Acquisition of Subjective Adjectives with Limited Resources

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

Abstract

This paper describes a bootstrapping algorithm for acquiring a lexicon of subjective adjectives which minimizes the recourse to external resources (such as lexical databases, parsers, manual annotation work). The method only employs a corpus tagged with part-of-speech information and a seed set of subjective adjectives. The list of candidate subjective adjectives is generated incrementally by looking at the head nouns they modify and computing their distribution-based semantic similarity (cosine) with respect to the seed set and its successive extensions. The advantages of a method using limited resources include the following: a) it can be used for languages other than English for which resources such as parsers and annotated corpora are not available, but a part-of-speech tagger is; b) it can be used for English as well when fast and low cost development is required in specific sub-domains of subjective language.

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, flame recognition in email messages and others. A recent and extensive overview of current research work in the area of subjectivity analysis is provided by (Wiebe *et al.* 2002).

Related work

We can distinguish two connected directions in the research on subjectivity: a) methods for acquiring subjective expressions (Hatzivassiloglou & McKeown 1997); b) methods for classifying documents or sentences as subjective or not (Hatzivassiloglou & Wiebe 2000; Tong 2001; Pang, Lee, & Vaithyanathan 2002; Yu & Hatzivassiloglou 2003). Most often, both perspectives can overlap and be combined in the same work as successive steps, i.e., using lists of acquired subjective expressions for classifying documents or sentences (Wiebe 2000; Turney 2002; Riloff, Wiebe, & Wilson 2003).

This distinction corresponds to two qualitatively different types of approach. The need for acquisition algorithms arises from the fact that existing lexical databases such as Wordnet typically do not provide subjectivity classification, hence this information has to be created anew. These algorithms typically belong to the family of lexical acquisition procedures based on distributional similarity (Pereira, Tishby, & Lee 1993; Lin 1998; Weeds & Weir 2003). The basic idea is that distributionally similar words are also semantically related. In particular, the hypothesis has been explored that specific types of syntactic contexts convey specific types of semantic relationships, for the purpose of acquiring semantic lexicons. This has been used to learn hyponymy relationships from patterns of the type "NP, NP and other NP" (Hearst 1992), semantically related words from contexts like conjunctions, lists, appositives and nominal compounds (Riloff & Shepherd 1997; Roark & Charniak 1998; Phillips & Riloff 2002).

Classification algorithms typically rely on previously coded data (either manually annotated or pre-existing document metadata) at various levels (document, sentence, phrase, word), possibly making use of resources generated through acquisition methods. The availability of annotated data provides a substantial amount of knowledge that can be used to compute the predictive value of a large array of simple and complex features that can be used to train appropriate classifiers. Relevant features include: single words, phrases, n-grams, various types of collocations, unique and low-frequency words, verbs (Wiebe *et al.* 2002), adjectives and adjective subtypes, like gradable and semantically oriented adjectives (Wiebe 2000; Hatzivassiloglou & Wiebe 2000).

Both acquisition and classification methods often achieve very good results. However, they are substantially dependent on the availability of knowledge-intensive resources, like annotated data and other pre-processing tools. In this latter regard, for example, Riloff and her colleagues applied bootstrapping algorithms that use automatically generated

patterns in order to acquire subjective nouns. These algorithms do not require annotated data, but nevertheless they rely on the availability of a previously developed tool for learning extraction patterns in information extraction applications (Riloff & Wiebe 2003). Other acquisition methods rely on the use of parsers to identify the relevant syntactic contexts (Roark & Charniak 1998).

The appeal to some even superficial level of syntactic analysis in acquisition methods based on distributional similarity stems from the need to optimize the trade-off between reliability (precision) and coverage (recall) of the syntactic contexts being used. A syntactic context needs to be reliable, in the sense that it is regularly correlated with a given, specific semantic relationship. At the same time it should cover as many instances of the semantic information being acquired as possible. Therefore the use of tools like a shallow parser or an extraction pattern learner allows the identification of syntactic contexts that are specific enough to be reliably associated with the same semantic relationship, while coverage may be ensured by combining together different specific syntactic contexts (Phillips & Riloff 2002).

This paper presents a bootstrapping algorithm for the semi-automatic acquisition of subjective adjectives, akin to the acquisition methods based on distributional similarity that have been mentioned above. The focus of this work is to investigate the possibility of learning useful resources while at the same time reducing to a minimum the use of knowledge-based resources like annotated data and preprocessing tools. The proposed method only requires a part-of-speech tagger and a small set of seed adjectives. It does not require annotated data or parsers. The advantages of a method using limited resources include the following: a) it can be used for languages other than English for which resources such as parsers and annotated corpora are not available, but a part-of-speech tagger is; b) it can be used for English as well when fast and low cost development is required in specific sub-domains of subjective language.

