DO WE NEED LINGUISTICS WHEN WE HAVE STATISTICS? A COMPARATIVE ANALYSIS OF THE CONTRIBUTIONS OF LINGUISTIC CUES TO A STATISTICAL WORD GROUPING SYSTEM

Vasileios Hatzivassiloglou

Department of Computer Science
450 Computer Science Building
Columbia University
New York, N.Y. 10027

vh@cs.columbia.edu

ABSTRACT

We present a comparative analysis of the performance of a statistics-based system for the formation of semantic groups of adjectives when various sources of linguistic knowledge are introduced. We identify four different types of shallow linguistic knowledge that are applicable to this system, and we quantify the performance gained by incorporating each such knowledge module. We perform experiments for different corpus sizes and different inputs (sets of adjectives to group), collect data on the usefulness of each linguistic module, assess the statistical significance of the results, and compare the contributions of the linguistic knowledge sources against each other. We also assess the overall effect linguistic knowledge has in our system. Our results show that linguistic knowledge causes a significant increase in the performance of the system. We conclude by discussing how these positive results can be generalized to other problems in statistical NLP.

1. INTRODUCTION

The idea of integrating statistical and knowledge-based approaches for natural language problems has been recently gaining ground in the computational linguistics community, as it is expected that a combined approach will offer significantly better performance over either methodology alone. This paper supplements this intuitive belief with actual evaluation data, obtained when several linguistics-based modules were integrated in a statistical system.

We used a system we previously developed for the separation of adjectives into semantic groups [Hatzivassiloglou and McKeown, 1993] as the basis for our comparative analysis. We identified several different types of shallow linguistic knowledge that can be efficiently introduced into our system. We evaluated the system with and without each such feature, obtaining an estimate of each feature's positive or negative contribution to the overall performance. By matching cases where all system parameters are the same except for one feature, we assess the statistical significance of the differences found. Also, a statistical model of the system's performance in terms of the active features for each run offers a view of the contributions of features from a different angle, contrasting the significance of linguistic features (or other modeled system parameters) against each other.

Our analysis of the experimental results showed that many forms of linguistic knowledge have a significant positive contribution to the performance of the system. We attribute to the combined effect of the linguistic knowledge modules the ability of our system to perform fine-tuned classification of adjectives into semantic classes. Other statistical systems that address word classification problems do not emphasize the use of linguistic knowledge and do not deal with a specific word class [Brown *et al.*, 1992], or do not exploit as much linguistic knowledge as we do [Pereira *et al.*, 1993]. As a result, a coarser classification is usually produced. In contrast, by limiting the system's input to adjectives, we can take advantage of specific syntactic relationships and additional filtering procedures that apply only to particular word classes. These sources of linguistic knowledge provide in turn the extra edge for discriminating among the adjectives at the semantic level.

Our adjective grouping system can be used for applications such as natural language generation (where knowledge of the semantic groups and of the ordering of the elements within them allows the precise lexicalization of semantic concepts [Elhadad, 1991]) and computational lexicography (by automatically compiling domain-dependent lists of synonyms and antonyms). The produced groups can also help correct erroneous usage of

multiple qualifiers that are superfluous or contradict each other, a phenomenon that has been observed in medical reports[1]. But in addition to the immediate applications of word classification, many other statistical NLP applications can be cast in a similar framework. Therefore, the positive effects of linguistic knowledge on our system indicate that the incorporation of linguistic knowledge will probably result in similar benefits for other applications as well.

In what follows, we briefly review our adjective grouping system, and then present the linguistic features we explored and the alternatives for each of them. In Section 5 we give the results of our evaluation on different combinations of features and we analyze their significance. We also present these results in a predictor-response framework, and we conclude by discussing the applicability of our results to other NLP problems.

