Predicting the Semantic Orientation of Adjectives

Vasileios Hatzivassiloglou and **Kathleen R. McKeown**
Department of Computer Science
450 Computer Science Building
Columbia University
New York, N.Y. 10027, USA
{vh, kathy}@cs.columbia.edu

Abstract

We identify and validate from a large corpus constraints from conjunctions on the positive or negative semantic orientation of the conjoined adjectives. A log-linear regression model uses these constraints to predict whether conjoined adjectives are of same or different orientations, achieving 82% accuracy in this task when each conjunction is considered independently. Combining the constraints across many adjectives, a clustering algorithm separates the adjectives into groups of different orientations, and finally, adjectives are labeled positive or negative. Evaluations on real data and simulation experiments indicate high levels of performance: classification precision is more than 90% for adjectives that occur in a modest number of conjunctions in the corpus.

1 Introduction

The semantic *orientation* or *polarity* of a word indicates the direction the word deviates from the norm for its semantic group or *lexical field* (Lehrer, 1974). It also constrains the word's usage in the language (Lyons, 1977), due to its *evaluative* characteristics (Battistella, 1990). For example, some nearly synonymous words differ in orientation because one implies desirability and the other does not (e.g., *simple* versus *simplistic*). In linguistic constructs such as conjunctions, which impose constraints on the semantic orientation of their arguments (Anscombre and Ducrot, 1983; Elhadad and McKeown, 1990), the choices of arguments and connective are mutually constrained, as illustrated by:

The tax proposal was
$$\left\{\begin{array}{l}\text{simple and well-received}\\ \text{simplistic but well-received}\\ \text{*simplistic and well-received}\end{array}\right\}$$
by the public.

In addition, almost all antonyms have different semantic orientations.[1] If we know that two words relate to the same property (for example, members of the same scalar group such as *hot* and *cold*) but have different orientations, we can usually infer that they are antonyms. Given that semantically similar words can be identified automatically on the basis of distributional properties and linguistic cues (Brown et al., 1992; Pereira et al., 1993; Hatzivassiloglou and McKeown, 1993), identifying the semantic orientation of words would allow a system to further refine the retrieved semantic similarity relationships, extracting antonyms.

Unfortunately, dictionaries and similar sources (theusari, WordNet (Miller et al., 1990)) do not include semantic orientation information.[2] Explicit links between antonyms and synonyms may also be lacking, particularly when they depend on the domain of discourse; for example, the opposition *bear–bull* appears only in stock market reports, where the two words take specialized meanings.

In this paper, we present and evaluate a method that automatically retrieves semantic orientation information using indirect information collected from a large corpus. Because the method relies on the corpus, it extracts domain-dependent information and automatically adapts to a new domain when the corpus is changed. Our method achieves high precision (more than 90%), and, while our focus to date has been on adjectives, it can be directly applied to other word classes. Ultimately, our goal is to use this method in a larger system to automatically identify antonyms and distinguish near synonyms.

2 Overview of Our Approach

Our approach relies on an analysis of textual corpora that correlates linguistic features, or *indicators*, with

[1] Exceptions include a small number of terms that are both negative from a pragmatic viewpoint and yet stand in an antonymic relationship; such terms frequently lexicalize two unwanted extremes, e.g., *verbose–terse*.

[2] Except implicitly, in the form of definitions and usage examples.

semantic orientation. While no direct indicators of positive or negative semantic orientation have been proposed[3], we demonstrate that conjunctions between adjectives provide indirect information about orientation. For most connectives, the conjoined adjectives usually are of the same orientation: compare *fair and legitimate* and *corrupt and brutal* which actually occur in our corpus, with **fair and brutal* and **corrupt and legitimate* (or the other cross-products of the above conjunctions) which are semantically anomalous. The situation is reversed for *but*, which usually connects two adjectives of different orientations.

