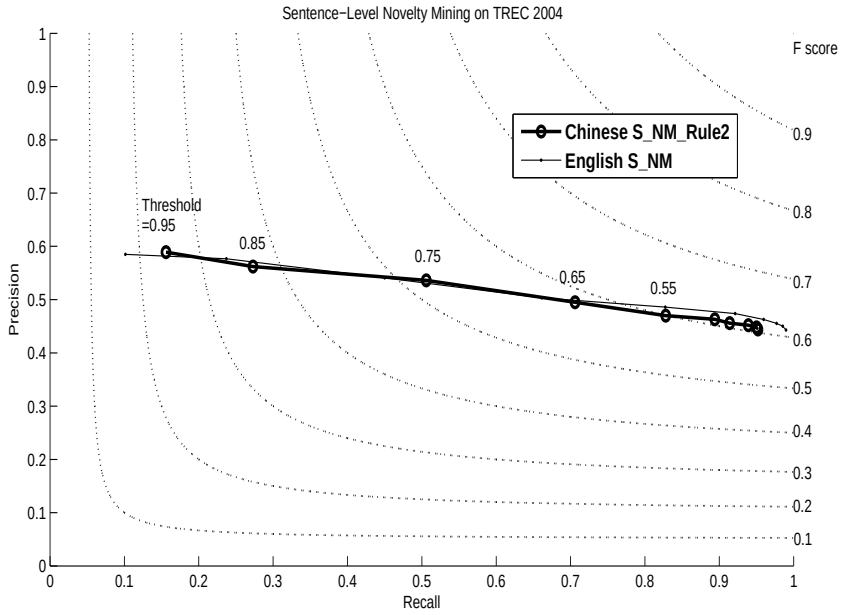

Figure 7: PR curves for sentence-level novelty mining on Chinese and English on TREC 2004. The grey dashed lines show contours at intervals of 0.1 points of F.

tics of each language so that the preprocessing influence on each language's novelty mining is dissimilar. Furthermore, the Chinese preprocessing quality is not as good as that on English so that it is difficult to obtain a good "bag of words" from a document. Moreover, the errors in word segmentation will influence the result of POS tagging. These issues make tokenizing and POS tagging extremely difficult for the Chinese text.

However, the performance of Chinese sentence-level novelty mining is almost the same as that on English. The reason is that the novelty mining performance at the sentence level is not so sensitive to the preprocessing steps as that at the document level. If the similarity computation is based on the sentence level, the word segmentation and POS tagging errors actually will not have a big influence on the result as that on documents.

6 Conclusion

This paper studied the preprocessing issues on mining novel Chinese text, which, to the best of our knowledge, have not been sufficiently addressed in previous studies. We described the Chinese preprocessing steps and discussed the influence when choosing different Part-of-Speech (POS) filtering rules. Then we applied novelty mining on Chinese and English documents/sentences and compared their performance.

The experimental results on APWSJ and TREC 2004 Novelty Track showed that after adopting a stricter POS filtering rule, the Chinese novelty mining performed better on both documents and sentences. This is because non-meaningful words have a negative influence on detecting novel text. However, compared to English, Chinese performed worse on the document level and similarly on the sentence level. The reason may be due to the lower sensitivity of preprocessing at the sentence level. The main contributions of this work are as follows:

1) We investigated the preprocessing techniques for detecting novel Chinese text on both document and sentence level.

2) The POS filtering rule, telling how to select words to represent one document/sentence, was discussed.

3) Several experiments were conducted to compare the novelty mining performance between Chinese and English. The novelty mining performance on Chinese can be improved as good as that on English if we can increase the preprocessing precision on Chinese text.

Our findings will be very helpful for developing a real-time Chinese novelty mining system at both the document and sentence level. In future work, we will try other word combinations and investigate better ways to represent the Chinese text. In addition, we will explore how to utilize the better Chinese sentence-level novelty mining result to improve the detection performance on documents.

References

James Allan, Ron Papka, and Victor Lavrenko. 1998. Online new event detection and tracking. In *SIGIR 1998, Melbourne, Australia*, pages 37–45.

James Allan, Courtney Wade, and Alvaro Bolivar. 2003. Retrieval and novelty detection at the sentence level. In *SIGIR 2003, Toronto, Canada*, pages 314–321. ACM, August.

Thorsten Brants, Francine Chen, and Ayman Farahat. 2003. A system for new event detection. In *SIGIR 2003, Toronto, Canada*, pages 330–337.

Yun Chen, Flora S. Tsai, and Kap Luk Chan. 2008. Machine learning techniques for business blog search and mining. *Expert Syst. Appl.*, 35(3):581–590.

D. Eichmann and P. Srinivasan. 2002. Novel results and some answers. In *TREC 2002 - the 11th Text REtrieval Conference*.

