Identifying Subjective Adjectives through Web-based Mutual Information

Marco Baroni
SSLMIT, University of Bologna
Corso della Repubblica 136
47100 Forlì (FC), Italy
baroni@sslmit.unibo.it

Stefano Vegnaduzzo
AskJeeves, Inc.
5858 Horton Street, Suite 350
Emeryville, CA 94608, USA
svegnaduzzo@askjeeves.com

Abstract

This paper describes a method for ranking a large list of adjectives according to a subjectivity score without resorting to any knowledge-intensive external resources (such as lexical databases, parsers or manual annotation). The method only requires a list of adjectives to be ranked and a small set of "seeds" (manually selected subjective adjectives). The subjectivity score is obtained by computing the mutual information of pairs of adjectives taken from each set, using frequency and cooccurrence frequency counts on the World Wide Web, collected through queries to the AltaVista search engine. The obtained results improve significantly over a comparable low-resource acquisition algorithm.

1 Introduction

In recent years an extensive body of research has addressed the general problem of the acquisition (manual and automatic) and evaluation of lexical resources. Within this broad domain, growing attention has been devoted to the acquisition of subjective expressions. These are linguistic terms or phrases which convey the point of view (opinion, evaluation, emotion, speculation) of the author or other source mentioned in a text (Wiebe, 1994).

NLP applications that could benefit from use of these resources include information extraction, summarization, text categorization/genre detection, lexicography, and others. A recent and extensive overview of current work in the area of subjectivity analysis is provided by Wiebe et al. (2002).

In this paper, we propose a method to identify new subjective adjectives among a set of candidates, using only a small set of manually selected subjective adjectives as knowledge source. Our method is based on the idea that subjective adjectives will tend to cooccur in subjective texts. Thus, given a sufficiently large corpus, we can look for new subjective adjectives simply by searching for adjectives that tend to cooccur with our seeds. Following the method of Turney (2001), we use the web as our corpus, computing cooccurrence statistics via automated queries to the AltaVista search engine.

The results of our experiments are very encouraging, both in absolute terms (with precision above 70% at 50% recall) and when the performance of our algorithm is compared to that of the similarly knowledge-poor subjective adjective acquisition method proposed by Vegnaduzzo (2004).

The rest of this paper is organized as follows: In section 2 we review related work; in section 3 we motivate and describe our application of web-based mutual information to the subjective adjective mining task; our experiments are presented in section 4; we conclude the paper by discussing current problems and sketching directions for further work in section 5.

2 Related work

2.1 Acquisition of subjective expressions

We can distinguish two connected directions in the research on subjectivity: a) methods for acquiring subjective expressions (Hatzivassiloglou and McKeown, 1997); b) methods for classifying documents or sentences as subjective or not (Hatzivassiloglou and Wiebe, 2000; Tong, 2001; Pang et al., 2002), often relying on subjective expressions acquired with the former methods. Our research falls into the first category. The need for algorithms to acquire subjective expressions arises from the fact that existing lexical databases such as WordNet typically do not provide subjectivity information.

Wiebe (2000) is a landmark study on the acquisition of subjective adjectives that exemplifies the typical approach. Human judges were asked to classify sentences in a corpus as subjective or objective, and to identify the subjective expressions. Then the adjectives from the subjective sentences were collected, and used as seeds for an automatic thesaurusbuilding process based on a broad-coverage parser. This process yields 20 new adjectives for each original seed extracted from the manually tagged corpus.

Acquisition methods along these lines often achieve very good results. However, they are substantially dependent on the availability of knowledge-intensive resources, such as parsing tools and annotated data.

As a first step in the direction of acquiring subjective adjectives using limited resources, Vegnaduzzo (2004) proposed a distribution-based bootstrapping method that only needs a set of seed adjectives and a part-of-speech tagger. Since this work is the closest in spirit to the knowledge-poor approach presented in this paper, we will use it as a term of comparison in the discussion and evaluation of our results (see section 4.3.3 below).