2. AN OVERVIEW OF THE ADJECTIVE GROUPING SYSTEM

Our adjective grouping system [Hatzivassiloglou and McKeown, 1993] starts with a set of adjectives to be clustered into semantically related groups. Ideally, we want highly related words such as synonyms, antonyms, and hyponyms to be the only ones placed in the same group. The system is given the number of groups to form as an input parameter[2], and has access to a text corpus. No semantic information about the adjectives is available to the system. The system operates by extracting pairs of modified nouns for each adjective, and, optionally, pairs of adjectives that we can expect to be semantically unrelated on linguistic grounds[3]. From the estimated distribution of modified nouns for each adjective, a similarity score is assigned to each possible pair of adjectives. This is based on Kendall's τ, a non-parametric, robust estimator of correlation [Kendall, 1938]. Using the similarity scores and, optionally, the established relationships of non-relatedness, a non-hierarchical clustering method [Späth, 1985] assigns the adjectives to groups in a way that maximizes the within-group similarity (and therefore also maximizes the between-group dissimilarity).

1. deadly fatal
2. capitalist socialist
3. clean dirty dumb
4. hazardous toxic
5. insufficient scant
6. generous outrageous unreasonable
7. endless protracted
8. plain
9. hostile unfriendly
10. delicate fragile unstable
11. affluent impoverished prosperous
12. brilliant clever energetic smart stupid
13. communist leftist
14. astonishing meager vigorous
15. catastrophic disastrous harmful
16. dry exotic wet
17. chaotic turbulent
18. confusing misleading
19. dismal gloomy
20. dual multiple pleasant
21. fat slim
22. affordable inexpensive
23. abrupt gradual stunning
24. flexible lenient rigid strict stringent

Figure 1: Example clustering found by the system using all linguistic modules.

To evaluate our system, we have developed extended versions of the standard information retrieval measures precision, recall, and fallout. These extended versions score the grouping produced by the system against a set of model groupings (instead of just one) for the same adjec-

[1]We thank Johanna Moore for pointing out this application to us.

[2]Determining this number from the data is probably the hardest problem in cluster analysis in general; see [Kaufman and Rousseeuw, 1990].

[3]These are adjectives that either modify the same noun in the same NP (e.g. *big white house*) or one of them modifies the other (e.g. *light blue coat*); see [Hatzivassiloglou and McKeown, 1993] for a detailed analysis.

tives, supplied by humans. In the experiments reported in this paper, we employ 8 or 9 human-constructed models for each adjective set. We base our comparisons on and report the F-measure scores [Van Rijsbergen, 1979] which combine precision and recall in a single number. In addition, since the correct number of groupings is something that the system cannot yet determine (and, incidentally, something that human evaluators disagree about), we run the system for the five cases in the range -2 to +2 around the average number of clusters employed by the humans and average the results. This smoothing operation prevents an accidental high or low score being reported when a small variation in the number of clusters produces very different scores.

It should be noted here that the scores reported should not be interpreted as linear percentages. In other words, a score of 40 is not just twice as good as a score of 20, and going from 30 to 40 is much harder than going from 20 to 30. The latter is true for most applications, but the problem of interpreting the scores is exacerbated in our context because of the structural constraints imposed by the clustering and the presence of multiple models. Furthermore, even the best clustering that could be produced would not receive a score of 100, because of disagreement among humans on what is the correct answer. To clarify the meaning of the scores, we accompany them with lower and upper bounds for each adjective set we examine. These bounds are obtained by the performance of a system that creates random groupings (averaged over many runs) and by the average score of the human-produced partitions when evaluated against the other human models respectively.

Figure 1 shows an example clustering produced by our system for one of the adjective sets analyzed in this paper.

3. THE LINGUISTIC FEATURES BEING TESTED

We have identified several sources of linguistic knowledge that can be incorporated in our system, augmenting the statistical component. Each such source represents a parameter of the system, i.e. a feature that can be present or absent or more generally take a value from a predefined set. We selected features that can be efficiently computed in a completely automatic way for unrestricted text and do not require extensive amounts of knowledge to be available to the system. Almost all of these features can be generalized to other applications as well, as we discuss in Section 6. In this section we discuss first one of these parameters that can take several values, namely the method of extracting data from the corpus, and then several other binary-valued features.

3.1 Extracting data from the corpus

Our adjective clustering system determines the distribution of related (modified) nouns for each adjective and eventually the similarity between adjectives from pairs of the form (adjective, modified noun) observed in the corpus. Direct information about incompatible adjectives (in the form of appropriate adjective-adjective pairs) can also be collected from the corpus. Therefore, a first parameter of the system and a possible dimension for comparisons is the method employed to identify such pairs in free text. This is hardly a unique feature of our system: all word-based statistical systems must first collect data from the corpus about the words of interest, on which the subsequent statistics operate[4].

