ISI Novelty Track System for TREC 2004

Soo-Min Kim **Deepak Ravichandran** **Eduard Hovy**

Information Science Institute
University of Southern California
4676 Admiralty Way
Marina del Rey, CA 90292-6695
{skim,ravichan,hovy}@isi.edu

Abstract

We describe our system developed at ISI for the Novelty track at TREC 2004. The system's two modules recognize relevant event and opinion sentences respectively. We focused mainly on recognizing relevant opinion sentences using various opinion-bearing word lists. Of our 5 runs submitted for task 1, the best run provided an F-score of 0.390 (precision 0.30 and recall 0.71).

1 Introduction

The Novelty Track is designed to investigate systems' abilities on two separate tasks: locating *relevant* or *new* information within a set of documents relevant to a TREC topic. Identifying relevant sentences is a sentence retrieval task similar to passage retrieval, where relevance is defined as both relating to a given topic and bearing an opinion about the topic. Identifying new sentences is defined as retrieving sentences about a given topic that contain information that has not appeared previously in this topic's set of documents. This year, opinion topics include "Gay Boy Scouts Banned", "Military Action Kosovo", and "Banning Tobacco Advertisements", while event topics include "India and Pakistan Nuclear Tests", "East Timor Independence", and "Payne Steward Plane Crash".

For both tasks, systems are given the topic and a set of relevant documents ordered by date, and must return only the appropriate sentences. Unlike last year's task, the initial documents set may include irrelevant documents, which systems must identify and ignore. We used a probabilistic Bayesian inference network model, as implemented in the search engine software package INQUERY (Callan et al. 1992), to identify the relevant documents.

There are four subtasks that vary the kinds of data available to systems and the kinds of results that must be returned. Among these four tasks, ISI participated in the first: identify all relevant and novel sentences given the full set of documents for the given topic.

We describe document filtering in Section 2 and our methods for identifying relevant sentences from opinion topic documents in Section 3. Section 4 explains the method we used for event topics. System results are reported in Section 5 and conclusions appear in Section 6.

2 Document Filtering

Each topic contained 25 relevant documents, possibly mixed with additional irrelevant documents. Thus, before proceeding to the next phase we had to separate relevant documents from irrelevant documents. We treat this problem as a standard Information Retrieval (IR) procedure. This approach is motivated by years of research in IR. It is a well known truism that simple techniques in IR often yield better results than sophisticated and deep linguistic analysis. We are interested in analyzing the results from other participants to empirically verify whether this belief holds in the novelty track document filtering task.

We use the example in Figure 1 to illustrate out document filtering process. To perform document filtering, we used the description <desc> field as our IR query. (We initially tried various other fields, including narrative <narr> and title <title>, but quickly abandoned them after noticing that the narrative field contains information which is very hard for computational treatment, such as negation.) To perform IR we use a probabilistic Bayesian inference network model as implemented in the search engine software package INQUERY. For each query we perform the standard procedure of stop-word removal and stemming. Using an OR query, we select the top 25 documents returned by the search engine as the relevant set.

<num> Number: N51
<title> General Pinochet Arrested

<toptype> Event

<desc> Description:
Arrest of former Chilean dictator, General Augusto Pinochet, in London. He was charged with murder, torture, genocide, and terrorism during his regime in Chile.

<narr> Narrative:
Information about Pinochet's arrest and evidence of charges of murdering, torturing and the disappearance of people in Chile while he was head of state is relevant. Specifically relevant are mention of charges against him.

Figure 1: Topic example

3 Opinion Topics

Identifying relevant sentences from opinion topic documents is different from identifying sentences from event topic documents. Relevant sentences from opinion documents should be relevant to the topic and opinion-bearing at the same time. Unlike event topics, we assume that whether a sentence is opinion-bearing or not is more important than its relevance to the topic in the opinion topic case, assuming that most opinions expressed a document are relevant to its topic. In other words, opinion documents such as editorials could contain sentences that describe irrelevant facts but the opinion sentences in these documents are likely relevant to the topic.

The most important part of opinion-bearing sentence recognition is identifying so-called subjectivity clues in a sentence. There are many approaches to building and using subjectivity clues. Turney (2002) and Wiebe (2000) focused on learning adjectives and adjectival phrases. Riloff et al. (2003) extracted nouns and Riloff and Wiebe (2003) extracted patterns for subjective expressions using a bootstrapping process.

We used unigrams as subjectivity clues and built four different systems to generate opinion-bearing word lists. After building these unigram lists, we checked each sentence in the relevant documents for the presence of opinion-bearing words. Sections 3.1 through 3.4 describe the four methods.

3.1 ISI Opinion-Bearing Word List

We developed a system to classify words as either opinion-bearing or non-opinion-bearing using WordNet and Wall Street Journal data.