2.2 Web-based Mutual Information

The idea of computing an association measure by using statistics obtained from an Internet search engine was first introduced by Turney (2001), who proposed the Web-based Mutual Information (WMI) method.

The (pointwise) mutual information of two words w_1 and w_2 is:

$$MI(w_1, w_2) = \log_2 \frac{\mathrm{P}(w_1, w_2)}{\mathrm{P}(w_1)\mathrm{P}(w_2)} \qquad (1)$$

The mutual information between two words can be seen as the ratio between the probability of seeing one of the two words if we saw the other and the context-independent probability of seeing the word. On mutual information see, e.g., Church and Hanks (1989).

Turney computed the mutual information of target word pairs by using frequency and cooccurrence frequency data extracted from the web via the AltaVista search engine.[1] In particular, cooccurrence frequencies were computed using the AltaVista NEAR operator, which returns pages in which the two target words occur within 10 words of one another, in either order. Turney showed that WMI greatly outperforms a more sophisticated method based on Latent Semantic Analysis in a synonym identification task.

Other studies (Terra and Clarke, 2003; Baroni and Bisi, 2004; Grefenstette et al., 2004) have confirmed the effectiveness of WMI and related methods in various tasks, even when compared to more sophisticated association measures such as log-likelihood ratio and cosine similarity. This is remarkable, since it is well known that mutual information is prone to overestimation if frequency counts are too small (Manning and Schütze, 1999, chapter 5). Evidently, the sheer size of the web as a dataset solves most of the problems related to low frequency counts.

Of particular interest to us are the experiments with WMI presented by Turney and Littman (2003). They show that the semantic orientation (positive or negative) of a set of subjective words, in particular adjectives, can be determined by computing WMI scores with respect to a small set of (positive or negative) paradigm words. Our task, i.e., to extract subjective adjectives from a larger adjective set, is complementary to the one of Turney and Littman. Indeed, our method to identify subjective adjectives can be seen as a preprocessing step for the algorithm of Turney and Littman, which can then be applied to the extracted list of subjective adjectives in order to determine their semantic orientation.

3 Mining subjective adjectives with WMI

We focus on the automated discovery of subjective adjectives. Adjectives are a well-known linguistic means to express point of view. In particular, Bruce and Wiebe (1999) have shown a statistically significant positive correlation of adjectives with subjective sentences in a tagged corpus.

The basic intuition motivating our method is that subjective adjectives will tend to occur in the near of other subjective adjectives. Consider a typical subjective text, such as a record review. It is extremely likely that it will contain not one, but a whole set of different subjective adjectives (e.g., in the case of a positive review: *great, gorgeous, stunning*...) Thus, if we start with a small list of known subjective adjectives, we can enlarge it by looking for other adjectives that tend to occur in their vicinity.

Cooccurrence statistics of this sort must be collected from a very large corpus, because of data sparseness issues. For example, of the 3047 adjectives in our test set (see section 4.1 below), only 19 occurred at least once within 10 words of a seed in the 1.2M-word corpus they were extracted from.

Thus, we decided to collect web-based cooccurrence statistics using the WMI method. Of the same 3047 adjectives, only 40 (all typos) did not cooccur at least once with at least one seed in the AltaVista-based dataset.

[1] `http://www.altavista.com`

We compute WMI using the following formula:[2]

$$\mathrm{WMI}(w_1, w_2) = \log_2 N \frac{\mathrm{hits}(w_1 \,\mathrm{NEAR}\, w_2)}{\mathrm{hits}(w_1)\mathrm{hits}(w_2)} \quad (2)$$

In (2), $\mathrm{hits}(w_1 \,\mathrm{NEAR}\, w_2)$ is the number of documents retrieved by AltaVista for a query in which the two target words are connected by the NEAR operator and $\mathrm{hits}(w_n)$ is the number of documents retrieved for a single word query. We also report (in section 4.3.4) results obtained without using the NEAR operator, i.e., counting as cooccurrences all documents in which both target words appear, independently of their proximity.