The system identifies and uses this indirect information in the following stages:

1. All conjunctions of adjectives are extracted from the corpus along with relevant morphological relations.
2. A log-linear regression model combines information from different conjunctions to determine if each two conjoined adjectives are of same or different orientation. The result is a graph with hypothesized same- or different-orientation links between adjectives.
3. A clustering algorithm separates the adjectives into two subsets of different orientation. It places as many words of same orientation as possible into the same subset.
4. The average frequencies in each group are compared and the group with the higher frequency is labeled as positive.

In the following sections, we first present the set of adjectives used for training and evaluation. We next validate our hypothesis that conjunctions constrain the orientation of conjoined adjectives and then describe the remaining three steps of the algorithm. After presenting our results and evaluation, we discuss simulation experiments that show how our method performs under different conditions of sparseness of data.

3 Data Collection

For our experiments, we use the 21 million word 1987 Wall Street Journal corpus[4], automatically annotated with part-of-speech tags using the PARTS tagger (Church, 1988).

In order to verify our hypothesis about the orientations of conjoined adjectives, and also to train and evaluate our subsequent algorithms, we need a set of adjectives with predetermined orientation labels. We constructed this set by taking all adjectives appearing in our corpus 20 times or more, then removing adjectives that have no orientation. These are typically members of groups of complementary, qualitative terms (Lyons, 1977), e.g., *domestic* or *medical*.

Positive:	adequate central clever famous intelligent remarkable reputed sensitive slender thriving
Negative:	contagious drunken ignorant lanky listless primitive strident troublesome unresolved unsuspecting

Figure 1: Randomly selected adjectives with positive and negative orientations.

We then assigned an orientation label (either + or −) to each adjective, using an evaluative approach. The criterion was whether the use of this adjective ascribes in general a positive or negative quality to the modified item, making it better or worse than a similar unmodified item. We were unable to reach a unique label out of context for several adjectives which we removed from consideration; for example, *cheap* is positive if it is used as a synonym of *inexpensive*, but negative if it implies inferior quality. The operations of selecting adjectives and assigning labels were performed before testing our conjunction hypothesis or implementing any other algorithms, to avoid any influence on our labels. The final set contained 1,336 adjectives (657 positive and 679 negative terms). Figure 1 shows randomly selected terms from this set.

To further validate our set of labeled adjectives, we subsequently asked four people to independently label a randomly drawn sample of 500 of these adjectives. They agreed with us that the positive/negative concept applies to 89.15% of these adjectives on average. For the adjectives where a positive or negative label was assigned by both us and the independent evaluators, the average agreement on the label was 97.38%. The average inter-reviewer agreement on labeled adjectives was 96.97%. These results are extremely significant statistically and compare favorably with validation studies performed for other tasks (e.g., sense disambiguation) in the past. They show that positive and negative orientation are objective properties that can be reliably determined by humans.

To extract conjunctions between adjectives, we used a two-level finite-state grammar, which covers complex modification patterns and noun-adjective apposition. Running this parser on the 21 million word corpus, we collected 13,426 conjunctions of adjectives, expanding to a total of 15,431 conjoined adjective pairs. After morphological trans-

[3] Certain words inflected with negative affixes (such as *in-* or *un-*) tend to be mostly negative, but this rule applies only to a fraction of the negative words. Furthermore, there are words so inflected which have positive orientation, e.g., *independent* and *unbiased*.

[4] Available form the ACL Data Collection Initiative as CD ROM 1.

Conjunction category	Conjunction types analyzed	% same-orientation (types)	% same-orientation (tokens)	P-Value (for types)
All conjunctions	2,748	77.84%	72.39%	$< 1 \cdot 10^{-16}$
All *and* conjunctions	2,294	81.73%	78.07%	$< 1 \cdot 10^{-16}$
All *or* conjunctions	305	77.05%	60.97%	$< 1 \cdot 10^{-16}$
All *but* conjunctions	214	30.84%	25.94%	$2.09 \cdot 10^{-8}$
All attributive *and* conjunctions	1,077	80.04%	76.82%	$< 1 \cdot 10^{-16}$
All predicative *and* conjunctions	860	84.77%	84.54%	$< 1 \cdot 10^{-16}$
All appositive *and* conjunctions	30	70.00%	63.64%	0.04277

Table 1: Validation of our conjunction hypothesis. The P-value is the probability that similar or more extreme results would have been obtained if same- and different-orientation conjunction types were actually equally distributed.

formations, the remaining 15,048 conjunction *tokens* involve 9,296 distinct pairs of conjoined adjectives (*types*). Each conjunction token is classified by the parser according to three variables: the conjunction used (*and*, *or*, *but*, *either-or*, or *neither-nor*), the type of modification (attributive, predicative, appositive, resultative), and the number of the modified noun (singular or plural).