Data and task definition

Two subsets of the Reuters collection from the American News Corpus were used. Most of the experiments were carried out on the smaller subset (1,200,000 words), which served as a development corpus to improve parameter tuning; the larger subset (4,800,000 million words) was used to check how corpus size affects the performance of the algorithm. Both corpora were tagged for part-of-speech information using an efficient implementation of the the Brill tagger (Ngai & Florian 2001).

The target is the acquisition of subjective adjectives. Adjectives are a well-known linguistic means to express point of view. In particular, (Bruce & Wiebe 1999) have shown a statistically significant positive correlation of adjectives with subjective sentences in a tagged corpus.

For the algorithm proposed here we needed a set of subjective adjectives a) to be used as gold standard for the evaluation of the results and b) to extract a subset to be used as seed set in order to bootstrap the learning process. Now, the domain of subjectivity has the same decidability difficulties as many other areas of natural language semantics, in the sense that it is often unclear whether to classify an expression as subjective or not. In particular, many expressions may be subjective in some contexts but not in others (Wiebe 1994). Studies in subjectivity tagging typically rely on more than one annotator and evaluate inter-annotator agreement (Wiebe *et al.* 2002).

For this work we used as a starting point the list of adjectives learned using the process presented in (Wiebe 2000)[1]. Two judges were instructed to manually select from the original list those adjectives that they would rate as subjective (even if not necessarily in all contexts) with a certainty level from medium to high. The goal was to obtain a high-quality data set for both seeding and evaluation. To give an idea of what this filtering process achieved, items like *administrative*, *Colombian*, *eighth*, *red* were excluded, whereas items like *worthy*, *trashy*, *superior*, *abysmal*, *enjoyable* were selected.

In this manner, a gold list of 332 subjective adjectives was obtained. In different experiments, different subsets of adjectives were used as seed sets, and the part that was not used as a seed set was used for testing, as detailed later. Two seed set sizes were used: the default size for most experiments was 35; 100 adjectives were used to control for the effects that seed set size has on the algorithm performance.

Given as input only a tagged corpus and a seed set, the bootstrapping algorithm yields as output a (much larger) ranked list of subjective adjectives. The algorithm falls within the family of lexical acquisition procedures based on distributional similarity.

However, the approach presented here exploits the correlation between syntactic context and semantic relationships under a slightly different angle than the previously mentioned methods of this kind. Such methods make the assumption that a particular syntactic configuration is a good heuristics to predict the semantic relationship among items that fill certain slots in that configuration, often quite independently of the specific meaning of words in those slots.

For example (Phillips & Riloff 2002) pay great attention to selecting reliable syntactic heuristics like a particular type of nominal compound that they call "GN PNP", where one or more general nouns modify a proper name, as in "violinist James Braum", "software maker Microsoft". These compounds can be used to identify terms that have the same immediate hypernym (e.g., "software maker Oracle"). The general problem here is that many syntactic contexts, while widespread in corpora, are also too generic to be reliably associated with a particular kind of semantic relationship. For example, as (Phillips & Riloff 2002) correctly note, nominal compounds are very common, but they also exhibit a wide variety of semantic relationships. Hence, in order to strike a balance between precision and coverage, it is necessary to choose more restrictive contexts (like the "GN PNP" phrases).

In order to address this problem that is typical of distribution-based methods, in this paper we shift the focus on the word meaning of the seed adjectives. We choose as syntactic context the sequence of an adjective and a noun.

[1] It is available at http://www.cs.pitt.edu/ wiebe/pubs/aaai00/.

This is a very generic context which can hardly be associated with any specific semantic relationship, besides very abstract and general notions like restricting (e.g., *tall, beautiful, intelligent*) or intersecting (e.g., *male*, *American*, *red*) modification (Keenan & Faltz 1985).

Instead of adding further restrictions to the chosen syntactic context, we make the hypothesis that subjective adjectives tend to modify nouns that are oriented towards subjectivity themselves, in the sense that they denote referents that easily lend themselves to be the object of some subjective attitude. For example, we expect that nouns like *book, movie, experience*, are more likely to be modified by subjective adjectives than nouns like *semiconductors*, *calipers*, *antifreeze*.