We take N, the number of documents indexed by AltaVista, to be 1 billion. This figure reflects AltaVista's self-reported index size as of September 2003. However, the N term is constant and it has no effect on the relative rank of pairs.

Given a list of unclassified adjectives and a list of seeds, we compute the WMI of each unlabeled adjective with each seed. We then rank these pairs on the basis of WMI, and we transform the WMI values into ranks. From now, whenever we use the term *WMI score*, we are referring to these ranks.

Thus, following standard practice in association measure studies, we work with an ordinal scale. In our case, this seems particularly sensible because, in absence of a reliable estimate for N in the WMI formula, we can trust relative ranks but not absolute differences between WMI scores. Using ranks also solves the problem of how to treat pairs that never cooccur: we can safely assign them the lowest rank, without worrying about how to estimate their WMI. We also experimented, albeit less extensively, with non-transformed WMI, obtaining results that are similar to the ones we report for rank-based scores.

Given that, for each unlabeled adjective, we have a set of WMI scores (one for each seed), we must decide how to determine the overall subjectivity value we will assign to that adjective. We can follow two strategies: we can pick one single score (for example, selecting the highest WMI score for each adjective), or we can take multiple scores into considerations (for example, summing the WMI scores of a test adjective with all the seeds). We will report results obtained with several variants of both strategies.

4 Experiments

4.1 Data

The same initial data as in Vegnaduzzo (2004) were used: A corpus of 1.2M words from the Reuters collection in the American News Corpus, and a manually created seed set. The corpus was tagged for part-of-speech information using an implementation of the Brill tagger (Ngai and Florian, 2001).

We used the same seed set of 35 adjectives that gave the best results in Vegnaduzzo's experiments. These adjectives had low frequency in the corpus and were manually selected on the basis of high subjectivity ratings from two judges.

We used as test set all the 3047 items tagged as adjectives that occur in the corpus and are not in the seed set. Notice that this list is rather noisy, containing typos (e.g., *authoratative*) and forms that have been mistagged as adjectives (e.g., *the*). However, since we are interested in a realistic testing scenario, we did not clean the list manually.

While the list of adjectives was obtained from the same corpus used by Vegnaduzzo (2004), the test set for the experiments carried out in that work was generated by obtaining subjectivity ratings only for a small subset of this list. Consequently, that test set turned out to be inadequate in evaluating precision and, especially, recall. In order to overcome these shortcomings, we rated for subjectivity all the adjectives in the list (in section 4.3.3, we present a re-evaluation of Vegnaduzzo's algorithm using our versions of the test set).

Both authors rated each adjective in the test set on a three-value scale: *3: Strongly Subjective*, *2: Possibly Subjective* and *1: Non-Subjective*. While there was no previous agreement on the rating criteria, we both used the intermediate value for weakly subjective adjectives (such as *charismatic* and *cautious*) and for adjectives that may be used with a clearly subjective meaning only in some contexts (e.g., *clear*, *commercial*, *sharp*). The inter-annotator agreement rate was at 79.2%. If strongly and weakly subjective ratings were merged, agreement rate grew to 84.2%.[3]

We evaluate the performance of the algorithm against two different labelings of the test set: in one version, we treat as subjective all adjectives that

[2] This formula can be straightforwardly derived from (1) by using the following joint and marginal probability estimates:
$\mathrm{P}(w_1, w_2) = \mathrm{hits}(w_1 \,\mathrm{NEAR}\, w_2)/N$
$\mathrm{P}(w) = \mathrm{hits}(w)/N$

[3] We also assigned subjectivity ratings to the seeds. See the brief discussion at the end of section 4.3.2.

were rated at least "possibly subjective" by at least one rater; in the other version, we consider subjective only those adjectives that were rated "strongly subjective" by both. We refer to the two versions of the test set as *lax* and *strict*, respectively. In the lax version, 972 adjectives (31.9%) are labeled as subjective; in the strict version, 104 adjectives (3.4%) are labeled as subjective.