There are several alternate models for this task of data collection, with different degrees of linguistic sophistication. A first model is to use no linguistic knowledge at all: we collect for each adjective of interest all words that fall within a window of some predetermined size. Naturally, no negative data (adjective-adjective pairs) can be collected with this method. However, the method can be implemented easily and does not require the identification of any linguistic constraints so it is completely general. It has been used for diverse problems such as machine translation and sense disambiguation [Gale *et al.*, 1992, Schütze, 1992].

A second model is to restrict the words collected to the same sentence as the adjective of interest and to word class(es) that we expect on linguistic grounds to be relevant to adjectives. For our application, we collect all nouns in the vicinity of an adjective without leaving the current sentence. We assume that these nouns have some relationship with the adjective and that semantically different adjectives will exhibit different collections of such nouns. This model requires only part-of-speech information (to identify nouns) and a method of detecting sentence boundaries. It uses a window of fixed length to define the neighborhood of each adjective. Such a model incorporates minimal linguistic knowledge, namely in determining what constitutes the informative class(es) of words collected (nouns in our problem). Again, negative knowledge such as incompatible adjective pairs cannot be collected with this model. Nevertheless, it has also been widely used, e.g. for collocation extraction [Smadja, 1993] and sense disambiguation [Liddy and Paik, 1992].

A third model uses a simple linguistic rule to identify pairs of interest that is even more restrictive and informative than the ‘‘nouns in vicinity’’

[4]Although frequently details of the statistical model employed receive more consideration.

approach. Since we are interested in nouns modified by adjectives, such a rule is to collect a noun immediately following an adjective, assuming that this implies a modification relationship. Pairs of consecutive adjectives can also be collected.

Up to this point we have successively restricted the collected pairs on linguistic grounds, so that less but cleaner data is collected. For the fourth model, we extend the simple rule given above, using linguistic information to catch more valid pairs without sacrificing accuracy. We employ a pattern matcher that retrieves any sequence of one or more adjectives followed by any sequence of zero or more nouns. These sequences are then analyzed with heuristics based on linguistics to obtain pairs. For example, it can be shown that all adjectives in such a sequence must be semantically unrelated, and that it is best to attach all the adjectives to the final noun.

The regular expression and pattern matching rules of the previous model can be extended further, forming a grammar for the constructs of interest. This approach can detect more pairs, and at the same time address known problematic cases not detected by the previous models.

We implemented the above five data extraction models, using typical window sizes for the first two methods (50 and 5 on each side of the window respectively) which have been found useful in other problems before. For the fifth model, we developed a finite-state grammar for NPs which is able to handle both predicative and attributive modification of nouns, conjunctions of adjectives, adverbial modification of adjectives, quantifiers, and apposition of adjectives to nouns or other adjectives[5]. Unfortunately, the resources required to perform our tests for the first model were too great (e.g. 12,287,320 pairs in a 151 MB file were extracted for the 21 adjectives in our smallest test set) so we dropped that model from further consideration and we use the second model as the baseline of minimal linguistic knowledge. Other researchers have also reported similar problems of excessive resource demands with the ‘‘collect all neighbors’’ model [Gale *et al.*, 1992].

3.2 Other linguistic features

In addition to the data extraction method, we identified three other areas where linguistic knowledge can be introduced in our system. First, we can employ morphology to convert plural nouns to the corresponding singular ones and adjectives in comparative or superlative degree to their base form. Almost all adjectives and nouns that appear in multiple forms have no semantic difference from their base form except for the number or degree feature. This conversion combines counts of similar pairs, thus raising the expected and estimated frequencies of each pair in any statistical model. We developed a morphology component that produces the singular form of nouns using rules plus a large table of exceptions. For adjectives, a set of rules is again employed but because of the vowel in the suffix *-er* or *-est*, many base forms look plausible without a lexicon (e.g. *bigger* could have been produced from *big*, *bigg*, or *bigge*). We solve this problem by counting the occurrences of each candidate form in our corpus and selecting the one with non-zero frequency.

antitrust	new
big	old
economic	political
financial	potential
foreign	real
global	serious
international	severe
legal	staggering
little	technical
major	unexpected
mechanical	

Figure 2: Test set 1; from an earlier corpus.