This hypothesis is operationalized as follows, in two main stages. In the first stage, a seed set of subjective adjectives is used to identify the set of all nouns that they modify. These nouns are expected to denote referents that may be object of a subjective attitude. Then the adjectives that modify those nouns and are not in the initial seed set are collected. In the second stage, the newly found adjectives are ranked by computing their average semantic similarity to the adjectives in the seed set. This procedure is repeated until a termination point is reached.

The discovery procedure

In this section we describe in detail the acquisition method step by step. In the initial data preparation phase, the raw text corpus is tokenized, tagged for part-of-speech, and the 200 most frequent words are removed. The purpose of part-of-speech tagging is only to identify sequences of adjective-noun pairs; once this is done no other syntactic information is used, and thus high frequency words can be discarded to reduce noise in the later stages.

Next, a lexical association measure is computed for all the bigrams consisting of an adjective and a noun. This step yields a ranking of the adjective noun pairs in terms of how closely associated they are. Experiments were carried out with different measures: log-likelihood ratio, mutual information, chi-square, and left Fisher coefficient, using the N-gram Statistics Package (Banerjee & Pedersen 2003). The best overall algorithm performance was obtained using the log-likelihood ratio; all the results mentioned later on are based on this measure. This concludes the one-off data preparation step.

At this point, the iterative procedure begins. For each noun the average lexical association value with the list of seed adjectives is computed, and the nouns are ranked according to such averages. This operation is intended to identify those nouns that most typically occur with subjective adjectives from the seed set. The use of association measures instead of pure frequency counts favors nouns that are strongly associated with subjective adjectives (i.e., that are likely to denote potentially subjective referents), even if their frequency counts are low or very low.

Then from the list of nouns ranked by average association value the top portion is selected. For the development corpus the best choice was to get the top 40 nouns. Performance deteriorates both with smaller and larger values; however, this value is dependent on corpus size. Next, from the full list of adjective noun pairs in the corpus all the adjectives are collected that modify the top portion of the noun list just obtained. According to the working hypothesis, these are candidate subjective adjectives, since they modify nouns that have been found to be closely associated with the subjective adjectives in the seed set.

At this point we have to decide which of these candidates should be classified as subjective adjectives. We hypothesize that the adjectives in the candidate list that are most likely to be subjective are those that are most *similar* to the adjectives in the seed set.

The key point here is how to interpret and implement the notion of similarity. The decision on whether to classify a candidate adjective as subjective or not is based on its similarity to the seed set as whole. This raises the issue of computing the similarity of a single word to a set of words. There are two important factors here. On one hand several options are possible as to how to carry out such one-to-many computation. On the other hand, holding the computation method fixed, it is reasonable to expect that the particular composition of the seed set will affect classification decisions. Moreover, different computation methods might be affected in different ways by seed sets with different internal composition.

Here, similarity is computed using the vector cosine measure. This is done by first collecting all bigrams such that a) one of the members is an adjective either in the seed set or in the candidate set; b) the two members co-occur in a window of 10 words in the same document. This window is the size of the context for which cosine similarity is computed.

Next, the cross-product of the seed set and the candidate set is generated and using the list of bigrams just described the vector-based cosine similarity for each pair seed-candidate is computed. In order to compute the similarity of each candidate adjective with respect to the entire seed set, we choose to calculate the average of the cosine values of each candidate with respect to all the seeds and then the candidates are ranked on the basis of cosine averages. In this way candidates that got high cosine values only with very few seeds should be winnowed out. This is a welcome consequence only under the assumption being made here that all the adjectives in the original seed set are equally relevant (i.e., the set is homogenous) to classify a candidate adjective as subjective or not. There are various ways to relax the dependency on this assumption, for example by discarding the lowest value(s) in computing the cosine averages.

At this point we have a list of candidate subjective adjectives ranked by cosine averages with respect to the seed set. At the very first iteration the top portion of this list (selected according to the criteria detailed later on) is added directly to the seed set. This new set of adjectives is used as seed set for the next iteration.

For all iterations after the first, the process is slightly different. The list of candidate adjectives ranked by cosine averages is merged with the list of candidates obtained up to the previous iteration and then the new list is re-ranked by cosine average. Only at this point is the top portion of the list selected to be added to the seed set for the next iteration.