The lax and strict sets offer evaluation perspectives at two opposite ends of a recall/precision continuum: How good the algorithm is at finding a larger, possibly noisier set of subjective adjectives vs. how good the algorithm is at finding a smaller high quality set. What is the most meaningful evaluation standard will depend on the intended application of the WMI method.

4.2 Procedure

We performed single word queries for all 3082 target adjectives (35 seeds + 3047 test items) and NEAR queries for each adjective in the test set paired with each seed. In total, 109,727 AltaVista queries were issued.

Using these data, we computed WMIs for each test/seed pair and we transformed them into ranks. We ended up with 35 rank-based WMI scores for each adjective in the test set (one per seed). We then needed to derive a single subjectivity score from these values. Although we experimented with many different ways to derive an overall score for each test adjective, here we will report only the results we obtained with the following methods, that are representative of general trends:

1. Measures based on single WMI scores:
   (a) Top WMI (*Top*)
   (b) First quartile WMI (*1Q*)
   (c) Median (second quartile) WMI (*Med*)
   (d) Third quartile WMI (*3Q*)
2. Measures based on composite WMI scores:
   (a) Sum of all WMI scores (*Tot*)
   (b) Sum of WMI scores in top fourth (*1QSum*)
   (c) Sum of WMI scores in top half (*1HSum*)
   (d) Sum of WMI scores in median-centered half (*MedSum*)

In the list above, single WMI scores for a test adjective are obtained by choosing a single score out of the 35 available WMI rank-based scores. For example, the median WMI score of test adjective *x* is the 18th WMI score in the ordered list of the 35 scores obtained for all the pairs comprised of adjective *x* and an adjective in the seed set.

Composite WMI scores for a test adjective are determined as a function (sum) of a subset (possibly all) of the 35 available WMI rank-based scores. In particular, since we report only composite values based on sums of the same number of scores for each candidate adjective, the composite score measures are equivalent to averages.

Once we compute a unique subjectivity score for all adjectives in the test set using one of these methods, we rank the adjectives on the basis of this score, and we calculate precision/recall profiles on the resulting ranked list.

4.3 Results

4.3.1 Lax set

Tables 1 and 2 report percentage precision at ten recall levels for the single and composite score measures presented in section 4.2. Precision and recall counts are based on the lax set described in section 4.1.

Recall	Top	1Q	Med	3Q
10	66.44	88.18	75.78	70.29
20	64.88	81.51	77.29	70.80
30	61.09	76.84	74.30	65.62
40	59.30	74.24	71.64	63.05
50	53.82	70.95	67.59	60.07
60	53.14	66.70	63.16	56.71
70	48.82	60.39	58.67	52.63
80	45.52	54.79	52.21	49.05
90	40.57	47.58	46.25	43.73
100	32.36	31.90	31.90	31.90

Table 1: Percentage lax set precision of single score measures at different percentage recall levels

Recall	Tot	1QSum	1HSum	MedSum
10	74.62	83.62	83.62	75.78
20	76.38	83.98	80.17	77.91
30	72.28	80.44	79.35	76.04
40	69.09	75.83	75.53	71.77
50	66.30	70.54	71.68	68.26
60	61.69	66.78	66.40	63.72
70	57.97	62.39	61.59	57.87
80	53.18	55.14	55.06	52.93
90	46.82	47.76	48.21	47.17
100	32.38	32.38	32.38	31.90

Table 2: Percentage lax set precision of composite score measures at different percentage recall levels

Overall, the results are very encouraging. With most measures, we can retrieve 40% of the subjective adjectives in the set (about 390 forms) with precision well above 70%. With three measures (first quartile score, top fourth sum, top half sum), precision is still above 70% at 50% recall (486 hits). At 90% recall, precision is still well above chance level (31%) for all measures.

