Discovering Subjectivity Using Multi-document Summaries

Michele Banko and **Lucy Vanderwende**

Microsoft Research
One Microsoft Way
Redmond, WA 98052 USA
{mbanko, lucyv}@microsoft.com

Abstract

The ability to separate opinion from fact within text requires accurate detection of subjective language. Most work in this area has relied on some level of human supervision in the form of hand-tagging or word-list construction. Drawing from the unique nature of the multi-document summarization task, we present an unsupervised method for discovering subjectivity clues using freely available multi-document summarization corpora.

Introduction

In the context of multi-document summarization, we expect that a good summary will provide a synthesis of multiple views of an event being described over a set of documents. A summarist is required to generalize, condense and merge information coming from multiple sources. Frequently, the summarist offers a high-level view of an event that is not explicitly reflected in any single document. A useful multi-document summary will also indicate the presence of new or distinct information contained within a set of documents describing the same topic (McKeown et. al., 1999). Furthermore, humans, for whatever reason, will sometimes use artistic licensing when creating summaries, which often lends an air of opinion to the text.

As a result of the requirements and phenomena described above, humans often generate new text when forming a summary of multiple documents. It is not sufficient to simply cut-and-paste existing document text as it might be when summarizing a single document (Jing and McKeown, 2000). In order to explore this idea, we took freely available summarization corpora and compared the text from human-written multi-document summaries to the source documents from which they were constructed. We found that the significant percentage of the terms present in the summaries, but not in the full text, belongs to the class of words previously attributed to identifying subjective states.

In the following sections, we review what it means for a term to be subjective, using examples we extracted from multi-document summarization data. We next describe our fully-automatic method for locating these terms, and then provide an evaluation and analysis. Finally we touch on related research and ideas for future applications of our findings.

Expressing Subjectivity in Summaries

Subjective language is that which is used to express opinions, evaluations, emotions and speculations (Banfield, 1982). Such *private states* cannot be directly observed, (Quirk, et. al. 1985), but can be expressed within language in two ways: they can be specifically mentioned as in (1) and (2) below, or they can be indirectly described via the style and type of language used as in (3).

(1) A UN envoy mediating the dispute *appeared optimistic* about a plan for broad autonomy for East Timor.

(2) *Revered* by Croats as an anti-communist martyr, Stepinac is *regarded* by most Serbs as a Nazi sympathizer.

(3) The airport will ease the *claustrophobia* of Gazans and provide a *boost* to the *troubled* Palestinian economy.

Taken from the DUC 2003 multi-document summarization corpus, the text in Figures 1 and 2, which recounts the 1998 Leonid meteor shower, illustrates in several ways how the pressure to provide a short account of multiple views of an event can cause human summarizers to introduce subjectivity into their descriptions. We have italicized the novel terms detected in the human summaries.

In the first snippet, we see the human summarizer highlighting the collective disappointment of those observing the showers, drawing this conclusion from both quoted anticipation and scientific facts. In the second example, we notice the country of Norway and the region of Southern Europe have been combined into a more general entity, Europe. Restrictions on summary length force the human to find a description of the events that encompasses the variety of experiences found in both areas. This pressure to combine different accounts of the event results in a new description of the event -- “more impressive.”

Figure 1

Source Document: ``Some people think the Leonid storm this year will be as good as the one in 1966…” But nowhere did the reported rate of meteor sightings greatly exceed 2,000 per hour - barely one-tenth the rate at which meteors hit the atmosphere during the great 1966 Leonid storm. … [T]heir enjoyment was tempered by the sight of clouds moving in.

Human Summary: The storm was *disappointingly* light compared to 1966.

Figure 2

Source Text: The meteor display was unexpectedly good over southern Europe, and came several hours earlier than predicted. … The view apparently was clearer in Norway…

Human Summary: In *fact* the showers were more *impressive* viewed from Europe than East Asia and there was no damage to satellites.

Experiments

For our experiments we used data made available from the 2003 Document Understanding Conference (DUC), an annual large-scale evaluation of summarization systems sponsored by the National Institute of Standards and Technology (NIST). In this corpus, NIST has provided documents describing 30 TDT events, taken from the Associated Press, New York Times, and English Xinhua newswires. On average, an event is described by 10 separate (but not necessarily unique) documents, which together consist of around 5700 words. Additionally, a total of four summaries is provided for every event, each hand-written by a distinct individual; each summary is approximately 100 words in length.

In order to compare the text of human-authored multi-document summaries to the full-text documents describing the events, we construct a minimal tiling of each summary sentence. More specifically, for each sentence in the summary, we search for all n-grams that are present in both the summary and the documents, placing no restrictions on n. We then cover each summary sentence with the n-grams, using as few n-grams as possible (i.e. favoring n-grams that are longer in length). Text that is not covered by an n-gram from the source documents is then placed into our list of terms.

We found that 22.5% of n-grams found by constructing a minimal tiling appeared in the human-generated summaries but not in the source documents. 89.5% were unigrams, and 8.3% were bigrams, with the remaining handful consisting of trigrams and four-grams.