This operation of merging and re-ranking is intended to add an extra level of control and filtering over the adjectives that make it to the seed set that starts each iteration. Since by design the original seed set of manually selected adjectives is expanded with those learned at each iteration, the merging and re-ranking steps ensure that at any point in time only adjectives with the highest cosine averages are added to the seed set, thus limiting the reduction of its quality. The first iteration is treated differently because there is no need to compare the cosine averages of the first batch of acquired adjectives with the original seed set, which should anyway be kept separate from the lists that are acquired at each iteration.

In different experiment sets we tried out various parameters to select the top portion of the candidates that are to be added to the seed set for the next iteration: a) different fractions of the candidate list were promoted to seeds: 5% or 10%; b) at each iteration, either the same fraction is always selected, or the initial value is decreased by some percentage: 1% or 3%, until the procedure terminates. The reason for the latter option is that using a smaller top portion of the candidate list at each iteration is a way to balance the overall decreasing quality of candidates, which is a typical drawback of bootstrapping algorithms.

At the end of the last iteration, the top portion of the candidate list, determined according to one of the parameter configurations above, is the final outcome of the learning process.

Results and evaluation

In evaluating the results of the learning algorithm one cannot rely on the same methods used by annotation-based approaches, since no annotated data are available to begin with. In order to have an objective benchmark against which to compare outcomes of different experiments, we used for testing the portion of the initial list that was not used for the seed set. This choice is clearly not optimal, since there might well be other subjective adjectives in the corpus that are not present in the test set. Hence the results we give in terms of precision and recall with respect to such test set probably represent a lower bound of the performance of the algorithm.

In this context recall and precision are defined as follows:

$$recall = \frac{output\ set \cap test\ set}{test\ set}$$

$$precision = \frac{output\ set \cap test\ set}{output\ set}$$

where *output set* is the final set of adjectives generated by the learning procedure, and *test set* is the part of the manually selected subjective adjective list that excludes the seeds (297 adjectives). Several experiments were run with different parameter settings, and for each setting precision, recall and F-measure were computed. Note that after each iteration recall and precision were computed for the overall set of adjectives discovered up to that point, and excluding the original seed set.

As a baseline, we take the score obtained by the trivial acceptor, treating all the adjectives in the corpus as subjective, which yields a recall of 1.000 and a precision of 0.083, for an F-measure of 0.153. The choice of this baseline as a meaningful benchmark makes sense with respect to the nature of the learning task and to the difficulty of evaluating it. The point is that the use of recall and precision scores as explained above was dictated by the need of some form of objective measurement in driving development. However, those scores are not necessarily a good indicator of the quality of the results obtained, since the learning procedure outputs many adjectives which are not present in the test set but which we would intuitively judge as subjective.

So, for example, the second learning iteration for the set of seed adjectives with the lowest frequency[2] in the corpus yields the following candidate set, where italicized adjectives are missing from the test set against which recall and precision scores are computed:

> *fickle*, *uphill*, *fruitful*, phenomenal, dangerous, *speedy*, *unpredictable*, gloomy, *extensive*, enormous, skilled, *staple*, tremendous, *sustainable*, *seasonal*, *protracted*, *wealthy promotable*, urgent, *unenthusiastic*, *legendary*, *political*, *social*, *staid*, excessive, *timely*, fundamental, *technological*, *apparent*, *multinational*, *rapid*, *unrevised*, *lukewarm*, *inevitable*, *feverish* and *monetary*

However, it seems that several of the missing adjectives should be considered subjective (possibly at different levels). Also, recall that many adjectives may be interpreted as subjective in some contexts but not in others.

More generally, what this suggests is that recall is much more difficult to assess than precision, and at the same time less relevant. The problem is that we do not know in advance what the subjective adjectives of a language are (and in many cases the classification is context-dependent anyway), so that there might always be some subjective adjectives that the algorithm might uncover and that yet are not present in the test set. On the other hand, as the size of the test set increases, the precision score becomes more and more reliable.

For this reasons, two human judges were asked to rate the degree of subjectivity of all adjectives in the output set obtained from the lowest frequency seeds after five iterations (in order to have a sample large enough) using a three-value scale: subjective (level 1), possibly subjective (level2), non-subjective (level 3). Then the adjectives rated subjective or possibly subjective were added to the test set, and the scores were recomputed. Table 1 shows the scores of the first four iterations for the original test set and for its increasing extensions with level 1 and level 2 adjectives.

The scores for the extended sets show slightly better recall, and a much more significant increase in precision. We conclude from this discussion that, from the point of view of the effectiveness of the proposed algorithm, precision is more important than recall, and thus precision should be the main score to look at in assessing the quality of the results. Also, from now on, the recall and precision scores we pro-

[2]We will discuss this choice in detail later on.