Thus, here is clear evidence that a very simple method to look for subjective expressions combined with a very large database can lead to very good results (see section 4.3.3 below for a direct comparison with Vegnaduzzo's more sophisticated algorithm).

In both tables, the first quartile measures (1Q and 1QSum) achieve the best overall performance. This indicates that subjective adjectives are characterized by high cooccurrence with a smaller set of seeds, rather than with the full seed list. Consider for example the case of *imaginative*, one of the top subjective adjectives retrieved by the first quartile measures. This adjective has highly-ranked WMIs (within the top 2000) with seeds such as *clever, lively, eclectic*, that are clearly semantically related to it and have the same (positive) polarity. On the other hand, its WMI scores with unrelated and negatively connotated adjectives such as *poisonous, horrendous, unseemly* are low (second half of ranked WMI list). Clearly, high correlation with the first set is a much more meaningful fact than low correlation with the second set. In other words, a true subjective adjective is more likely to be highly correlated with *some* of the seeds, rather than being somewhat correlated with *all* the seeds. Not surprisingly, then, the list based on the sum of all scores is infested with highly ranked "generic" adjectives (such as *new* at rank 14) that tend *not* to be subjective.

Independently of the ranking method, many of the top false positives are forms that were wrongly tagged as adjectives, e.g. *the, as, betrayal, very*. This suggests that the performance of the WMI method can be further improved simply by improving POS tagging quality, or, more in general, by performing more careful preprocessing, in order to minimize the occurrence of false adjectives in the candidate set.

4.3.2 Strict set

Tables 3 and 4 report percentage precision at ten recall levels for the single and composite score measures. This time, performance statistics are computed using the strict set of section 4.1.

Recall	Top	1Q	Med	3Q
10	6.37	20.41	23.26	13.70
20	5.88	16.94	18.10	16.94
30	6.39	19.50	18.56	16.67
40	7.53	15.27	15.38	16.09
50	7.29	15.03	13.00	11.58
60	6.90	13.60	12.25	10.53
70	6.95	12.74	10.31	9.36
80	6.45	9.51	8.74	7.90
90	5.61	7.54	6.72	6.77
100	3.88	4.24	3.41	3.41

Table 3: Percentage strict set precision of single score measures at different percentage recall levels

Recall	Tot	1QSum	1HSum	MedSum
10	15.87	17.24	19.23	20.41
20	19.27	19.44	19.44	17.50
30	19.25	15.98	15.05	19.02
40	17.65	13.42	15.44	17.72
50	12.81	14.33	14.29	13.94
60	11.38	13.81	13.69	13.48
70	10.58	11.99	13.08	10.43
80	9.44	9.08	9.95	9.09
90	6.92	7.16	7.96	7.00
100	4.14	4.23	4.24	4.13

Table 4: Percentage strict set precision of composite score measures at different percentage recall levels

At first sight, these results look decidedly less impressive. However, they must be put into perspective. This was a much harder task, with chance level at 3.4% (vs. 31.9% with the lax set). In this respect, consider that the strict set precision around 15% obtained by the best measures at 50% recall is more than 4 times the chance level. Notice also that, like in the case of the lax set, the performance of all measures is still well above chance at 90% recall. Nevertheless, the results are clearly unsatisfactory in absolute terms, and further work is needed to improve the performance of the WMI method on the strict set task.

It is interesting to observe that the average rating assigned by the two authors to the seeds was of 2.4 and 2.2, respectively, suggesting that the seeds were much closer in subjectivity strength to the lax set adjectives than to the strict set adjectives. Thus, in the near future we plan to test whether performance on the strict set can be improved simply by choosing a "stricter" seed set.