Another potential application of linguistic knowledge is the use of a spell-checking procedure, combined with a word list, to eliminate typographical errors from the corpus. Such errors can produce wrong estimates for the frequencies of modified nouns for an adjective, but most importantly introduce ‘‘unique’’ nouns appearing only with one adjective, skewing the comparison of noun distributions. We implemented this component using the Unix *spell* program and associated word list, with extensions for hyphenated compounds. Unfortunately, since a fixed and domain independent lexicon is used for this process, some valid but overspecialized words may be discarded too.

Finally, we can use additional sources of knowledge which supplement the primary similarity relationships and are justified on linguistic grounds. We identified several potential sources of additional knowledge that can be extracted from the corpus (e.g. conjunctions of adjectives). In this comparison study we implemented and consider the significance of one of these knowledge sources, namely the negative examples offered by adjective-adjective pairs.

[5]For efficiency reasons we did not consider a more powerful formalism.

4. THE COMPARISON EXPERIMENTS

In the previous section we identified four parameters of the system, the effects of which we want to analyze. But in addition to these parameters that can be directly varied and have predetermined possible values, several other variables can affect the performance of the system.

First, the performance of the system depends naturally on the adjective set that is to be clustered. Presumably variations in the adjective set can be modeled by several parameters, such as size of the set, number of semantic groups in it, and strength of semantic relatedness among its members, plus several parameters describing the properties of the adjectives in the set in isolation, such as frequency, specificity, etc.

A second variable that affects the clustering is the corpus that is used as the main knowledge source, through the observed cooccurrence patterns. Again the effects of different corpora can be separated into several factors, e.g. the size of the corpus, its generality, the genre of the texts, etc.

Since in this paper we are interested in quantifying the effect of the linguistic knowledge in our system, or more precisely of the linguistic knowledge that we can explicitly control through the four parameters discussed above, we did not attempt to model in detail the various factors entering the system as a result of the choice of adjective set and corpus. However, we are interested in measuring the effects of the linguistic parameters in a wide range of contexts, and in correlating these effects with variables originating from the choice of corpus and adjective set. For example, we would want to be able to detect that the linguistic parameter ‘‘morphology’’ is significant for small corpora but not for large ones, if that were the case. Therefore, we included in our model two additional parameters, representing the corpus and the adjective set used.

We used the Wall Street Journal articles from the ACL-DCI as our corpus. We selected four subcorpora of decreasing size to study the relationship of corpus size with linguistic feature effects: all the 1987 articles (21 million words), every third of these articles (7 million words), every twenty-first (1 million words), and articles no. 50 and 100 (330,000 words). Since we use subsets of the same corpus, we are essentially modeling the corpus size parameter only.

annual	negative
big	net
chief	new
commercial	next
current	old
daily	past
different	positive
difficult	possible
easy	pre-tax
final	previous
future	private
hard	public
high	quarterly
important	recent
initial	regional
international	senior
likely	significant
local	similar
low	small
military	strong
modest	weak
national	

Figure 3: Test set 2; high frequency words.

abrupt	hazardous
affluent	hostile
affordable	impoverished
astonishing	inexpensive
brilliant	insufficient
capitalist	leftist
catastrophic	lenient
chaotic	meager
clean	misleading
clever	multiple
communist	outrageous
confusing	plain
deadly	pleasant
delicate	prosperous
dirty	protracted
disastrous	rigid
dismal	scant
dry	slim
dual	smart
dumb	socialist
endless	strict
energetic	stringent
exotic	stunning
fat	stupid
fatal	toxic
flexible	turbulent
fragile	unfriendly
generous	unreasonable
gloomy	unstable
gradual	vigorous
harmful	wet

Figure 4: Test set 3; low frequency words.

Parameter	Value	Score
Extraction Model	Parsing	30.29
	Pattern Matching	28.88
	Observed Pairs	27.87
	Nouns in Vicinity	22.36
Morphology	Yes	28.60
	No	27.53
Spell-checking	Yes	28.12
	No	28.00
Use of negative knowledge	Yes	29.40
	No	28.63

Table 1: Average scores when only one feature is changed.