3.1.1 Using WordNet

In pursuit of accuracy, we first manually collected a set of opinion-bearing words (34 adjectives and 44 verbs). Early classification trials showed that precision was very high (the system found only opinion-bearing sentences), but since the list was so small, recall was very low. We therefore used this list as seed words for expansion using WordNet. Our assumption was that synonyms and antonyms of an opinion-bearing word could be opinion-bearing as well, as for example "nice, virtuous, pleasing, well-behaved, gracious, honorable, righteous" as synonyms for "good", or "bad, evil, disreputable, unrighteous" as antonyms. However, not all synonyms and antonyms could be used: some such words seemed to exhibit both opinion-bearing and non-opinion-bearing senses, such as "solid, hot, full, ample" for "good". This indicated the need for a scale of opinion valence (*good* or *bad*) strength. If we can measure the 'opinion-based closeness' of a synonym or antonym to a known opinion-bearing word, then we can determine whether to include it in the expanded set. To develop such a scale, we first created a non-opinion-bearing word list manually and produced related words for it using WordNet. To avoid collecting uncommon words, we started with a basic/common English word list compiled for foreign students preparing for the TOEFL test. From this we randomly selected 462 adjectives and 502 verbs for human annotation. Human1 and human2 annotated 462 adjectives and human3 and human2 annotated 502 verbs, labeling each word as either opinion-bearing or non-opinion-bearing.

Now, to obtain a measure of opinion/non-opinion strength, we measured the WordNet distance of a target (synonym or antonym) word to the two sets of manually selected seed words plus their current expansion words. We assigned the new word to the closer category. The following equation represents this approach:

$$\begin{aligned} & \underset{c}{\operatorname{argmax}} P(c \mid w) \\ & \cong \underset{c}{\operatorname{argmax}} P(c \mid syn_1, syn_2 \ldots. syn_n) \qquad (1) \end{aligned}$$

where c is a category (opinion-bearing or non-opinion-bearing), w is the target word, and syn_n is the synonyms or antonyms of the given word by WordNet. To compute equation (1), we built a classification model, equation (2):

$$\begin{aligned}\arg\max_{c} P(c \mid w) &= \arg\max_{c} P(c)P(w \mid c) \\ &= \arg\max_{c} P(c)P(syn_1 syn_2 syn_{3 \ldots} syn_n \mid c) \\ &= \arg\max_{c} P(c)\prod_{k=1}^{m} P(f_k \mid c)^{count(f_k, synset(w))} \quad (2)\end{aligned}$$

where f_k is the k^{th} feature of category *c* which is also a member of the synonym set of the target word *w*, and *count(f_k,synset(w))* means the total number of occurrences of f_k in the synonym set of *w*. The motivation for this model is document classification. (Although we used the synonym set of seed words derived from WordNet, we could instead have obtained word features from a corpus.) After expansion, we obtained 2682 opinion-bearing and 2548 non-opinion-bearing adjectives, and 1329 opinion-bearing and 1760 non-opinion-bearing verbs, with strength values. Using these words as features we built a Naive Bayesian classifier and classified 32373 words.

3.1.2 Using WSJ Data

Experiments with the above set did not provide very satisfactory results on arbitrary text. For one reason, WordNet's synonym connections are simply not extensive enough. However, if we know the relative frequency of a word in opinion-bearing texts compared to non-opinion-bearing text, we can use the statistical information instead of lexical information. For this, we collected a huge amount of data in order to make up for the limitations of collection 1.

Following the insight of Yu and Hatzivassiloglou (2003), we made the basic and rough assumption that words that appear more often in newspaper editorials and letters to the editor than in non-editorial news articles could be potential opinion-bearing words (even though editorials contain sentences about factual events as well). We used the TREC collection to collect data, extracting and classifying all Wall Street Journal documents from it either as Editorial or nonEditorial based on the occurrence of the keywords "Letters to the Editor", "Letter to the Editor" or "Editorial" present in its headline. This produced in total 7053 editorial documents and 166025 non-editorial documents.

We separated out opinion from non-opinion words by considering their relative frequency in the two collections, expressed as a probability, using SRILM, SRI's language modeling toolkit (http://www.speech.sri.com/projects/srilm/). For every word W occurring in either of the document sets, we computed

$$Editorial\,\Pr ob(W) = \frac{\#\text{W in Editorial documents}}{\text{total words in Editorial documents}}$$

$$nonEditorial\,\Pr ob(W) = \frac{\#\text{W in nonEditorial docs}}{\text{total words in nonEditorial docs}}$$

We used Kneser-Ney smoothing (Kneser and Ney, 1995) to handle unknown/rare words. Having obtained the above probabilities we calculated the score of W as the following ratio:

$$Score(W) = \frac{\text{EditorialProb(W)}}{\text{nonEditorialProb(W)}}$$

Score(W) gives an indication of the bias of each word towards editorial or non-editorial texts. We computed scores for 86,674,738 word tokens.