Table 1: Best 10%, seed set of adjectives with lowest frequency

	With no extra adj.				With level 1 adj.				With level 1 and 2 adj.			
Iteration	1st	2nd	3rd	4th	1st	2nd	3rd	4th	1st	2nd	3rd	4th
Recall	0.020	0.030	0.040	0.050	0.026	0.045	0.068	0.081	0.030	0.064	0.098	0.131
Precision	0.353	0.250	0.200	0.180	0.470	0.388	0.350	0.301	0.588	0.583	0.533	0.518
F-measure	0.038	0.054	0.067	0.078	0.049	0.081	0.114	0.128	0.058	0.116	0.165	0.210

vide are based on the larger extended version (level 1 and level 2 adjectives).

We will discuss now how different parameter settings affect the outcome of the algorithm. First of all, experiments were run in two versions: always selecting the same top portion of the candidate list (10% or 5%) or selecting a smaller portion at each iteration (e.g., first 10%, then 9%, then 8% etc.). The second version always outperformed the first, suggesting that the candidate quality decreases quite rapidly. In the second version we obtained somewhat complementary results when starting from 10% and 5% (with a reduction of one percentage point for both): in the first case we get lower precision scores, but many more adjectives are acquired (83 at the fourth iteration, see Table 1 for scores); in the second case we get higher precision, but fewer adjectives are acquired (22 at the fourth iteration, see Table 2 for scores).

We also tried a reduction of three percentage points at a time, starting from 10%, obtaining precision scores that are just slightly better than with a reduction of only 1 percentage point at a time, but acquiring much fewer adjectives (only 37 at the fourth iteration, see Table 3). The best overall balance is then obtained starting with the top 10% portion and reducing it by one percentage point at every iteration. In all the other experiments reported from now on we use these parameters.

Table 2: Best 5%, lowest frequency

Iteration	1st	2nd	3rd	4th
Recall	0.021	0.037	0.046	0.049
Precision	0.875	0.705	0.714	0.727
F-measure	0.041	0.070	0.086	0.091

Table 3: Best 10%, reduction by 3 pct. points

Iteration	1st	2nd	3rd	4th
Recall	0.030	0.055	0.064	0.067
Precision	0.588	0.620	0.583	0.594
F-measure	0.058	0.101	0.116	0.121

Now we move on to consider parameters pertaining to the internal composition of the seed set. This is the most important factor in determining the outcome of the algorithm. The single most relevant dimension turned out to be frequency and its correlation with the semantic specificity of the seed adjectives.

We compared the results obtained using as seed set the adjectives with the highest, random and lowest frequency, keeping every other parameter constant. As can be seen from Table 4, the seeds with lowest frequency got the best scores, and those with the highest got the worst[3], with the random selection somewhat in the middle.

On one hand this is in line with the observation by (Wiebe *et al.* 2002) that low frequency words and subjectivity are strongly correlated, in the sense that the former can be used as features for predicting the latter. (Wiebe *et al.* 2002) suggest that this correlation is due to the fact that people are creative when they express their opinion, and thus may use unusual or rare words to do so. However, in the context of this work, a closer look at the actual seed sets, as reported in Box 1 and 2, suggests an alternative or complementary insight.

Box 1: Highest frequency seeds

good right important poor significant hard positive competitive great serious successful bad tough popular powerful solid volatile fair sure responsible normal severe healthy dramatic fine safe true hot cheap appropriate crucial unfair sophisticated essential surprising

Box 2: Lowest frequency seeds

frivolous flashy dismal clever astonishing versatile unskilled unseemly unorthodox unforgiving unbelievable unbearable trivial sinister regrettable profane poisonous obsessive neat mundane lively intolerable informative imperfect horrendous hollow heartless glamorous fancy enjoyable eclectic ebullient brutal breathtaking bland

From the point of view of the algorithm presented here we can interpret this result as suggesting that high frequency subjective adjectives such as *good*, *right*, *important*, *poor*, *significant*, *great*, *bad*, etc. are too generic in meaning to be strongly associated with nouns that may typically denote referents that are object of a subjective attitude, and thus they cannot work as good seeds to help identify new subjective adjectives. However, the low frequency seeds like *dismal*, *intolerable*, *sinister*, *flashy*, etc. are more specialized exactly in a subjective direction, so that a noun that is modified by one of these adjectives may be more likely to refer to something that can be the object of a subjective attitude, and thus low frequency adjectives work as better seeds for the task at stake here.