For each corpus, we analyzed three different sets of adjectives, listed in figures 2-4. The first of them was selected from a similar corpus, contains 21 frequent and ambiguous words that all associate strongly with a particular noun (*problem*), and was analyzed in [Hatzivassiloglou and McKeown, 1993]. The second set (43 adjectives) was selected with the constraint that it contain high frequency adjectives (more than 1,000 occurrences in the 21 million word corpus). The third set (62 adjectives) satisfies the opposite constraint containing adjectives of relatively low frequency (between 50 and 250). Figure 1 shows a typical grouping found by our system for the third set of adjectives, when the full corpus and all linguistic modules were used.

These three sets of adjectives represent various characteristics of the adjective sets that the system may be called to cluster. First, they explicitly represent increasing sizes of the grouping problem. The second and third sets also contrast the independent frequencies of their member adjectives. Furthermore, we have found that the less frequent adjectives of the third set tend to be more specific than the more frequent ones. The human evaluators reported that the task of classification was easier for the third set, and their models exhibited about the same degree of agreement for the second and third sets although the third set is significantly larger. We plan to investigate the generality of this inverse correlation between frequency and specificity in the future.

By including the parameters ‘‘corpus size’’ and ‘‘adjective set’’, we have six parameters that we can vary in our experiments. Any remaining factors affecting the performance of our system are modeled as random noise, so statistical methods are used to evaluate the effects of the selected parameters. The six chosen parameters are completely orthogonal, with the exception that parameter ‘‘negative knowledge’’ must have the value ‘‘not used’’ when parameter ‘‘extraction model’’ has the value ‘‘nouns in vicinity’’. In order to avoid introducing imbalance in our experiment, we constructed a complete designed experiment [Hicks, 1973] for all their $(4\times2-1)\times2\times2\times4\times3=336$ valid combinations[6].

5. RESULTS

5.1 Average effect of each linguistic parameter

Space limitations do not allow us to present the scores for every one of the 336 individual experiments performed, corresponding to all valid combinations of the six modeled parameters. Instead we present several summary measures. We measured the effect of each particular setting of each linguistic parameter of Section 3 by averaging the scores obtained in all experiments where that particular parameter had that particular value. In this way, Table 1 summarizes the differences in the performance of the system caused by each parameter. Because of the complete design of the experiment, each value in Table 1 is obtained in runs that are identical to the runs used for estimating the other values of the same parameter except for the difference in the parameter itself[7].

Table 1 shows that there is indeed improvement with the introduction of any of the proposed linguistic features, or with the use of a linguistically more sophisticated extraction model. To assess the statistical significance of these differences, we compared each run for a particular value of a parameter to the corresponding identical (except for that parameter) run for a different value of the parameter. Each pair of values for a parameter produces in this way a set of paired observations. On each of these sets, we performed a sign test [Gibbons and Chakraborti, 1992] of the null hypothesis that there is no real difference in the system’s performance between the two values, i.e. that any observed difference is due to chance. We counted the number of times that the first of the two compared values led to superior performance relative to the second, distributing ties equally between the two cases. Under the null hypothesis, the number of times that the first value

[6]Recall that a designed experiment is *complete* when at least one trial, or *run*, is performed for every valid combination of the modeled predictors.

[7]The slight asymmetry in parameters ‘‘extraction model’’ and ‘‘negative knowledge’’ is accounted for by leaving out non-matching runs.

Parameter tested	Test		Comparisons	First value better than second	Probability
	First Value	Second Value			
Extraction model	Parsing	Pattern matching	96	64	0.0014
	Parsing	Observed pairs	96	66	0.0003
	Parsing	Nouns in vicinity	48	42	10^{-7}
	Pattern matching	Observed pairs	96	61	0.0104
	Pattern matching	Nouns in vicinity	48	41	$6.24 \cdot 10^{-7}$
	Observed pairs	Nouns in vicinity	48	36	0.0007
Morphology	Used	Not used	168	107	0.0005
Spell-checking	Used	Not used	168	94	0.1425
Negative knowledge	Used	Not used	144	97	$3.756 \cdot 10^{-5}$

Table 2: Statistical tests of the difference in performance offered by each linguistic feature.

performs better follows the binomial distribution with parameter p=0.5. Table 2 gives the results of these tests along with the probabilities that the same or more extreme results would be encountered by chance. We can see from the table that all types of linguistic knowledge except spell-checking have a beneficial effect that is statistically significant at, or below, the 1% level.