4 Validation of the Conjunction Hypothesis

Using the three attributes extracted by the parser, we constructed a cross-classification of the conjunctions in a three-way table. We counted types and tokens of each conjoined pair that had both members in the set of pre-selected labeled adjectives discussed above; 2,748 (29.56%) of all conjoined pairs (types) and 4,024 (26.74%) of all conjunction occurrences (tokens) met this criterion. We augmented this table with marginal totals, arriving at 90 categories, each of which represents a triplet of attribute values, possibly with one or more "don't care" elements.

We then measured the percentage of conjunctions in each category with adjectives of same or different orientations. Under the null hypothesis of same proportions of adjective pairs (types) of same and different orientation in a given category, the number of same- or different-orientation pairs follows a binomial distribution with $p = 0.5$ (Conover, 1980).

We show in Table 1 the results for several representative categories, and summarize all results below:

- Our conjunction hypothesis is validated overall and for almost all individual cases. The results are extremely significant statistically, except for a few cases where the sample is small.
- Aside from the use of *but* with adjectives of different orientations, there are, rather surprisingly, small differences in the behavior of conjunctions between linguistic environments (as represented by the three attributes). There are a few exceptions, e.g., appositive *and* conjunctions modifying plural nouns are evenly split between same and different orientation. But in these exceptional cases the sample is very small, and the observed behavior may be due to chance.
- Further analysis of different-orientation pairs in conjunctions other than *but* shows that conjoined antonyms are far more frequent than expected by chance, in agreement with (Justeson and Katz, 1991).

5 Prediction of Link Type

The analysis in the previous section suggests a baseline method for classifying links between adjectives: since 77.84% of all links from conjunctions indicate same orientation, we can achieve this level of performance by always guessing that a link is of the same-orientation type. However, we can improve performance by noting that conjunctions using *but* exhibit the opposite pattern, usually involving adjectives of different orientations. Thus, a revised but still simple rule predicts a different-orientation link if the two adjectives have been seen in a *but* conjunction, and a same-orientation link otherwise, assuming the two adjectives were seen connected by at least one conjunction.

Morphological relationships between adjectives also play a role. Adjectives related in form (e.g., *adequate–inadequate* or *thoughtful–thoughtless*) almost always have different semantic orientations. We implemented a morphological analyzer which matches adjectives related in this manner. This process is highly accurate, but unfortunately does not apply to many of the possible pairs: in our set of 1,336 labeled adjectives (891,780 possible pairs), 102 pairs are morphologically related; among them, 99 are of different orientation, yielding 97.06% accuracy for the morphology method. This information is orthogonal to that extracted from conjunctions: only 12 of the 102 morphologically related pairs have been observed in conjunctions in our corpus. Thus, we

Prediction method	Morphology used?	Accuracy on reported same-orientation links	Accuracy on reported different-orientation links	Overall accuracy
Always predict same orientation	No	77.84%	–	77.84%
	Yes	78.18%	97.06%	78.86%
But rule	No	81.81%	69.16%	80.82%
	Yes	82.20%	78.16%	81.75%
Log-linear model	No	81.53%	73.70%	80.97%
	Yes	82.00%	82.44%	82.05%

Table 2: Accuracy of several link prediction models.

add to the predictions made from conjunctions the different-orientation links suggested by morphological relationships.