[3]Recall and precision were both 0 using the original test set (without level 1 and level 2 adjectives) for evaluation.

Table 4: Scores by seed frequency, extended test set

	Highest frequency				Random frequency				Lowest Frequency			
Iteration	1st	2nd	3rd	4th	1st	2nd	3rd	4th	1st	2nd	3rd	4th
Recall	0.012	0.015	0.015	0.015	0.023	0.036	0.043	0.051	0.030	0.064	0.098	0.131
Precision	0.137	0.111	0.090	0.078	0.220	0.198	0.170	0.158	0.588	0.583	0.533	0.518
F-measure	0.022	0.027	0.026	0.025	0.042	0.061	0.068	0.077	0.058	0.116	0.165	0.210

Frequency turned out to be more important than other dimensions of meaning we explored. One potential problem with using the lowest frequency seeds might be that it is not semantically homogeneous, for example in the sense that it might contain adjectives of both positive and negative orientation (Hatzivassiloglou & McKeown 1997). Given the low frequency itself of the seeds, one might expect that this might be a source of noise and thus have adverse effects on the final results.

For this reasons, we ran the algorithm using seed sets of manually selected positive and negative adjectives, in both cases starting to choose from those with the lowest frequency. However, in both cases the result of the selection was that the two seed sets, while semantically homogeneous in a positive or negative sense, also contained adjectives that were not any more those with the lowest frequency. Somewhat surprisingly, the scores for these experiments are inferior to those obtained by the lowest frequency seeds (see Table 5 and 6), suggesting that frequency is more relevant than the basic level of semantic homogeneity given by polarity orientation.

Table 5: Best 10%, seed set of positive adjectives

Iteration	1st	2nd	3rd	4th
Recall	0.015	0.042	0.061	0.067
Precision	0.357	0.358	0.333	0.318
F-measure	0.029	0.076	0.101	0.111

Table 6: Best 10%, seed set of negative adjectives

Iteration	1st	2nd	3rd	4th
Recall	0.024	0.036	0.052	0.061
Precision	0.380	0.292	0.223	0.224
F-measure	0.046	0.065	0.084	0.096

Next, we tried to assess the relevance of frequency under a different angle. We ran another experiment using a seed set of adjectives that occur at least 3 times. This choice was meant to select a seed set of low but not lowest frequency. The hypothesis was that the slightly higher number of occurrences might help identify a larger number of relevant nouns, while the low frequency might still be enough to guarantee a good quality of final results. The scores, as reported in Table 7, show that this is not the case, and once again the best scores are those obtained with the lowest frequency seeds.

Finally we also controlled the effects of seed set size, by running an experiment with a set of 100 seed adjectives (selected in order of increasing frequency), obtaining the scores in Table 8. In this case precision is even more relevant than usual with respect to recall, since increasing the seed set size reduces significantly the size of the test test against which scores are computed (since we always use the same set of manually selected gold adjectives to get seed set and test set). Apart from this, what is more interesting is that the 100-seed set allows the acquisition of fewer adjectives than the usual 35-seed set: 44 for the former against 83 for the latter, after the fourth iteration. This might be an indication that a seed set that is too large introduces too much dispersion, since in this case for an adjective to be classified as subjective it must exhibit similarity to 100 seeds, and getting high cosine averages is likely to be more difficult than in the case of the 35-seed set.

Table 7: Best 10%, seeds with frequency greater than 2

Iteration	1st	2nd	3rd	4th
Recall	0.009	0.021	0.040	0.067
Precision	0.200	0.212	0.240	0.247
F-measure	0.017	0.039	0.068	0.106

Table 8: Best 10%, seed set of 100 adjectives

Iteration	1st	2nd	3rd	4th
Recall	0.031	0.043	0.055	0.059
Precision	0.421	0.407	0.424	0.428
F-measure	0.058	0.078	0.097	0.104

The most surprising conclusion is then that the proposed algorithm obtains consistently the best performance using the lowest frequency seed set. Of this set, as reported on Box 2 above, the first 7 adjectives occur 2 times in the corpus, and the remaining 28 occur once. These are very low counts, and the good results they yield are at odds with the general tendency to discard low frequency words that is common in natural language processing and especially information retrieval (Weeber, Vos, & Bayeen 2000). For example, mutual information gives good results in lexical association measurement tasks with word frequencies above 5. (Church & Hanks 1990).

