Collocations as Cues to Semantic Orientation

Faye Baron and **Graeme Hirst**
Department of Computer Science
University of Toronto
Toronto, Ontario, Canada M5S 3G4
faye@cs.toronto.edu, gh@cs.toronto.edu

Introduction

Techniques to classify opinion or sentiment in text as either *positive* or *negative* employ features such as individual words, bigrams, and part-of-speech patterns. When individual words and bigrams are employed, both the frequency of their presence and their predetermined probable polarity are used to decide the polarity of the text to be classified; neutral words and bigrams are not useful in this task. We refer to words that provide polarity cues as *nuance-bearing*.

Joanna Channell (2000) claims that through the systematic manual examination of corpus data, properties of evaluative expressions that cannot be intuitively understood can be assigned. Specifically she looks at concordances and identifies expressions or collocations in which the words individually do not carry evaluative nuance but which collectively are predominantly used in a positive or negative connotation. She provides examples to illustrate her theory:

1. *par for the course*: Though the word *par* may have mild positive connotation, the remaining words *for the course* are not imbued with any semantic polarity. Channell found that this collocation was usually found in a negative context.
2. *off the beaten track*: The word *beaten* has many negative connotations relating to *defeat* and *assault*. The remaining words in this collocation are neutral. Yet this expression usually refers to a positive, idyllic place and is most often found in a positive context.

Channell's observations have interesting implications in determining the polarity of a sentence or text: There are words that are neutral and would be ignored in isolation as cues to evaluative orientation, but which, when appearing in a collocation, provide a positive or negative polarity cue. Unigrams and bigrams have been implemented as affect-bearing polarity cues. Clearly, the set must be expanded to include longer and more complex collocations. In particular, we need to look beyond simple strings of adjacent words (or *n*-grams), and consider both phrasal patterns and templates with variable slots as potential cues to semantic orientation.

Here, we follow Smadja's (1993) view of collocations as lexical clusters that are *domain-specific*, *context-recurrent*, and *cohesive*. They include two-word *predicative relations* that appear in a fixed syntactic relationships to each other, but not necessarily contiguously; for example, the collocation *make ... decision* may be realized as *to make a decision*, *decisions to be made*, and *make an important decision*). Collocations also include *phrasal templates*: phrase-length structures that contain one or more variable slots that must be filled in a particular way. These slots may contain different values: *The Dow Jones average fell NUMBER points to NUMBER*. McKeown and Radev (2000) view collocations as lying somewhere on the continuum between rigid idioms and free word associations. In this research, we expand this view to regard idiomatic expressions as collocations.

The primary goal of this research is to test the hypothesis that there are nuance-bearing collocations that can be used to determine the subjective orientation of text when other orientation cues are not present, and to determine whether these collocations exist in significant numbers. While Channell claims that in her manual examination of text that she has found this to be true, it remains to be seen whether these collocations can be automatically extracted, orientation-tagged, and used as feature sets to classify text orientation, and whether their use augments current sets and improves classification standards. In this paper, we describe research that is presently under way to answer these questions.

In addition, we want to provide a repeatable technique (perhaps even a toolkit) for extracting new nuance-bearing collocations. While executing this technique once for the purpose of this research may provide a single feature set to assist in the determination of orientation, language continues to evolve and new collocations will frequently appear that should be added to the set. As well, this technique may be useful in extracting domain-specific nuance-bearing collocations from a corpus for the purpose of a domain-specific orientation classification.

Related Work

Subjectivity of text

The classification of text as either subjective or objective is clearly a precursor to determining the orientation of evaluative text since objective text is not evaluative by definition, and needs to be eliminated from this exercise. Wiebe's research, often in collaboration with other researchers, focuses

on subjectivity tagging, and identifying the characteristics of subjectivity. Subjectivity tagging and differences of opinion about subjectivity have been explored by Bruce and Wiebe (1999) and Wiebe et al. (2001). Hatzivassiloglou and Wiebe (2000) have explored different forms of adjectives and their usefulness as subjectivity clues. More recent collaborations between Riloff, Wiebe and Wilson (Riloff, Wiebe and Wilson 2003, Riloff and Wiebe 2003) have incorporated Riloff's *bootstrapping* techniques (1996) to extract words and patterns which are useful subjectivity classifiers. The significance of this later work (apart from the use of bootstrapping) is that it moves away from simple unigrams and bigrams as classification features and uses phrase patterns to identify subjective text — a step in the direction of using collocations as classifiers.

