Proceedings of the Second International Conference on Quantitative Linguistics,
Moscow, Russia, September 20-24, 1994

Corpus-Based Analyses of Adjectives: Automatic Clustering

Kuang-hua Chen and Hsin-Hsi Chen

Department of Computer Science and Information Engineering

National Taiwan University

Taipei, Taiwan, R.O.C.

Tel: +886-2-3625336 ext. 301; Fax: +886-2-3628167

e-mail: khchen@nlg.csie.ntu.edu.tw; hh_chen@csie.ntu.edu.tw

Abstract

Similarity analysis is a substantial issue in both corpus-based researches and language usages. This paper focuses on the semantic usages of adjectives, and analyzes the similarities among adjectives. The adjective and the semantic tag of the head noun that it modifies in a noun phrase form a co-occurrence. A two-stage algorithm is applied to clustering the adjectives according to these co-occurrence relationships. Experimental results show that we break even the two issues of large data clustering and meaningful clustering.

Paper Category: Topical Paper.
Topic Area: Corpus Linguistics, Similarity Analysis, Clustering.

1. Introduction

Since the importance of real-world applications is committed in recent years, corpus-based researches become the core of the field of computational linguistics. Many models, such as hidden Markov model, word association model, cache-based model, *etc.*, have been proposed to deal with practical applications. An important problem in these models is how to calculate the reliable probabilities of events. Many smoothing methods are reported. Most of these methods use low degree Markov probabilities to replace the unreliable high degree probabilities. A thorough resolution to this problem is to cluster the events. Brown *et al.* (1992) propose class-based language models to group words directly on the training table. Clustering not only reduces the memory needed in corpus-based tasks, but also smoothes the probabilities of events. In addition, word groups could be further investigated for language usage and lexicography. A method to group nouns according to the predicate-argument structures is described by Hindle (1990). Hatzivassiloglou and McKeown (1993) analyze the co-occurrences of adjectives and nouns, and then cluster the adjectives. All these methods investigate the relationships among word surface forms. In this paper, we intensify the semantic usages of words. The idea is to examine the noun phrases in text corpora, and assign semantic tags to head nouns. Then, the co-occurrences of the premodifying adjectives and the semantic tag of head noun provide the clues for clustering.

Section 2 will describe semantic tags provided by Roget's Thesaurus and the overall analysis procedure for adjectives. Section 3 will touch on the clustering algorithm. The experimental results will be discussed in Section 4. Section 5 will conclude the remarks.

2. The Work

The proposed method uses a probabilistic chunker (Chen and Chen, 1993) to generate chunked texts, and determines the possible boundaries of phrase structures. All the noun chunks that contain adjectives are considered for further analyses. The head nouns of these noun chunks are assigned semantic tags, and then the relationships of adjectives and the semantic tags are investigated. The semantic tags of nouns are defined by Roget's Thesaurus. Roget's Thesaurus defines 1000 tags, and these tags are the leaves of a tree-like structure. Six classes are given. Various sections are defined under the classes and the 1000 tags are characterized. Table 1 shows the plan of classification of Roget's Thesaurus.

The overall analysis procedures are summarized in Figure 1. The test texts are LOB Corpus which contains 9123 different adjectives. The component CHUNKER is trained from LOB Corpus underlying bigram language model, i.e., two probabilities are considered. One is the probability of a chunk; the other is the probability of a chunk given its previous chunk. The NOUN-TRACTOR, which is a finite state mechanism containing 9 states, extracts the maximum length noun phrases from chunks sequence. The chunker has 98% chunk correct rate and 94% sentence correct rate in outside test (Chen and Chen, 1993). The average precision of NOUN-TRACTOR is 95% (Chen and Chen, 1994). Due to the high performance of these two components, the extracted noun phrases are suitable for clustering the adjectives.

Two similarity modules are considered in Hatzivassiloglou and Mckeown (1993), one is the co-occurrences of adjectives and nouns; the other is those of adjectives and adjectives. Their postulation is: it is impossible that two adjectives modifying the same nouns belong to the same cluster. However, it is not true in many real examples. Some of these shown as follows are quoted from LOB Corpus.