Naturally, words with scores close to 1 were untrustworthy markers of opinion valence. To eliminate these words we applied a simple filter as follows. We divided the Editorial and the non-Editorial collections each into 3 subsets. For each word in each {Editorial, non-Editorial} subset pair we calculated Score(W). We retained only those words for which the scores in all three subset pairs were all greater than 1 or all less than 1. In other words, we only kept words with a repeated bias towards Editorial or non-Editorial. This procedure helped eliminate some of the noisy words, resulting in 15568 words. Table 1 shows the top 10 words with strong biases in each direction.

Editorial	Non-Editorial
M.D	+
Senator	Ltd.
Gigot	composite
rea	Volume
gerrymandering	NYSE
Crovitz	Holdings
gerrymander	Inc.
Quotable	surged
Manuela	redemption
Coordinator	Analysts

Table 1: Top 10 words of collection 2 for each category.

3.1.3 Merger of WordNet and WSJ Lists

So far, we have classified words as either opinion-bearing or non-opinion-bearing by two different methods. The first method calculates the degrees of closeness to manually chosen sets of opinion-bearing and non-opinion-bearing words in WordNet and decides its class and strength. When the word is equally close to both classes, it is hard to decide its subjectivity, and when WordNet does not contain a word or its synonyms, such as the word "antihomosexsual", we fail to classify it.

The second method, classification of words using WSJ texts, is less reliable than the lexical method. However, it does for example successfully handle "antihomosexual". Therefore, we combined the results of the two methods, since their different characteristics compensate for each other.

After merging two opinion-bearing word lists, we experimented with a cut-off parameter value to select only strong opinion bearing words. Finally, 10682 words were selected and applied for our run ISIRUN204.

3.2 Using the General Inquirer Dictionary

We created a third list of words similarly, by selecting positive and negative valence words from the *General Inquirer Dictionary*[1]. Any sentence that contains either positive or negative words was selected as a relevant opinion.

Initial results using this list were unsatisfactory since it contained only 1,915 positive and 2,291 negative words. This motivated us to apply the same method we used for Section 3.1.1, namely collecting synonyms and antonyms of positive and negative words from WordNet. The result was 6047 words.

Algorithm	Precision	Recall	Fscore
Inquirer only + Stemmer	0.32	0.02	0.05
Inquirer + WN expansion	0.48	0.67	0.56
Inquirer + WN expansion + Stemmer	0.46	0.95	0.62

Table 2: Results Using Inquirer Dictionary

We tested this list on the TREC 2003 Novelty Track sentences of opinion topics. Table 2 shows the results. We also applied a stemmer to avoid strict string matching. (It is not suprising that only using Inquirer words without WordNet expansion performed poorly given its small size.) We applied this list for the run ISIRUN304.

3.3 Using TREC 2003 Data

We used the official relevant sentences from the 2003 Novelty Track as training data for our run ISIRUN404. Using the file "qrels.relevant.03.txt", which contains all relevant sentences that were selected by human annotators at NIST, we divided all sentences in relevant documents into two categories: relevant (*R*) and non-relevant (*NR*). Then we applied the Brill tagger to locate all verbs, adjectives, nouns and modal verbs that we believed played important roles in determining relevance and subjectivity of a sentence. For each word, we calculated the probability of the word being relevant and non-relevant based on a model described by the following formulas.

$$P_R(w) \cong \frac{count\ (w \in R)}{count\ (w)}$$

$$P_{NR}(w) \cong \frac{count(w \in NR)}{count(w)}$$

We subtracted $P_{NR}(w)$ from $P_R(w)$:

$$Diff(P_R(w), P_{NR}(w)) = P_R(w) - P_{NR}(w)$$

The larger $Diff(P_R(w), P_{NR}(w))$, the more likely w is a good representative of the *R* class instead of the *NR* class, and the hence more useful as an opinion-bearing indicator. To obtain only reliable words, we experimented with various cut-off values λ as threshold. Table 3 shows numbers of indicator words according to λ.

λ	# of words
0.0001	308
0.00001	4921
0.000001	12877
0.0000001	14267

Table 3: Number of subjectivity indicator words for various thresholds λ

After collecting these words, we checked each sentence in the test data and marked it as "relevant" if it contains at least one of them. For the official run, we experimentally selected λ=0.0001 based on development test data of Novelty 2003.

3.4 Combination of All Three

For the run ISIRUN504, we simply combined all words from 3.1 through 3.3 and used them to select relevant sentences from the test data.