Nuance-bearing words and bigrams

Techniques to classify evaluative text often rely on the presence of unigrams or bigrams which have been correlated to positive or negative orientation. Hatzivassiloglou and McKeown (1997) used the words *and*, *or*, and *but* as linguistic cues to extract adjective pairs, which they then clustered into positive and negative partitions. Turney (2002) assessed the semantic orientation of two-word phrases using their occurrence near the strongly-polarized words *excellent* and *poor*. Yu and Hatzivassiloglou (2003) have extended Turney's work and used the co-occurrence of a word and its part-of-speech tag with a set of previously classified nuance-bearing words to calculate the polarity measure of that word. This technique has significantly expanded the set of nuance-bearing words to be considered as features.

Collocation extraction

Though considerable study has been performed in the extraction of collocations, the space limitations of this paper limit our discussion to the work of Frank Smadja (1993), as it is his **Xtract** procedure that we implement.

Smadja's algorithm incorporates the technique of Choueka, Klein, and Neuwitz (1983) in which frequency identifies significant (contiguous) *n*-grams, and Church and Hanks's *mutual information* metric (1989), which looks at the cohesiveness of words that are *near* each other. The latter permits Smadja to expand from Choueka's rigid noun phrases to include phrasal templates. Through the application of filters with prespecified thresholds, Smadja effectively eliminates most of the insignificant lexical clusters retaining meaningful collocations. McKeown and Radev (2000) indicate that after passing through the filtering process in the third and final phase of Smadja's process, 80% of the collocations retrieved were considered *good* by a lexicographer. Smadja's technique for retrieving phrasal template collocations provides an effective method of identifying potentially nuance-bearing features that extend beyond simple words and *n*-grams.

Data

Our data is the 3,144 part-of-speech-tagged written texts in the British National Corpus, World Edition (BNC) (Burnard 2000).

McKeown and Radev (2000) point out that collocations tend to be dialect-specific. Collocations that are used in British English may not be used in American English. The implication of this is that if the corpus used for collocation extraction is from one dialect, the corpus to test their usefulness must come from the same dialect. As well, collocations are often technical and jargonistic, and therefore domain-specific extraction and application may be more effective. Experiments with the data, such as implementing domain constraints, will determine whether this technique can apply generally to the entire BNC.

Procedure

Text preparation

Input: PoS-tagged BNC texts.
Output: BNC texts suitably formatted for collocation extraction.

We filter the BNC texts, removing extraneous information and modifying the tags to ensure that they are in the format required by **Xtract**.

Collocation extraction

Input: BNC texts suitably tagged for collocation extraction.
Output: Syntax-tagged collocations containing the "largest subsuming statistical *n*-grams" (Smadja 1993).

Using Smadja's **Xtract** program, we extract both phrasal templates and rigid noun phrases. In the first stage, significant bigrams are extracted. Through the use of automated concordances using the bigrams produced in this stage, collocations are produced then filtered in the second stage. In the third and final stage, using a bottom-up parser, and a partial parse tree, the pairwise syntactic relations (e.g., *subject-verb* or *verb-object*) between the words in the collocation are identified and examined. When the syntactic (modifier-modifier) relation is inconsistent between occurrences, that collocation is considered to be insignificant, and is filtered out.

Determining the evaluative orientation of collocations

Input: Collocations extracted from BNC and the BNC sentences.
Output: Orientation-tagged collocations.

In this phase we attempt to identify and tag those collocations that are predominantly found with either a positive connotation or a negative one. To accomplish this, we examine the neighborhood of *n* sentences on either side of each occurrence of a collocation, including the remaining words in the sentence in which the collocation occurs. If the orientation of the neighborhood is predominantly negative, the connotation of that occurrence will be counted as negative; if it is predominantly positive, it will be counted as positive. We have not yet determined the value of *n* and the exact threshold for *predominantly*. These values will be established through experimentation.

To determine the orientation of the neighborhood, the sentences in it must be individually classified as either positive or negative. Each sentence is treated as a *bag of words*. Depending on the frequency of nuance-bearing words (classification features) in the sentence, it is considered either positive, negative, or undetermined. We use the orientation-tagged words of the General Inquirer (Stone 1966) as features. As well, we apply the techniques of Yu and Hatzivassiloglou (2003) to generate new classification features from a seed set.