We improve the accuracy of classifying links derived from conjunctions as same or different orientation with a log-linear regression model (Santner and Duffy, 1989), exploiting the differences between the various conjunction categories. This is a generalized linear model (McCullagh and Nelder, 1989) with a linear predictor

$$\eta = \mathbf{w}^{\mathrm{T}}\mathbf{x}$$

where $\mathbf{x}$ is the vector of the observed counts in the various conjunction categories for the particular adjective pair we try to classify and $\mathbf{w}$ is a vector of weights to be learned during training. The response y is non-linearly related to η through the *inverse logit* function,

$$y = \frac{e^{\eta}}{1+e^{\eta}}$$

Note that $y \in (0,1)$, with each of these endpoints associated with one of the possible outcomes.

We have 90 possible predictor variables, 42 of which are linearly independent. Since using all the 42 independent predictors invites overfitting (Duda and Hart, 1973), we have investigated subsets of the full log-linear model for our data using the method of *iterative stepwise refinement*: starting with an initial model, variables are added or dropped if their contribution to the reduction or increase of the residual deviance compares favorably to the resulting loss or gain of residual degrees of freedom. This process led to the selection of nine predictor variables.

We evaluated the three prediction models discussed above with and without the secondary source of morphology relations. For the log-linear model, we repeatedly partitioned our data into equally sized training and testing sets, estimated the weights on the training set, and scored the model's performance on the testing set, averaging the resulting scores.[5] Table 2 shows the results of these analyses. Although the log-linear model offers only a small improvement on pair classification than the simpler *but* prediction rule, it confers the important advantage of *rating* each prediction between 0 and 1. We make extensive use of this in the next phase of our algorithm.

[5]When morphology is to be used as a supplementary predictor, we remove the morphologically related pairs from the training and testing sets.

6 Finding Groups of Same-Oriented Adjectives

The third phase of our method assigns the adjectives into groups, placing adjectives of the same (but unknown) orientation in the same group. Each pair of adjectives has an associated dissimilarity value between 0 and 1; adjectives connected by same-orientation links have low dissimilarities, and conversely, different-orientation links result in high dissimilarities. Adjective pairs with no connecting links are assigned the neutral dissimilarity 0.5.

The baseline and *but* methods make qualitative distinctions only (i.e., same-orientation, different-orientation, or unknown); for them, we define dissimilarity for same-orientation links as one minus the probability that such a classification link is correct and dissimilarity for different-orientation links as the probability that such a classification is correct. These probabilities are estimated from separate training data. Note that for these prediction models, dissimilarities are identical for similarly classified links.

The log-linear model, on the other hand, offers an estimate of how good each prediction is, since it produces a value y between 0 and 1. We construct the model so that 1 corresponds to same-orientation, and define dissimilarity as one minus the produced value.

Same and different-orientation links between adjectives form a graph. To partition the graph nodes into subsets of the same orientation, we employ an iterative optimization procedure on each connected component, based on the *exchange method*, a non-hierarchical clustering algorithm (Späth, 1985). We define an *objective function* Φ scoring each possible partition $\mathcal{P}$ of the adjectives into two subgroups C_1 and C_2 as

$$\Phi(\mathcal{P}) = \sum_{i=1}^{2}\left(\frac{1}{|C_i|}\sum_{\substack{x,y\,\in\, C_i\\ x\neq y}} d(x,y)\right)$$

α	Number of adjectives in test set ($\vert A_\alpha \vert$)	Number of links in test set ($\vert L_\alpha \vert$)	Average number of links for each adjective	Accuracy	Ratio of average group frequencies
2	730	2,568	7.04	78.08%	1.8699
3	516	2,159	8.37	82.56%	1.9235
4	369	1,742	9.44	87.26%	1.3486
5	236	1,238	10.49	92.37%	1.4040

Table 3: Evaluation of the adjective classification and labeling methods.

where $|C_i|$ stands for the cardinality of cluster i, and $d(x, y)$ is the dissimilarity between adjectives x and y. We want to select the partition $\mathcal{P}_{\min}$ that minimizes Φ, subject to the additional constraint that for each adjective x in a cluster C,

$$\frac{1}{|C|-1} \sum_{\substack{y \in C \\ x \neq y}} d(x,y) \quad < \quad \frac{1}{|\overline{C}|} \sum_{y \in \overline{C}} d(x,y) \qquad (1)$$

where $\overline{C}$ is the complement of cluster C, i.e., the other member of the partition. This constraint, based on Rousseeuw's (1987) *silhouettes*, helps correct wrong cluster assignments.