The next natural step is to verify whether the lowest frequency seeds continue to yield the best results when other parameters that we did not control for so far are changed. First of all, the size of the corpus is important in determining the meaning that frequency has. It is to be expected that in a small enough corpus many unique words, or in any event words with low frequency, are actually "well-established" in language use, in the sense that they might be somewhat less

common words but still part of the general competence of language speakers. This might be a very good condition for the proposed algorithm to be effective, since low frequency words in a small corpus might have a degree of specificity that allows them to have good discriminatory power in identifying nouns that can lead to good candidates.

However, as the corpus size increases, word frequency distributions tend to yield a very long tail of unique words that include more and more technical terms, neologisms and misspellings (Baayen 2001). Under these circumstances lowest frequency seeds might not yield the optimal results any more.

For these reasons we ran an experiment on a larger corpus of 4,800,000 words, using as seed set the lowest frequency adjectives for this corpus that are also in the manually selected set of 332 adjectives discussed earlier on. This seed set actually has many adjectives in common with the seed set for the small corpus, since in both cases we decided by design to get them from the set of manually selected adjectives. However, the result of this choice is that in the seed set for the large corpus there are 10 adjectives of frequency 1, 9 of frequency 2, 13 of frequency 3 and 3 of frequency 4, i.e., these are not the lowest frequency adjectives in the corpus anymore. In fact, out of about 53,000 word types, about 16,000 have frequency 1 (30%), and we did not choose the seeds from this latter group, since this would have required a manual inspection and rating of all the adjectives in this group. Therefore, in practice the seed set used did not really contain the lowest frequency subjective adjectives for the corpus the algorithm was run on.

Nevertheless, it is interesting to compare the results, as reported in Table 9, to those obtained with the lowest frequency seed set on the small corpus, as reported in Table 1. The main observation is that on the large corpus recall increases and precision decreases with respect to the small corpus, with the F-measure slightly better across the board. The main reason for this results is probably another effect of corpus size: on the large corpus the algorithm acquires many more adjectives (359 against 83 on the small corpus, after the fourth iteration).

Limits of the method and future directions

(Wiebe 1994) is one of the earliest computational papers on subjectivity in natural language. It sets the stage for subsequent work concerning algorithms for acquisition and classification of subjective language resources.

As mentioned at the outset, the goal of many studies is to build a classifier for subjectivity tagging using features extracted from annotated data. However, the procedure we proposed here is more similar to algorithms like Basilisk (Thelen & Riloff 2002). However, Basilisk has access to a greater array of syntactic relations since it uses an auxiliary extraction pattern learner which can generate information extraction-style patterns in order to identify every noun phrase in the corpus. Here we use only adjacent adjective noun pairs without any additional tools beside the part-of-speech tagger, since our goal is to see how far we can go in learning subjective adjectives by relying on minimal resources.

Therefore the important question that arises here is how far the proposed resource-poor method can go, compared to approaches that rely on richer knowledge sources. The first issue is that the same property that allows the algorithm to work is also its main limitation: the acquisition methods crucially relies on the strict adjacency of adjective noun pairs, i.e., bypassing the employment of a parser or similar tool is possible only as long as purely linear relationships like immediate adjacency are exploited. So for example, it would not be possible to extend the approach, in its present form, to verb-direct object pairs for the purpose of learning subjective verbs, since this would require the identification of a noun phrase and its head noun (besides the direct object grammatical relationship within the verb phrase), and therefore the recourse to higher level syntactic information.

More in general, the algorithm cannot be used for all those cases (which are the typical focus of distribution-based methods) in which a particular semantic relationship that is the goal of the acquisition procedure is associated with a syntactic relationship that goes beyond strict adjacency. However, the modification relationship between adjectives and noun is probably the most relevant in the domain of subjectivity, given the crucial role that adjectives play in expressing a point of view. Now, the experiments discussed above show that the proposed algorithm is good at generating automatically quality candidates, which, after validation through a minimal human effort, can provide low cost lexicons of subjective adjectives. This can be very valuable for several languages other than English for which parsers or similar processing tools are not available. Since only a part-of-speech tagger is required, for many languages this might be the very first attempt towards the automated construction of a lexicon of subjective adjectives.