5.2 Comparison among the linguistic features

In order to measure the significance of the contribution of each linguistic feature relative to the other linguistic features, we fitted a linear regression model [Draper and Smith, 1981] to the data. We use the six parameters of our experiments as the predictors, and the measured F-score of the corresponding clustering as the response variable. In such a model the response Y is assumed to be a linear function of the predictors, i.e.

$$Y = b_0 + b_1 \cdot X_1 + b_2 \cdot X_2 + \ \ldots \ + b_n \cdot X_n \qquad (1)$$

where X_i is the i-th predictor and b_i is its corresponding weight[8]. The weights found by the fitting process (Table 3) indicate by their absolute magnitude and sign how important each predictor is and whether it contributes positively or negatively to the final result. Numerical values such as the corpus size enter formula (1) directly as predictors, so Table 3 indicates that each additional megabyte of text increases the performance of the system by 0.9417 on the average. For binary features, the weights in Table 3 indicate the increase in the system's performance when the feature is present, so introduction of morphology improves the system's performance by 0.5371 on the average. For the categorical variables ''extraction model'' and ''adjective set'', the weights show the change in score for the indicated value in contrast to the base case (minimal linguistic knowledge represented by extraction model ''nouns in vicinity'' and adjective set 1 respectively). For example, using the finite-state parser instead of the ''nouns in vicinity'' model improves

Variable	Weight
Intercept	18.7997
Corpus size (in megabytes)	0.9417
Extraction method (Pairs)	5.1307
Extraction method (Sequences)	6.1418
Extraction method (Parser)	7.5423
Morphology	0.5371
Spelling	0.0589
Adjective Set (2)	2.5996
Adjective Set (3)	-11.4882
Use of negative knowledge	0.3838

Table 3: Fitted coefficients for linear regression model.

[8]Such a model is appropriate for comparative purposes, although extrapolating response values for prediction outside the range of predictor values used in the fitting may give incorrect results. For example, the coefficients in Table 3 cannot be used to predict the score when the corpus is significantly smaller than 0.33 Mbytes or larger than 21 Mbytes.

	Adjective Set 1	Adjective Set 2	Adjective Set 3
Random partitions	9.66	6.21	3.80
No linguistic components active	24.51	38.51	33.21
All linguistic components active	39.06	44.73	46.17
Humans	53.98	64.27	63.07

Table 4: Performance of a random classifier, of the system on the 21 million word corpus, and of the humans.

the score by 7.5423 on the average, while going from adjective set 2 to adjective set 3 decreases the score by -(-2.5996 − 11.4882) = 14.0878 on the average. Finally the intercept b_0 gives a baseline performance of a minimal system that uses the base case for each parameter; the effects of corpus size are to be added to this system.

From Table 3 we can see that the data extraction model has a significant effect on the quality of the produced clustering, and among the linguistic parameters is the most important one. Increasing the size of the corpus also significantly increases the score. The adjective set that is clustered also has a major influence on the score, with rarer adjectives leading to worse clusterings. The two linguistic features ''morphology'' and ''negative knowledge'' have less pronounced although still significant effects, while spell-checking offers minimal improvement that probably does not justify the effort of implementing the module and the cost of activating it at run-time.

5.3 Overall effect of linguistic knowledge

Up to this point we have described averages of scores, taken over many combinations of features that are orthogonal to the one studied. These averages are good for describing the *existence* of a difference caused by the different values of each feature, across all possible combinations of the other features. They are not, however, representative of the performance of the system in a particular setting of parameters, nor are they suitable for describing the difference in features quantitatively, since they are averages taken over widely differing settings of the system's parameters. In particular, the inclusion of very small corpora drives the average scores down, as we have confirmed in a more detailed analysis where averages were computed separately for each value of the corpus size parameter. To give a feeling of how important the introduction of linguistic knowledge is quantitatively, we compare in Table 4 the results obtained for the full corpus of 21 million words for the two cases of having all or none of the linguistic components active. The scores obtained by a random system that produces partitions of the adjectives with no knowledge except the number of groups are included as a lower bound. These estimates are obtained after averaging the scores of 20,000 such random partitions for each adjective set. The average scores that each human model receives when compared to all the other human models are also included, since they provide an estimate of the maximum score that can be achieved by any system. That maximum depends on the disagreement between models for each adjective set. For these measurements we use a smaller smoothing window of size 3 instead of 5, which is fairer to the system when its performance is compared to the humans. We also give in Figure 5 the grouping produced by the system without using any of the linguistic modules for adjective set 3; this is to be contrasted with Figure 1.