To find $\mathcal{P}_{\min}$, we first construct a random partition of the adjectives, then locate the adjective that will most reduce the objective function if it is moved from its current cluster. We move this adjective and proceed with the next iteration until no movements can improve the objective function. At the final iteration, the cluster assignment of any adjective that violates constraint (1) is changed. This is a steepest-descent hill-climbing method, and thus is guaranteed to converge. However, it will in general find a local minimum rather than the global one; the problem is NP-complete (Garey and Johnson, 1979). We can arbitrarily increase the probability of finding the globally optimal solution by repeatedly running the algorithm with different starting partitions.

7 Labeling the Clusters as Positive or Negative

The clustering algorithm separates each component of the graph into two groups of adjectives, but does not actually label the adjectives as positive or negative. To accomplish that, we use a simple criterion that applies only to pairs or groups of words of opposite orientation. We have previously shown (Hatzivassiloglou and McKeown, 1995) that in oppositions of gradable adjectives where one member is semantically unmarked, the unmarked member is the most frequent one about 81% of the time. This is relevant to our task because semantic markedness exhibits a strong correlation with orientation, the unmarked member almost always having positive orientation (Lehrer, 1985; Battistella, 1990).

We compute the average frequency of the words in each group, expecting the group with higher average frequency to contain the positive terms. This aggregation operation increases the precision of the labeling dramatically since indicators for many pairs of words are combined, even when some of the words are incorrectly assigned to their group.

8 Results and Evaluation

Since graph connectivity affects performance, we devised a method of selecting test sets that makes this dependence explicit. Note that the graph density is largely a function of corpus size, and thus can be increased by adding more data. Nevertheless, we report results on sparser test sets to show how our algorithm scales up.

We separated our sets of adjectives A (containing 1,336 adjectives) and conjunction- and morphology-based links L (containing 2,838 links) into training and testing groups by selecting, for several values of the parameter α, the maximal subset of A, A_α, which includes an adjective x if and only if there exist at least α links from L between x and other elements of A_α. This operation in turn defines a subset of L, L_α, which includes all links between members of A_α. We train our log-linear model on $L - L_\alpha$ (excluding links between morphologically related adjectives), compute predictions and dissimilarities for the links in L_α, and use these to classify and label the adjectives in A_α. α must be at least 2, since we need to leave some links for training.

Table 3 shows the results of these experiments for $\alpha = 2$ to 5. Our method produced the correct classification between 78% of the time on the sparsest test set up to more than 92% of the time when a higher number of links was present. Moreover, in all cases, the ratio of the two group frequencies correctly identified the positive subgroup. These results are extremely significant statistically (P-value less than 10^{-16}) when compared with the baseline method of randomly assigning orientations to adjectives, or the baseline method of always predicting the most frequent (for types) category (50.82% of the adjectives in our collection are classified as negative). Figure 2 shows some of the adjectives in set A_4 and their classifications.

Classified as positive:
bold decisive disturbing **generous good honest important large mature patient peaceful positive proud sound stimulating straightforward** strange **talented vigorous witty**

Classified as negative:
ambiguous cautious **cynical evasive harmful hypocritical inefficient insecure irrational irresponsible minor** outspoken pleasant **reckless risky selfish tedious unsupported vulnerable wasteful**

Figure 2: Sample retrieved classifications of adjectives from set A_4. Correctly matched adjectives are shown in bold.

9 Graph Connectivity and Performance

A strong point of our method is that decisions on individual words are aggregated to provide decisions on how to group words into a class and whether to label the class as positive or negative. Thus, the overall result can be much more accurate than the individual indicators. To verify this, we ran a series of simulation experiments. Each experiment measures how our algorithm performs for a given level of precision P for identifying links and a given average number of links k for each word. The goal is to show that even when P is low, given enough data (i.e., high k), we can achieve high performance for the grouping.