Martin Franz, Abraham Ittycheriah, J.Scott McCarley, and Todd Ward. 2001. First story detection: combining similarity and novelty based approach. In *Topic Detection and Tracking Workshop*.

Jianfeng Gao, Mu Li, Andi Wu, and Chang-Ning Huang. 2005. Chinese word segmentation and named entity recognition: A pragmatic approach. *Computational Linguistics*, 31(4):531–574, December.

D. Harman. 2002. Overview of the TREC 2002 Novelty Track. In *TREC 2002 - the 11th Text Retrieval Conference*, pages 46–55.

Yu Hong, Yu Zhang, Jili Fan, Ting Liu, and Sheng Li. 2008. Chinese topic link detection based on semantic domain language model. *Journal of Software*, 19(9):2265–2275.

ICTCLAS. 2008. http://ictclas.org/index.html.

Agus Trisnajaya Kwee, Flora S Tsai, and Wenyin Tang. 2009. Sentence-level novelty detection in English and Malay. In *Lecture Notes in Computer Science (LNCS)*, volume 5476, pages 40–51.

Xiaoyong Li and W. Bruce Croft. 2005. Novelty detection based on sentence level patterns. In *CIKM 2005*, pages 744–751.

Xiaoyong Li and W. Bruce Croft. 2008. An information-pattern-based approach to novelty detection. *Information Processing and Management: an International Journal*, 44(3):1159–1188, May.

Robert M. Losee. 2001. Natural language processing in support of decision making: Phrases and part-of-speech tagging. *Information Processing and Management: an International Journal*, 37(6).

Kok Wah Ng, Flora S. Tsai, Kiat Chong Goh, and Lihui Chen. 2007. Novelty detection for text documents using named entity recognition. In *Information, Communications and Signal Processing, 2007 6th International Conference on*, pages 1–5, December.

PKU and CAS. 1999. Chinese POS tagging criterion. http://icl.pku.edu.cn/icl_groups/corpus/addition.htm.

M.F. Porter. 1997. An algorithm for suffix stripping. *Readings in information retrieval*, pages 313–316.

Ian Soboroff and D. Harman. 2003. Overview of the TREC 2003 Novelty Track. In *TREC 2003 - the 12th Text Retrieval Conference*.

Ian Soboroff. 2004. Overview of the TREC 2004 Novelty Track. In *TREC 2004 - the 13th Text Retrieval Conference*.

N. Stokes and J. Carthy. 2001. First story detection using a composite document representation. In *HLT 2001*, pages 134–141.

Wenyin Tang and Flora S Tsai. 2009. Threshold setting and performance monitoring for novel text mining. In *SIAM International Conference on Data Mining Workshop on Text Mining*.

Wenyin Tang, Agus Trisnajaya Kwee, and Flora S Tsai. 2009. Accessing contextual information for interactive novelty detection. In *European Conference on Information Retrieval (ECIR) Workshop on Contextual Information Access, Seeking and Retrieval Evaluation*.

Flora S. Tsai, Wenchou Han, Junwei Xu, and Hock Chuan Chua. 2009. Design and development of a mobile peer-to-peer social networking application. *Expert Syst. Appl.*, 36(8):11077 – 11087.

Mengqiu Wang, Kenji Sagae, and Teruko Mitamura. 2006. A fast, accurate deterministic parser for Chinese. In *ACL 2006, Sydney, Australia*, pages 425 – 432.

Youzheng Wu, Jun Zhao, and Bo Xu. 2003. Chinese named entity recognition combining a statistical model with human knowledge. In *ACL 2003 workshop on Multilingual and mixed-language named entity recognition*, pages 65–72.

Yiming Yang, Tom Pierce, and Jaime Carbonell. 1998. A study on retrospective and on-line event detection. pages 28–36. ACM Press.

Yiming Yang, Jian Zhang, Jaime Carbonell, and Chun Jin. 2002. Topic-conditioned novelty detection. In *SIGKDD 2002*, pages 688 – 693.

Huaping Zhang and Qun Liu. 2002. Model of Chinese words rough segmentation based on n-shortest paths method. *Journal of Chinese Information Processing*, 15:1–7.

Yi Zhang and Flora S. Tsai. 2009. Combining named entities and tags for novel sentence detection. In *ESAIR '09: Proceedings of the WSDM '09 Workshop on Exploiting Semantic Annotations in Information Retrieval*, pages 30–34.

Yi Zhang, Jamie Callan, and Thomas Minka. 2002. Novelty and redundancy detection in adaptive filtering. In *ACM SIGIR 2002, Tampere, Finland*, pages 81–88.

Le Zhao, Min Zheng, and Shaoping Ma. 2006. The nature of novelty detection. *Information Retrieval*, 9:527–541.

Wei Zheng, Yu Zhang, Bowei Zou, Yu Hong, and Ting Liu. 2008. Research of Chinese topic tracking based on relevance model.