4.3.3 Comparison with Vegnaduzzo's algorithm

We compared the results obtained with our strategy to those obtained by Vegnaduzzo (2004). This approach strives just like ours to minimize reliance on knowledge-intensive resources. However, it uses a fairly more complex algorithm than the simple method proposed here. In particular, Vegnaduzzo proposes a bootstrapping procedure based on the hypothesis that subjective adjectives tend to modify nouns that are themselves oriented towards subjectivity. Seed adjectives are used to identify all nouns they modify (the tagger is needed to find adjective-noun pairs in a corpus). Adjectives that modify those nouns and are not in the initial seed set are collected, and ranked by computing their average cosine similarity to the adjectives in the seed set. The most highly ranked adjectives are added to the seed set and the procedure restarts.

In order to carry out the comparison, we evaluated the results obtained by Vegnaduzzo's algorithm on the versions of the test set we created for our own experiments. We noted above how Vegnaduzzo's original test set was not adequate to assess performance (in particular, recall) properly.

Tables 5 and 6 report precision and recall scores for the lax set and the strict set as defined above, for 7 iterations of Vegnaduzzo's algorithm. The format of these tables is different from the one used above because of the different nature of the two methods. Our method generates a score for all adjectives in the test set, thus yielding a ranked list that allows to compute precision-recall profiles as we did above. On the other hand, at each iteration Vegnaduzzo (2004)'s method generates a discrete set of adjectives classified as subjective (this set includes new adjectives acquired at that iteration in addition to all those acquired in previous iterations).

Iteration	Recall	Precision
1	1.23	70.58
2	2.36	63.88
3	3.29	53.33
4	4.21	49.39
5	4.42	47.77
6	4.47	46.23
7	4.62	45.91

Table 5: Recall and precision scores at each iteration of Vegnaduzzo's algorithm (lax set)

As with WMI, the scores on the lax set are better than those on the strict set. However, WMI is clearly outperforming Vegnaduzzo's method, if we compare them at approximately the same data point on a precision-recall curve. For example (restricting now attention to the lax set), at 10% recall precision of the various WMI scores ranges between 66.44% and 88.18%. The best result with Vegnaduzzo's method is obtained in the first iteration, where precision is comparable to the lower WMI values (70.58%) but recall is much lower (1.23%). After the first iteration, recall increases only minimally and never even reaches the 10% level, whereas precision decreases quite rapidly. Analogous observations can be made for the strict set.

Iteration	Recall	Precision
1	2.88	17.64
2	2.88	8.33
3	4.80	8.33
4	4.80	6.02
5	4.80	5.55
6	4.80	5.37
7	4.80	5.10

Table 6: Recall and precision scores at each iteration of Vegnaduzzo's algorithm (strict set)

Notice also that the WMI-based method is much more robust and scalable. Vegnaduzzo's algorithm still relies on a part-of-speech tagger and it is limited to the acquisition of adjectives that immediately precede a noun. By design, it is limited to acquire terms of a particular syntactic category that are adjacent or in a linear order dependency with terms of another syntactic category. Moreover, it is based on various manually chosen parameter values (e.g., thresholds for selecting candidates) which must be experimentally re-adjusted for corpora and seed sets of different sizes. On the other hand, the WMI method does not incorporate any design constraint on the syntactic category of the terms it can process, since it only requires a list of seeds and a list of candidates, and it will yield a ranked list as output independently of data set size or any other parameter.

4.3.4 WMI without the NEAR operator

Around April 2004 (after we collected the data presented above), AltaVista stopped supporting the NEAR operator. Thus, if we still want to rely on this engine, we have to estimate cooccurrence frequency without NEAR, i.e., to search for pairs that occur on the same web-page independently of their distance.

The hypothesis behind our application of WMI is that subjective adjectives tend to occur in the same articles/documents. Given that web-pages will typically contain a single document or a set of related documents (e.g., reviews), it is not unreasonable to look for adjectives that occur in the same page, independently of their proximity. Thus, we repeated our experiments computing WMI without NEAR. Notice that, since other changes recently occurred in AltaVista, the comparison of current and previous results has to be taken with a grain of salt.