As we noted earlier, the corpus data is eventually represented in our system as a graph, with the nodes corresponding to adjectives and the links to predictions about whether the two connected adjectives have the same or different orientation. Thus the parameter P in the simulation experiments measures how well we are able to predict each link independently of the others, and the parameter k measures the number of distinct adjectives each adjective appears with in conjunctions. P therefore directly represents the precision of the link classification algorithm, while k indirectly represents the corpus size.

To measure the effect of P and k (which are reflected in the graph topology), we need to carry out a series of experiments where we systematically vary their values. For example, as k (or the amount of data) increases for a given level of precision P for individual links, we want to measure how this affects overall accuracy of the resulting groups of nodes. Thus, we need to construct a series of data sets, or graphs, which represent different scenarios corresponding to a given combination of values of P and k. To do this, we construct a random graph by randomly assigning 50 nodes to the two possible orientations. Because we don't have frequency and morphology information on these abstract nodes, we cannot predict whether two nodes are of the same or different orientation. Rather, we randomly assign links between nodes so that, on average, each node participates in k links and $100 \times P\%$ of all links connect nodes of the same orientation. Then we consider these links as identified by the link prediction algorithm as connecting two nodes with the same orientation (so that $100 \times P\%$ of these predictions will be correct). This is equivalent to the baseline link classification method, and provides a lower bound on the performance of the algorithm actually used in our system (Section 5).

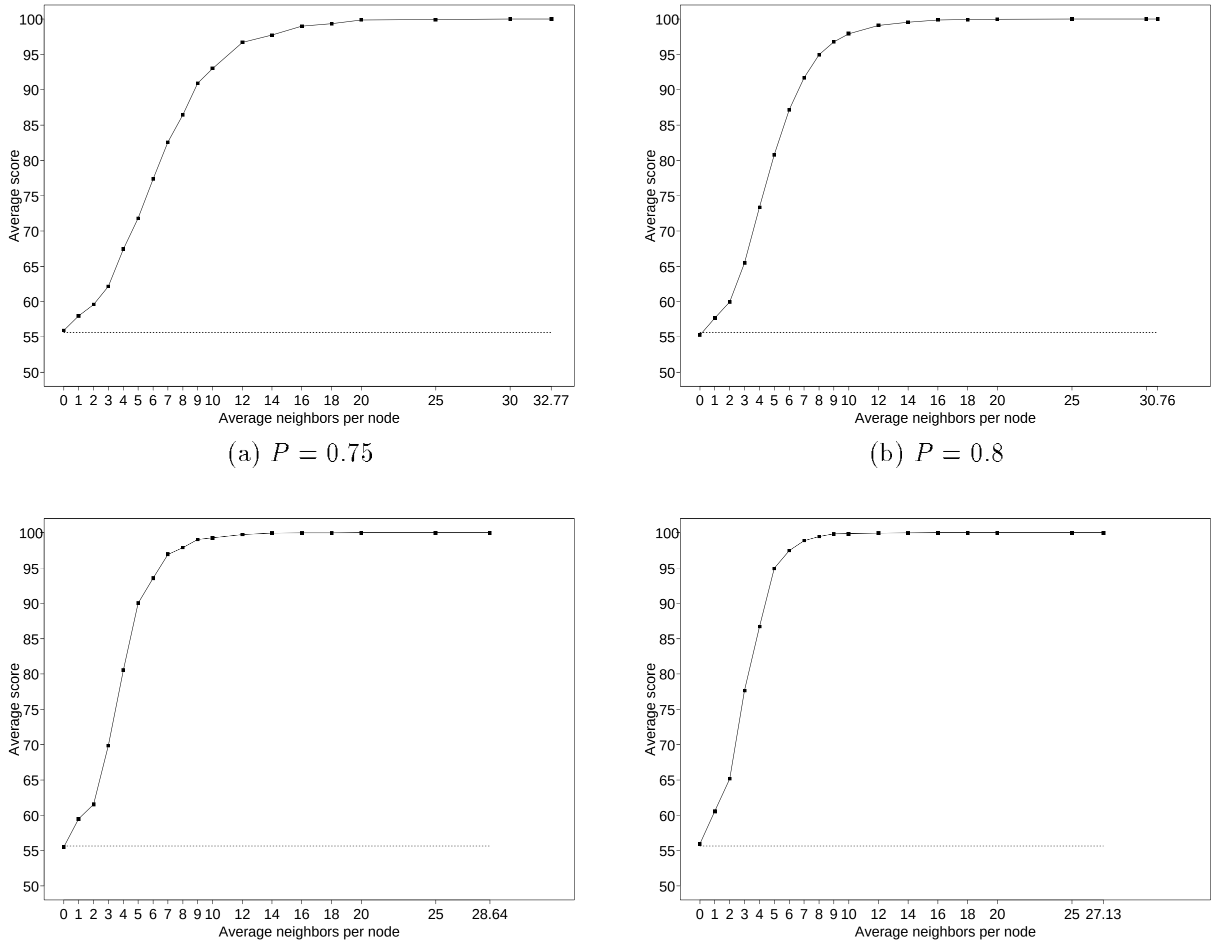
Because of the lack of actual measurements such as frequency on these abstract nodes, we also decouple the partitioning and labeling components of our system and score the partition found under the best matching conditions for the actual labels. Thus the simulation measures only how well the system separates positive from negative adjectives, not how well it determines which is which. However, in all the experiments performed on real corpus data (Section 8), the system correctly found the labels of the groups; any misclassifications came from misplacing an adjective in the wrong group. The whole procedure of constructing the random graph and finding and scoring the groups is repeated 200 times for any given combination of P and k, and the results are averaged, thus avoiding accidentally evaluating our system on a graph that is not truly representative of graphs with the given P and k.