As a final step, we filter out those terms consisting exclusively of stopwords or containing numbers. We found the latter to be necessary as a result of the observation that summarizers frequently replace words such as “today” or “Tuesday” with an actual date in order to provide a more lasting summary context. This filtering step caused us to remove 62 out of 468 terms from our initial list. Table 1 shows a sample of the words found by our analysis.

Table 1: Sample of words found by our analysis

alleged	deliberately	position
appeared optimistic	denounced	proposed
attack	disagreed	reaction
believing	fortunately	reflected
blamed	generally	reportedly
characterized	hoped	surprisingly
charged	intention	threatened
claims	obvious	unanimously disagrees
compromised	offensive	uncertain
declared	passed judgment	worst

Evaluation and Observations

From our list of 406 terms, we selected a random sample of 100, and provided the terms, along with the context in which they appeared in the human summary, to a human assessor for evaluation. Using two references for guidelines (Riloff et. al. 2003, and Wiebe et. al. 1999), the linguistic assessor was asked if the term indicates subjectivity; if yes, the assessor was further asked to indicate whether it explicitly refers to opinion, emotion or speculation, or if it reflects an indirect private state. In this pilot study, 71% of the terms in our random sample were judged to indicate subjectivity. Of those, 57.14% are explicit referents, with the remainder (42.86%) indirectly indicating private states.

We then had three annotators perform the evaluation for the whole set of terms, providing the annotators with same references and using a substantial portion of the random sample above as training material. The results count a term to be subjective if 2 out of 3 annotators agreed; annotators agreed unanimously on 80% of the terms judged to be subjective. The results obtained are different from the pilot study, as shown below in Table 2. We therefore had the first linguistic assessor perform the same evaluation as the annotators; the results, however, did not differ significantly from those of the annotators.

Table 2: Evaluation of Mined Terms

Category	Percent of Terms Extracted
Subjective	46.19
Explicit Mention	81.50
Private States	17.50

Upon looking at the terms in our sample that were not judged to indicate subjectivity, we found simple additions we could make to our filtering process. Adding names of days, months, and named entities would have improved accuracy of the terms in our sample another 5-6%. An analysis of 10 trials of 100 random terms showed the variation for each annotator to be 12-16% per trial, thus the set of 100 terms sampled does have an influence on the results. This suggests the pilot study may have been a particularly high yield trial run.

Related Work

In addition to what this analysis reinforces about the nature of useful multi-document summaries, the connection to work on identifying subjectivity patterns from text should not be overlooked.

Methods for extracting subjectivity patterns have mostly relied on some form of supervision such as hand-annotated data (Wiebe et. al. 1999; Bruce and Wiebe, 1999, Wiebe et. al. 2001) or the manual construction of a list of seed words to support an automated bootstrapping approach (Riloff et. al. 2003). Our fully automatic method for extracting an initial set of terms which indicate subjectivity might replace the hand-constructed input to such a system. The resulting output may then be used to classify input at either the sentence or document level as being either subjective or objective.

Conclusions and Future Work

We have presented an unsupervised method for extracting subjective terms from a multi-document summarization corpus. Our discovery is motivated by the nature of the multi-document summarization task, in which the need to generalize, condense and merge information frequently results in the use of subjective language.

Useful multi-document summaries are able to provide a level of abstraction that cannot easily be achieved by simply extracting text from any portion of a document. At the surface level of analysis, this can be realized by the presence of novel terms in the summary text. We are currently investigating whether or not the presence of such terms correlates with summary quality as determined by NIST evaluators.

Acknowledgements

We wish to thank Chris Brockett for encouraging us to formally pursue this idea, and Eric Ringger for providing tools pertaining to the computation of minimal tiles. We would also like to thank the Butler Hill Group and its annotators for providing the formal evaluation and data analysis.

References

Banfield, A. 1982. *Unspeakable Sentences*. Routledge and Kegan Paul, Boston.

Bruce, R. and Wiebe, J. 1999. Recognizing subjectivity: A case study of manual tagging. *Natural Language Engineering, 5(2).*

Document Understanding Conference. 2003. http://duc.nist.gov

Jing, H. and McKeown, K. 2000. Cut and paste based text summarization. *In the Proceedings of the 1st Conference of the North American Chapter of the Association for Computational Linguistics*. Seattle, Washington.

McKeown, K. Klavans, J, Hatzivassiloglou, V., Barzilay R., and Eskin, E. 1999. Towards multidocument summarization by reformulation: Progress and prospects. In *Proceedings of AAAI.*

Quirk, R., Greenbaum, S., Leech, G., and Svartvik, J. (1985). *A Comprehensive Grammar of the English Language.* Longman, New York.

Riloff, E., Wiebe, J., and Wilson, T. 2003. Learning Subjective Nouns using Extraction Pattern Bootstrapping. *Proceedings of the Seventh Conference on Natural Language Learning (CoNLL-2003).*

Wiebe, J., Bruce, R. and O'Hara, T. 1999. Development and use of a gold standard data set for subjectivity classifications. In *Proceedings of 37th Annual Meeting of the Association for Computational Linguistics*, pages 246-253, University of Maryland.

Wiebe, J., Wilson, T., & Bell, M. (2001). Identifying Collocations for Recognizing Opinions. *Proc. ACL 01 Workshop on Collocation.* Toulouse, France, July 2001.