(1) The two *rival African nationalist* parties of Northern Rhoodesia. (A01:25)
... crochet and tatting in *fine* and *medium*-weight cottons ... (E01:55)
... an electrical drill, *pure* and *simple* ... (E03:104)
They are very *simple*, *cheap* and *easy* to make. (E04:97)

Table 1. Classification of Roget's Thesaurus

CLASS	SECTION	TAG	CLASS	SECTION	TAG
ABSTRACT RELATIONS	Existence	1 - 8	SPACE	In General	180 - 191
	Relation	9 - 24		Dimensions	192 - 239
	Quantity	25 - 57		Form	240 - 263
	Order	58 - 83		Motion	264 - 315
	Number	84 - 105	MATTER	In General	316 - 320
	Time	106 - 139		Inorganic	321-356
	Change	140 - 152		Organic	357 - 449
	Causation	153 - 179	AFFECTIONS	In General	820 - 826
INTELLECT	Formation of Ideas	450 - 515		Personal	827 - 887
	Communication of Ideas	516 - 599		Sympathetic	888 - 921
VOLITION	Individual	600 - 736		Moral	922 - 975
	Intersocial	737 - 819		Religious	975 - 1000

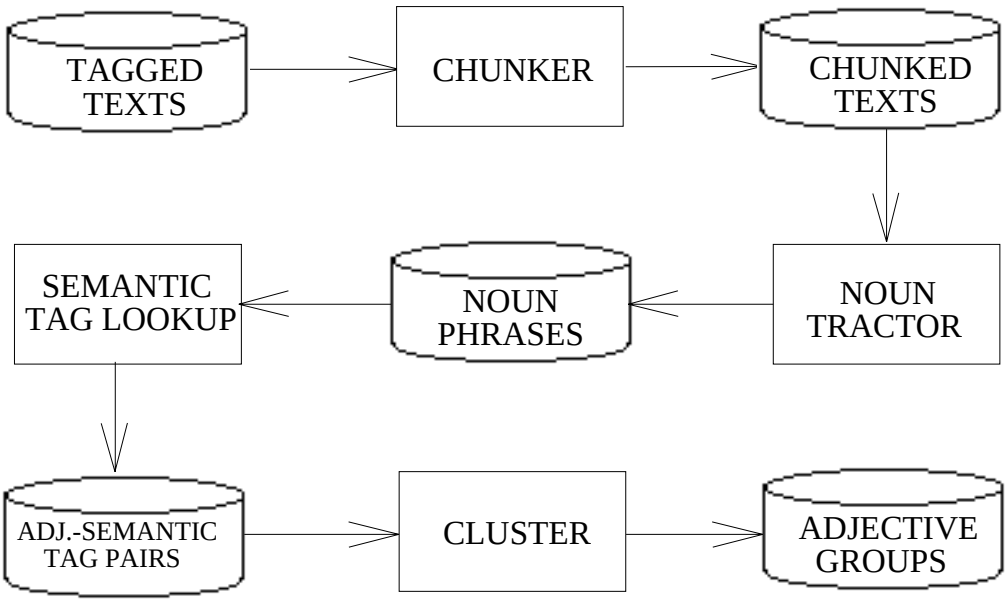

Figure 1. Experimental Procedures

As the result, we do not use the negative evidences for similarity analysis. Currently, only co-occurrences of adjectives and noun tags are considered.

3. Clustering

Our vocabulary consists of 9123 different adjectives and 1001 possible semantic tags for nouns (an extra semantic tag is used for unknown words). To cluster the 9123 adjectives is intractable in many clustering methods. The method proposed by Hatzivassiloglou and McKeown (1993) costs much computing time. This is why only 21 different adjectives are used as test set in their paper. Here, a two-stage clustering algorithm (Chen and Lee, 1994) is employed to cluster the adjectives. The co-occurrences of adjectives and semantic tags can be regarded as a matrix (say original matrix, OM) with 9123 rows and 1001 columns. Each entry indexed by (*i*,*j*) in the matrix is the frequency of co-occurrence of the i'th adjective and the j'th semantic tag in the testing corpus.

According to the entries in *OM*, a bit matrix *BM* is generated under the following rule.

(2) For each entry (*i*,*j*) in *OM*, if *OM*(*i*,*j*) > 0, then set *BM*(*i*,*j*) to 1.
Otherwise, set *BM*(*i*,*j*) to 0.

The similarity of two adjectives is measured by the respective row vectors in *BM*. (3) gives the similarity measure. *RV*[*k*] denotes the k'th element of row vector and $\oplus$ denotes the exclusive or.

(3) $$Sim(RV_{r_i}, RV_{r_j}) = 1001 - \sum_{k=1}^{1001} RV_{r_i}[k] \oplus RV_{r_j}[k]$$

The high similarity measure means the two row vectors are highly similar to each other. Based on the similarity measure, an optimal row index sequence (*ORIS*) is generated. The first element of the *ORIS* is obtained by choosing the row vector indexed by *ri* which has the highest similarity measure with zero vector. Then, the second element is obtained by choosing the row vector index *rj* which has the highest similarity measure with RV_{r_i}. The rest elements can be obtained in the same way. The above steps form the first stage of the clustering algorithm. ORIS will guide the clustering procedure in the next stage, and will reduce the complexity of the overall algorithm.