6. GENERALIZING TO OTHER APPLICATIONS

In the previous section, we showed that the introduction of linguistic knowledge in our system produces a performance difference, which is not only statistically observable but also quantitatively significant (cf. Table 4). We believe that these positive results should also apply to other corpus-based NLP systems that employ statistical methods. Many of the linguistic components of our system, including the extraction model that was shown to be the most important linguistic parameter, are not specific to the word grouping problem. They can thus be directly incorporated in systems designed for other problems but essentially following the same basic architecture as ours.

Many statistical approaches share the same basic methodology with our system: a set of words is preselected, related words are identified in a corpus, the frequencies of words and of pairs of related words are estimated, and a statistical model is used to make predictions for the original words. Across applications, there are differences in what words are selected, how related words are defined, and what kind of predictions is made. Nevertheless, the basic components stay the same. For example, in our application the original words are

1. catastrophic harmful
2. dry wet
3. lenient rigid strict stringent
4. communist leftist
5. flexible hostile protracted unfriendly
6. abrupt chaotic disastrous gradual turbulent vigorous
7. affluent affordable inexpensive prosperous
8. outrageous
9. capitalist socialist
10. dismal gloomy pleasant
11. generous insufficient meager scant slim
12. delicate fragile
13. brilliant energetic
14. dual multiple stupid
15. hazardous toxic unreasonable unstable
16. plain
17. confusing
18. clever
19. endless
20. clean dirty impoverished
21. deadly fatal
22. astonishing misleading stunning
23. dumb fat smart
24. exotic

Figure 5: Example clustering found by the system using no linguistic modules.

the adjectives and the predictions are their groups; in machine translation, the predictions are the translations of the words in the source language text; in sense disambiguation, the predictions are the senses assigned to the words of interest; in part-of-speech tagging or in classification the predictions are the tags or classes assigned to each word. Because of this underlying similarity, the comparative analysis presented in the paper is relevant to all these problems.

For a concrete example, we examine the case of collocation extraction that has been addressed with statistical methods in the past. Smadja [1993] describes a system that initially uses the ‘‘nouns in vicinity’’ extraction model to collect cooccurrence information about words, and then identifies collocations on the basis of distributional criteria. A later component filters the retrieved collocations, removing the ones where the participating words are not used consistently in the same syntactic relationship. This post-processing stage doubles the precision of the system. We believe that using from the start a more sophisticated extraction model to collect these pairs of related words will have similar positive effects. Other linguistic components, such as a morphology module that combines frequency counts, should also improve the performance of the system. In this way, we can benefit from linguistic knowledge without having to use a separate filtering process after expending the effort to collect the collocations.

Similarly, the sense disambiguation problem is typically attacked by comparing the distribution of the neighbors of a word’s occurrence to prototypical distributions associated with each of the word’s senses [Gale *et al.*, 1992, Schütze, 1992]. Usually, no explicit linguistic knowledge is used in defining these neighbors, which are taken as all words appearing within a window of fixed width centered at the word being disambiguated. Many words unrelated to the word of interest are collected in this way. In contrast, identifying appropriate word classes that can be expected on linguistic grounds to convey significant information about the original word should increase the performance of the disambiguation system. Such classes might be modified nouns for adjectives, nouns in a subject or object position for verbs, etc. As we have showed in Section 5, less but cleaner information increases the quality of the results.

An interesting topic is the identification of parallels of our linguistic modules for these applications, at least for those modules which, unlike morphology, are not ubiquitous. Negative knowledge for example improves the performance of our system, supplementing the positive information provided by adjective-noun pairs. It could be useful for other systems as well if an appropriate application-dependent method of extracting such information is identified.