Single and composite lax set data are reported in tables 7 and 8, respectively (strict set results follow the same pattern).

Recall	Top	1Q	Med	3Q
10	60.25	76.38	80.17	84.35
20	61.01	71.06	76.38	80.50
30	57.37	70.36	74.11	74.49
40	57.12	65.27	68.61	70.73
50	54.61	62.95	64.54	65.68
60	51.64	59.37	60.54	61.11
70	48.75	56.15	57.38	57.82
80	44.41	50.82	52.18	51.49
90	38.94	43.58	44.46	44.53
100	32.71	33.17	32.74	32.95

Table 7: Percentage lax set precision of single "no NEAR" score measures at different percentage recall levels

Recall	Tot	1QSum	1HSum	MedSum
10	82.91	76.98	77.60	81.51
20	79.84	70.04	71.85	76.68
30	74.49	68.07	71.92	73.92
40	71.38	65.16	65.93	69.09
50	64.54	61.83	62.87	64.97
60	60.54	59.01	59.37	61.18
70	57.14	55.19	56.11	56.95
80	51.32	49.84	50.95	51.22
90	43.86	42.03	43.27	44.08
100	33.03	32.87	32.97	32.80

Table 8: Percentage lax set precision of composite "no NEAR" score measures at different percentage recall levels

The tables show that removing NEAR does have a negative impact on performance. For example, at the 50% recall cut-off, the best no NEAR measures attain precision around 65%, whereas the best measures computed with NEAR-based counts have precision above 70%. Thus, if possible, WMI should be computed keeping target word proximity into account.

At the same time, the performance drop is stark but not catastrophic: The difference between the highest precisions computed with and without NEAR is never larger than 6%. The overall performance is still well above the one obtained with Vegnaduzzo's algorithm, indicating that, even without NEAR, WMI remains the best performing knowledge-poor method to acquire subjective adjectives.

Interestingly, when NEAR is not used, the measures that take into account cooccurrence with the whole seed set or a large portion of it (Med, 3Q, Tot, MedSum) outperform those that only look at the highest WMI values (Top, 1Q, 1QSum, 1HSum). This is the opposite of what happened with NEAR. We hypothesize that the difference reflects the fact that, with and without NEAR, we find correlations of a different nature: With NEAR, we find adjectives that are related to the seeds at the semantic level, and thus tend to cooccur with a small set of very similar seeds within sentences and paragraphs, whereas without NEAR we find words that are related to the seeds at the topic/genre level, and thus tend to cooccur with many of them within documents or sets of related documents. Obviously, evidence of the first type leads to better results. This preliminary hypothesis requires further investigation.

5 Conclusion

The results we obtained (especially with the lax set and using NEAR) show that WMI is a viable knowledge-poor method to mine subjective adjectives. Although there is obviously room for improvement, we can already foresee practical applications of our method at the current performance level. For example, WMI could be helpful in knowledge engineering and lexical acquisition tasks where human editors need to identify terms and phrases related to a given set of seeds in a very short time-frame. The high precision levels especially at the top of the ranked output list should be more than adequate to speed up the the editors' manual review process, saving a considerable amount of time.

Following up on the observation that subjective adjectives seem to be picked up more reliably by subsets of highly correlated seeds, rather than by the full seed list, a promising direction for further work will be to identify different classes of subjective adjectives (perhaps using automated clustering techniques) and to build separate seed sets to mine adjectives from the various classes. Experimenting with different seed sets will also help clarifying whether the poor performance with the strict seed set is simply due to poor seed selection, or whether the problem is caused by other factors. More in general, it will be interesting to study whether the WMI method can be extended to other types of related items, functioning as a general purpose lexical acquisition tool.