If *x* percent of a neighborhood of a collocation is counted as positive or counted as negative, then that collocation will be classified accordingly. Again, the threshold value of *x* will be established through experimentation.

Testing the usefulness of collocations

Input: Syntax- and orientation-tagged collocations; untagged sentences.
Output: Measure of success of collocations in determining evaluative orientation of text.

In the final phase, we look at the usefulness of our orientation-tagged collocations in determining the orientation of text. Using a either a secondary corpus, we extract a sample number of our collocations and their *n*-sentence neighborhoods. We will ask judges to manually tag the nuance of each collocation as it occurs in its extracted neighborhood. We will compare this to our automated classification to establish a measure of success. We will also look at those neighborhoods whose connotation (positive or negative) did not agree with that established for the collocation from its other occurrences in the corpus. It will be interesting to find out if these occurrences differ because of effects such as irony or sarcasm.

Conclusion

Collocations communicate more than the words that they are composed of. In this research we are seeking collocations that express affect, serving as cues that identify a positive or negative stance. We believe that the use of nuance-bearing collocations will be an important factor in improving the accuracy of systems that determine the orientation of text.

We expect that our technique will provide a tool for the on-going discovery and classification of new nuance-bearing collocations, both general and domain-specific, that arise as a result of the continual evolution of language.

References

Bruce, R. F., and Wiebe, J. 1999. Recognizing subjectivity: A case study of manual tagging. *Natural Language Engineering* 5(2):187–205.

Burnard, L., ed. 2000. *The British National Corpus Users Reference Guide*. Oxford: Oxford University Computing Services.

Channell, J. 2000. Corpus-based analysis of evaluative text. In Hunston, S., and Thompson, G., eds., *Evaluation in Text: Authorial Stance and the Construction of Discourse*, Oxford Linguistics. Oxford University Press. 38–55.

Choueka, Y.; Klein, S. T.; and Neuwitz, E. 1983. Automatic retrieval of frequent idiomatic and collocational expressions in large corpus. *Association for Literary and Linguistic Computing Journal* 4(1):34–38.

Church, K. W., and Hanks, P. 1989. Word association norms, mutual information and lexicography. In *Proceedings of the 27th Annual Conference of the Association for Computational Linguistics*, 76–83.

Hatzivassiloglou, V., and McKeown, K. R. 1997. Predicting the semantic orientation of adjectives. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and the 8th Conference of the European Chapter of the Association for Computational Linguistics*, 174–181.

Hatzivassiloglou, V., and Wiebe, J. 2000. Effects of adjective orientation and gradability on sentence subjectivity. In *Proceeding of the 18th International Conference on Computational Linguistics*.

McKeown, K. R., and Radev, D. R. 2000. Collocations. In Dale, R.; Moisl, H.; and Somers, H., eds., *A Handbook of Natural Language Processing*. Marcel Dekker.

Riloff, E., and Wiebe, J. 2003. Learning extraction patterns for subjective expressions. In *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing*, 105–112.

Riloff, E.; Wiebe, J.; and Wilson, T. 2003. Learning subjective nouns using extraction pattern bootstrapping. In Daelemans, W., and Osborne, M., eds., *Seventh Conference on Natural Language Learning (CoNLL-03)*, 25–32. Association for Computational Linguistics.

Riloff, E. 1996. Automatically generating extraction patterns from untagged text. In *Proceedings of the Thirteenth National Conference on Artificial Intelligence (AAAI-96)*. American Association of Artificial Intelligence.

Smadja, F. 1993. Retrieving collocations from text: Xtract. *Computational Linguistics* 7(4):143–177.

Stone, P. J.; Dunphy, D. C.; Smith, M. S.; Ogilvie, D. M.; and associates. 1966. *The General Inquirer: A Computer Approach to Content Analysis*. The MIT Press.

Turney, P. 2002. Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, 417–424.

Wiebe, J.; Bruce, R.; Bell, M.; Martin, M.; and Wilson, T. 2001. A corpus study of evaluative and speculative language. In *Second Association for Computational Linguistics SIGDIAL Workshop on discourse and Dialogue*.

Yu, H., and Hatzivassiloglou, V. 2003. Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences. In *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing*, 129–136.