We observe (Figure 3) that even for relatively low P, our ability to correctly classify the nodes approaches very high levels with a modest number of links. For $P = 0.8$, we need only about 7 links per adjective for classification performance over 90% and only 12 links per adjective for performance over 99%.[6] The difference between low and high values of P is in the rate at which increasing data increases overall precision. These results are somewhat more optimistic than those obtained with real data (Section 8), a difference which is probably due to the uniform distributional assumptions in the simulation. Nevertheless, we expect the *trends* to be similar to the ones shown in Figure 3 and the results of Table 3 on real data support this expectation.

10 Conclusion and Future Work

We have proposed and verified from corpus data constraints on the semantic orientations of conjoined adjectives. We used these constraints to automatically construct a log-linear regression model, which, combined with supplementary morphology rules, predicts whether two conjoined adjectives are of same

[6] 12 links per adjective for a set of n adjectives requires $6n$ conjunctions between the n adjectives in the corpus.



(a) $P = 0.75$

(b) $P = 0.8$

(c) $P = 0.85$

(d) $P = 0.9$

Figure 3: Simulation results obtained on 50 nodes. In each figure, the last x coordinate indicates the (average) maximum possible value of k for this P, and the dotted line shows the performance of a random classifier.

or different orientation with 82% accuracy. We then classified several sets of adjectives according to the links inferred in this way and labeled them as positive or negative, obtaining 92% accuracy on the classification task for reasonably dense graphs and 100% accuracy on the labeling task. Simulation experiments establish that very high levels of performance can be obtained with a modest number of links per word, even when the links themselves are not always correctly classified.

As part of our clustering algorithm's output, a "goodness-of-fit" measure for each word is computed, based on Rousseeuw's (1987) *silhouettes*. This measure ranks the words according to how well they fit in their group, and can thus be used as a *quantitative* measure of orientation, refining the binary *positive–negative* distinction. By restricting the labeling decisions to words with high values of this measure we can also increase the precision of our system, at the cost of sacrificing some coverage.

We are currently combining the output of this system with a semantic group finding system so that we can automatically identify antonyms from the corpus, without access to any semantic descriptions. The learned semantic categorization of the adjectives can also be used in the reverse direction, to help in interpreting the conjunctions they participate. We will also extend our analyses to nouns and verbs.