7. CONCLUSIONS AND FUTURE WORK

We have showed that all linguistic features considered in this study had a positive contribution to the performance of the system. Except for spell-checking, all these contributions were both statistically significant and large enough to make a difference in practical situations. Furthermore, the results can be expected to generalize to a wide variety of corpus-based systems for different applications.

The cost of incorporating the linguistics-based modules in the system is not prohibitive. The effort needed to implement all the linguistic modules was about 5 person-months, in contrast with 7 person-months needed to develop the basic statistical system. Furthermore, the run-time overhead caused by the linguistic modules is not significant. Each takes from 1 to 7 minutes on a Sun SparcStation 10 to process a million entries (words or pairs) and all except the negative knowledge module need process a corpus only once, reusing the same information for different adjective sets. This should be compared to the approximately 15 minutes needed by the statistical component for grouping about 40 adjectives.

In the future, we plan to extend the results discussed in this paper by an analysis of the dependence of the effects of each parameter on the values of the other parameters. We are currently stratifying the experimental data obtained to study trends in the magnitude of parameter effects as other parameters vary in a controlled manner, and we will examine the interactions with corpus size and specificity of clustered adjectives. We are also interested in providing similar quantitative results for other applications, to corroborate our belief in the generality of the importance of easily obtainable linguistic knowledge for statistical systems.

ACKNOWLEDGEMENTS

This work was supported jointly by ARPA and ONR under contract N00014-89-J-1782, by NSF GER-90-24069, and by New York State Center for Advanced Technology contract NYSSTF-CAT(91)-053. I wish to thank Kathy McKeown, Jacques Robin, and the workshop organizers for providing useful comments on earlier versions of the paper.

REFERENCES

Brown P., Della Pietra V., deSouza P., Lai J., and Mercer R. (1992). Class-based n-gram Models of Natural Language. *Computational Linguistics, 18:4*, 467-479.

Draper, N. R. and Smith, H. (1981). *Applied Regression Analysis* (2nd ed.). New York: Wiley.

Elhadad, Michael. (1991). Generating Adjectives to Express the Speaker's Argumentative Intent. *Proceedings of the 9th National Conference on Artificial Intelligence (AAAI 91)*. Anaheim.

Gale, W. A., Church, K. W., and Yarowsky, D. (1992). Work on Statistical Methods for Word Sense Disambiguation. *Probabilistic Approaches to Natural Language: Papers from the 1992 Fall Symposium*. AAAI.

Gibbons, Jean Dickinson and Chakraborti, Subhabrata. (1992). *Nonparametric Statistical Inference* (3rd ed.). New York: Marcel Dekker.

Hatzivassiloglou, Vasileios and McKeown, Kathleen. (June 1993). Towards the Automatic Identification of Adjectival Scales: Clustering Adjectives According to Meaning. *Proceedings of the 31st Annual Meeting of the ACL*. Columbus, Ohio: Association for Computational Linguistics.

Hicks, C. R. (1973). *Fundamental Concepts in the Design of Experiments*. New York: Holt, Rinehart, and Wilson.

Kaufman, L. and Rousseeuw, P.J. (1990). *Finding Groups in Data: An Introduction to Cluster Analysis*. New York: Wiley.

Kendall, M.G. (1938). A New Measure of Rank Correlation. *Biometrika, 30*, 81-93.

Liddy, Elizabeth D. and Paik, Woojin. (1992). Statistically-Guided Word Sense Disambiguation. *Probabilistic Approaches to Natural Language: Papers from the 1992 Fall Symposium*. AAAI.

Pereira F., Tishby N., and Lee L. (June 1993). Distributional Clustering of English Words. *Proceedings of the 31st Conference of the ACL*. Columbus, Ohio: Association for Computational Linguistics.

Schütze, Hinrich. (July 1992). Word Sense Disambiguation With Sublexical Representations. *Proceedings of the AAAI-92 Workshop on Statistically-Based NLP Techniques*. AAAI.

Smadja, Frank. (March 1993). Retrieving Collocations from Text: Xtract. *Computational Linguistics, 19:1*, 143-177.

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

Van Rijsbergen, C.J. (1979). *Information Retrieval* (2nd ed.). London: Butterwoths.