Another area that needs further elaboration is the evaluation procedure. In this paper the computation of the recall and precision scores was based on a kind of "closed world assumption", whereby the only adjectives accepted as subjective for evaluation purposes are those inserted in the seed set. Of course, many legitimate subjective adjectives are missing, and actually the output sets of the algorithm itself might help to integrate the original list. A larger test set constructed in this way should be able to provide a more precise indication of the effectiveness of the algorithm. Moreover, there are at least two classes of adjectives that need further attention: those that might be interpreted as subjective in some contexts but not in others, and those whose status is intrinsically uncertain. Using ratings by human judges and controlling for their agreement will help in this latter direction.

References

Baayen, R. H. 2001. *Word Frequency Distributions*. Dordrecht: Kluwer.

Banerjee, S., and Pedersen, T. 2003. The design, implementation, and use of the Ngram Statistics Package. In *Proceedings of the Fourth International Conference on Intelligent Text Processing and Computational Linguistics*.

Bruce, R., and Wiebe, J. 1999. Recognizing subjectivity:

Table 9: Best 10%, seed set of adjectives with lowest frequency, larger corpus

	With no extra adj.				With level 1 adj.				With level 1 and 2 adj.			
Iteration	1st	2nd	3rd	4th	1st	2nd	3rd	4th	1st	2nd	3rd	4th
Recall	0.020	0.053	0.124	0.185	0.026	0.071	0.140	0.198	0.024	0.070	0.138	0.220
Precision	0.136	0.120	0.158	0.153	0.181	0.165	0.183	0.169	0.181	0.173	0.192	0.200
F-measure	0.033	0.074	0.139	0.167	0.045	0.100	0.158	0.183	0.043	0.100	0.160	0.210

a case study in manual tagging. *Natural Language Engineering* 5(2).

Church, K. W., and Hanks, P. 1990. Word association norms mutual information, and lexicography. *Computational Linguistics* 16(1).

Hatzivassiloglou, V., and McKeown, K. R. 1997. Predicting the semantic orientation of adjectives. In *35th Annual Meeting of the Association for Computational Linguistics (ACL-97)*.

Hatzivassiloglou, V., and Wiebe, J. 2000. Effects of adjective orientation and gradability on sentence subjectivity. In *Proc. of International Conference on Computational Linguistics (COLING-2000)*.

Hearst, M. 1992. Automatic acquisition of hyponyms from large text corpora. In *Proc. of the 14th International Conference on Computational Linguistics (COLING-1992)*.

Keenan, E. L., and Faltz, L. 1985. *Boolean Semantics for Natural Language*. Dordrecht: Reidel.

Lin, D. 1998. Automatic retrieval and clustering of similar words. In *Proceedings of COLING-ACL98*.

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

Pang, B.; Lee, L.; and Vaithyanathan, S. 2002. Thumbs up? Sentiment classification using machine learning techniques. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002)*.

Pereira, F.; Tishby, N.; and Lee, L. 1993. Distributional clustering of English words. In *31st Annual Meeting of the Association for Computational Linguistics (ACL-93)*.

Phillips, W., and Riloff, E. 2002. Exploiting strong syntactic heuristics and co-training to learn semantic lexicons. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002)*.

Riloff, E., and Shepherd, J. 1997. A corpus-based approach for building semantic lexicons. In *Proc. Conference on Empirical Methods in Natural Language Processing (EMNLP-97)*.

Riloff, E., and Wiebe, J. 2003. Learning extraction patterns for subjective expressions. In *Proc. Conference on Empirical Methods in Natural Language Processing (EMNLP-03)*.

Riloff, E.; Wiebe, J.; and Wilson, T. 2003. Learning subjective nouns using extraction pattern bootstrapping. In *Proceedings of the Seventh CoNLL conference held at HLT-NAACL 2003*.

Roark, B., and Charniak, E. 1998. Noun phrase co-occurrence statistics for semi-automatic semantic lexicon construction. In *36th Annual Meeting of the Association for Computational Linguistics (ACL-98)*.

Thelen, M., and Riloff, E. 2002. A bootstrapping method for learning semantic lexicons using extraction pattern contexts. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002)*.

Tong, R. M. 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*.

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 (ACL-02)*.

Weeber, M.; Vos, R.; and Bayeen, R. H. 2000. Extracting the lowest frequency words: Pitfalls and possibilities. *Computational Linguistics* 26(3).

Weeds, J., and Weir, D. 2003. A general framework for distributional similarity. In *Proc. Conference on Empirical Methods in Natural Language Processing (EMNLP-03)*.

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

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

Wiebe, J. 2000. Learning subjective adjectives from corpora. In *Proc. 17th National Conference on Artificial Intelligence (AAAI-2000)*.

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