Acknowledgements

This work was supported in part by the Office of Naval Research under grant N00014-95-1-0745, jointly by the Office of Naval Research and the Advanced Research Projects Agency under grant N00014-89-J-1782, by the National Science Foundation under grant GER-90-24069, and by the New York State Center for Advanced Technology under contracts NYSSTF-CAT(95)-013 and NYSSTF-CAT(96)-013. We thank Ken Church and the AT&T Bell Laboratories for making the PARTS part-of-speech tagger available to us. We also thank Dragomir Radev, Eric Siegel, and Gregory Sean McKinley who provided models for the categorization of the adjectives in our training and testing sets as positive and negative.

References

Jean-Claude Anscombre and Oswald Ducrot. 1983. *L'Argumentation dans la Langue*. Philosophie et Langage. Pierre Mardaga, Brussels, Belgium.

Edwin L. Battistella. 1990. *Markedness: The Evaluative Superstructure of Language*. State University of New York Press, Albany, New York.

Peter F. Brown, Vincent J. della Pietra, Peter V. de Souza, Jennifer C. Lai, and Robert L. Mercer. 1992. Class-based n-gram models of natural language. *Computational Linguistics*, **18**(4):467–479.

Kenneth W. Church. 1988. A stochastic parts program and noun phrase parser for unrestricted text. In *Proceedings of the Second Conference on Applied Natural Language Processing (ANLP-88)*, pages 136–143, Austin, Texas, February. Association for Computational Linguistics.

W. J. Conover. 1980. *Practical Nonparametric Statistics*. Wiley, New York, 2nd edition.

Richard O. Duda and Peter E. Hart. 1973. *Pattern Classification and Scene Analysis*. Wiley, New York.

Michael Elhadad and Kathleen R. McKeown. 1990. A procedure for generating connectives. In *Proceedings of COLING*, Helsinki, Finland, July.

Michael R. Garey and David S. Johnson. 1979. *Computers and Intractability: A Guide to the Theory of NP-Completeness*. W. H. Freeman, San Francisco, California.

Vasileios Hatzivassiloglou and Kathleen R. McKeown. 1993. Towards the automatic identification of adjectival scales: Clustering adjectives according to meaning. In *Proceedings of the 31st Annual Meeting of the ACL*, pages 172–182, Columbus, Ohio, June. Association for Computational Linguistics.

Vasileios Hatzivassiloglou and Kathleen R. McKeown. 1995. A quantitative evaluation of linguistic tests for the automatic prediction of semantic markedness. In *Proceedings of the 33rd Annual Meeting of the ACL*, pages 197–204, Boston, Massachusetts, June. Association for Computational Linguistics.

John S. Justeson and Slava M. Katz. 1991. Co-occurrences of antonymous adjectives and their contexts. *Computational Linguistics*, **17**(1):1–19.

Adrienne Lehrer. 1974. *Semantic Fields and Lexical Structure*. North Holland, Amsterdam and New York.

Adrienne Lehrer. 1985. Markedness and antonymy. *Journal of Linguistics*, **31**(3):397–429, September.

John Lyons. 1977. *Semantics*, volume 1. Cambridge University Press, Cambridge, England.

Peter McCullagh and John A. Nelder. 1989. *Generalized Linear Models*. Chapman and Hall, London, 2nd edition.

George A. Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine J. Miller. 1990. Introduction to WordNet: An on-line lexical database. *International Journal of Lexicography (special issue)*, **3**(4):235–312.

Fernando Pereira, Naftali Tishby, and Lillian Lee. 1993. Distributional clustering of English words. In *Proceedings of the 31st Annual Meeting of the ACL*, pages 183–190, Columbus, Ohio, June. Association for Computational Linguistics.

Peter J. Rousseeuw. 1987. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, **20**:53–65.

Thomas J. Santner and Diane E. Duffy. 1989. *The Statistical Analysis of Discrete Data*. Springer-Verlag, New York.

Helmuth Späth. 1985. *Cluster Dissection and Analysis: Theory, FORTRAN Programs, Examples*. Ellis Horwood, Chichester, West Sussex, England.