4 Event Topics

We treat event identification as a traditional document IR task. The goal here is to identify those event sentences that are relevant to the given event from the given list of documents. For the IR part we treat each sentence independently of other sentences and index them accordingly. We thus reduce the problem of event identification to that of sentence retrieval.

We choose the description <desc> field for formulating the query. To perform IR we use a probabilistic Bayesian inference network model as implemented in the search engine software package INQUERY. For each query we perform

[1] http://www.wjh.harvard.edu/~inquirer/
http://www.wjh.harvard.edu/~inquirer/homecat.htm

the standard procedure of stop-word removal and stemming. Having performed the search we return all sentences that have non-zero scores as the final answer. For example, issuing the query "*Arrest of former Chilean dictator, General Augusto Pinochet, in London. He was charged with murder, torture, genocide, and terrorism during his regime in Chile.*" we obtain the following sentences as output:

1. Garzon's request for Pinochet's arrest and extradition on charges of genocide, torture and terrorism led to the dictator's detention in London last October.
2. The warrant said the general was wanted for questioning for "crimes of genocide and terrorism that includes murder''.
3. Prosecutors are arguing that the genocide, mass murder and torture charges against the former dictator should overrule Britain's immunity for former heads of state.
4. Pinochet was arrested at the instigation of a Spanish magistrate, who is seeking the general's extradition on charges of genocide, murder and torture during his 17-year-rule.

.
.
.

5 Performance

5.1 Submitted 5 Runs

We submitted 5 runs for Task 1. As a baseline system to compare our methods with, we produced ISIALL04 by marking every sentence in relevant documents as RELEVANT. Surprisingly, this baseline performed relatively well. For other 4 runs, we combined the event topic result described in Section 4 with the 4 opinion topic methods described in Section 3, since we focused mainly on opinion topics. Table 4 shows the performances of the 5 runs.

Run	Precision	Recall	F-score
ISIALL04	0.26	0.84	0.371
ISIRUN204	0.30	0.74	0.385
ISIRUN304	0.30	0.73	0.387
ISIRUN404	0.30	0.71	0.390
ISIRUN504	0.30	0.74	0.385

Table 4: Performance of 5 submitted runs

6 Conclusion

In this paper we presented our work on the Novelty track at TREC 2004. This track presents several challenges in terms of treating a sentence as information unit. We focused on recognizing relevant sentences from opinion type topics. Unlike event topics, we did not consider whether a sentence contains any phrase referring to the topic. We assumed that whether a sentence is opinion-bearing or not is much more important. We are curious to learn about other methods for document filtering and the effect on the overall performance of a system.

References

Callan, J.P., W.B. Croft, and S.M. Harding. 1992. The INQUERY Retrieval System. *Proceedings of DEXA-92, 3rd International Conference on Database and Expert Systems Applications.*

Fellbaum, C., D. Gross, and K. Miller. 1993. Adjectives in WordNet. http://www.cogsi.princeton.edu/~wn.

Kneser, R. and H. Ney. 1995. Improved Backing-off for n-gram Language Modeling. *Proceedings of ICASSP*, vol. 1, 181–184.

Miller, G.A., R. Beckwith, C. Fellbaum, D. Gross, and K. Miller. 1993. Introduction to WordNet: An On-Line Lexical Database. http://www.cogsi.princeton.edu/~wn.

Riloff , E. and J. Wiebe. 2003. Learning Extraction Patterns for Opinion-bearing Expressions. *Proceedings of the EMNLP-03*.

Riloff, E., J. Wiebe, and T. Wilson 2003. Learning Subjective Nouns Using Extraction Pattern Bootstrapping. *Proceedings of CoNLL-03.*

Soboroff, I. and D. Harman. 2003. Overview of the TREC 2003 Novelty Track. *Proceedings of TREC-2003.*

Stolcke, A. 2002. *SRILM — An Extensible Language Modeling Toolkit*. *Proceedings of the Intl. Conf. Spoken Language Processing*, Denver, CO.

Turney, P. 2002. Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews. *Proceedings of the 40th Annual Meeting of the ACL*, Philadelphia, 417–424.

Wiebe, J.M. 2000. Learning subjective adjectives from corpora. *Proceedings of the 17th National Conference on Artificial Intelligence.* Menlo Park, CA: AAAI press.

Wilson, T. and J. Wiebe. 2003. Annotating Opinions in the World Press. *Proceedings of the ACL SIGDIAL-03.*

Yu, H. and V. Hatzivassiloglou. 2003. Towards Answering Opinion Questions: Separating Facts from Opinions and Identifying the Polarity of Opinion Sentences. *Proceedings of EMNLP-2003*.