However, before we can tackle these issues, we have to deal with the fact that AltaVista is no longer supporting NEAR queries. Our preliminary experiments (section 4.3.4 above) indicate that not using NEAR does affect the quality of the results (although not dramatically).

What happened is a symptom of a more general problem with Internet-based data mining methods relying on commercial search engines, namely that the query options offered by such engines, as well as their terms of service and availability, can change in ways that have nothing to do with satisfaction of the linguists' community. Thus, the most obvious solution to the problem will be to follow the example of Terra and Clarke (2003), who used web-crawling to construct their own large (53 billion words) corpus. Besides the obvious advantages in terms of stability, having their own web corpus also allows them to collect a wider range of cooccurrence statistics and to experiment with parameters such as the size of the cooccurrence window. An even better long term solution could be for the linguistic community to build its own search engine, as advocated by Fletcher (2004) and Kilgarriff (2003).

In the meantime, we believe that our data provided a valid example of how, given a small set of seeds and a list of candidates, mutual information computed on a very large corpus (the web) can be effectively used in the domain of automated acquisition of subjective expressions.

References

M. Baroni and S. Bisi. 2004. Using cooccurrence statistics and the web to discover synonyms in a technical language. In *Proceedings of LREC 2004*.

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

K. W. Church and P. Hanks. 1989. Word association norms, mutual information, and lexicography. In *Proceedings of ACL 1989*.

W. Fletcher. 2004. Facilitating the compilation and dissemination of ad-hoc web corpora. In *Proceedings of TALC 2002*.

G. Grefenstette, Y. Qu, D.A. Evans, and J.G. Shanahan. 2004. Validating the coverage of lexical resources for affect analysis and automatically classifying new words along semantic axes. In *AAAI Spring Symposium Technical Report: Exploring Affect and Attitude in Text*.

V. Hatzivassiloglou and K. R. McKeown. 1997. Predicting the semantic orientation of adjectives. In *Proceedings of ACL 1997*.

V. Hatzivassiloglou and J. Wiebe. 2000. Effects of adjective orientation and gradability on sentence subjectivity. In *Proceedings of COLING 2000*.

A. Kilgarriff. 2003. Linguistic search engine. In *Proceedings of Corpus Linguistics 2003*.

Ch. Manning and H. Schütze. 1999. *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, MA.

G. Ngai and R. Florian. 2001. Transformation-based learning in the fast lane. In *Proceedings of NAACL 2001*.

B. Pang, L. Lee, and S. Vaithyanathan. 2002. Thumbs up? Sentiment classification using machine learning techniques. In *Proceedings of EMNLP 2002*.

E. Terra and C. L. A. Clarke. 2003. Frequency estimates for statistical word similarity measures. In *Proceedings of HLT-NAACL 2003*.

R. M. Tong. 2001. An operational system for detecting and tracking opinions in on-line discussions. In *Working Notes of the SIGIR 2001 Workshop on Operational Text Classification*.

P. Turney and M. Littman. 2003. Measuring praise and criticism: Inference of semantic orientation from association. *ACM Transactions on Information Systems (TOIS)*, 21(4):315–346.

P. Turney. 2001. Mining the web for synonyms: PMI-IR versus LSA on TOEFL. In *Proceedings of ECML 2001*.

S. Vegnaduzzo. 2004. Acquisition of subjective adjectives with limited resources. In *AAAI Spring Symposium Technical Report: Exploring Affect and Attitude in Text*.

J. Wiebe, T. Wilson, R. Bruce, M. Bell, and M. Martin. 2002. Learning subjective language. Technical report TR-02-100, Department of Computer Science, University of Pittsburgh, Pittsburgh, PA.

J. Wiebe. 1994. Tracking point of view in narrative. *Computational Linguistics*, 20(2).

J. Wiebe. 2000. Learning subjective adjectives from corpora. In *Proceedings of AAAI 2000*.