The second stage finds the clusters according to a predefined threshold value and the *ORIS* generated in the first stage. The threshold value determines how many clusters will be generated. At first, the zero row vectors in *OM* are clustered into an initial class *RC0*, say m_0 zero vector. Assume *k-1* clusters, *RC1*, *RC2*, ..., and *RCk-1*, are formed, and these clusters consist of m_1, m_2, ..., and m_{k-1} adjectives, respectively. Let $m = m_0 + m_1 + m_2 + ... + m_{k-1}$. Initiate a new cluster *RCk* with only one row vector $RV_{ORIS_{m+1}}$. Add $RV_{ORIS_{m+2}}$, $RV_{ORIS_{m+3}}$, ..., into the new cluster *RCk* until the information loss is larger than the predefined threshold. The information loss is defined below.

(4) $$IL_k = \sum_{i=1}^{m_k} \sum_{j=1}^{1001} abs(RV_{ORIS_{m+i}}[j] - RV_{RC_k}[j])$$

where $RV_{RC_k} = \sum_{i=1}^{m_k} RV_{ORIS_{m+i}}$, and function *abs* returns an absolute value.

The complexity is shown to be $O(M^2N)$ for an M×N matrix, so that the complexity of this clustering algorithm is tractable for large vocabulary application like our work.

4. Experimental Results

The first experiment we conduct is a repetition in (Hatzivassiloglou and McKeown, 1993), i.e., cluster the 21 adjectives listed in Table 2. Because the adjective *antitrust* does not occur in the test LOB Corpus, we exclude it in our experiment.

Table 2. Adjectives Used in Experiment I

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

Table 3 demonstrates the experimental result. In general, the clusters correspond to the common usages. The adjectives *old* and *new* are grouped into the same cluster, not two different clusters shown in (Brown *et al.*, 1990). This is because the negative evidences are not used in our experiment. From the viewpoint of language usage, *new* and *old* are used to modify the same kind of nouns. As the result, they belong to the same group in our model.

Table 3. Experimental Results of Experiment I

Cluster	Words
1	economic, financial, unexpected, potential, legal
2	mechanical, technical, international, foreign, global
3	major, serious, severe
4	real, big
5	political, old, new
6	little, staggering

Economic and *financial* are grouped together, but *political* are included into other cluster. These words are expected to be in the same cluster. However, due to the limited size of LOB Corpus (only 1M words), some bad clusters is unavoidable.

The second experiment we consider is to cluster some hyphened adjectives selected from test corpus. Total 21 adjectives are included in the experiment and segmented into 12 clusters. These words and the clusters to which they belong are listed in Table 4.

Table 4. Some Results of Experiment II

Cluster	Words

1	public-opinion
2	American-Indian, British-Caribbean
3	best-known, little-known, present-day
4	important-looking, politico-economic
5	main-line, right-angled
6	white-armed, younger
7	whole-hearted
8	long-ago
9	small-bowled, large-scale
10	old-age, new-born
11	good-humoured
12	great-power, high-backed

The clusters in Table 4 show many uncommon hyphened adjectives are also grouped into some meaningful clusters.

The last experiment is to cluster all of the adjectives in the test corpus, i.e., the 9123 adjectives. The resulting clusters are 367. This experiment takes about 24 hours. Because many low-frequency adjectives involve in, the experimental result is heavily disturbed. The meaningful clustering is not attained in the global viewpoint. But we still receive some good clusters. For example, the cluster 256 consists of *Nazi-style*, *socialistic*, *nationalistic* and *prohibitive*. Cluster 342 consists of *German-French*, *French-Canadian*, *American-Indian*, *British-Caribbean* and *Soviet-American*.

5. Concluding Remarks

For reliable estimation of probabilities in corpus-based researches, clustering is indispensable. From viewpoint of language usages, to cluster words is a good way for comparing words. Usually, the researches in grouping words focus on the surface forms of words, and try to find the implicit lexical-semantic relationship. In this paper, we cluster adjectives according to their semantic usages directly. The associations of the semantic tags of head nouns and their adjective modifiers are considered. Namely, a link is built between word and semantic tag, not just word and word. Since clustering is an NP problem, experiments on large volume of data seem to be intractable. However, practical applications are important and unavoidably this kind of researches involve very large data. To make the clustering for large data tractable, a two-stage clustering algorithm is applied. Comparing to 21 adjectives tested in Hatzivassiloglou and McKeown (1993), 9123 adjectives occurring in the LOB Corpus form the test set. Due to the training size of LOB Corpus, the results shown in our work demonstrate both good clusters and bad clusters. Very large corpus should be used to prove the effectiveness. Another problem is how to assign a unique semantic tag to head noun. Future works should focus on these two issues.

Acknowledgments

We are thankful to Ren-Feng Chang and Yue-Shi Lee for their helps in programming.

References

Brown, P.F., Pietra, V.J. *et al.* (1992), "Class-Based N-Gram Models of Natural Language". *Computational Linguistics*, 18(4), 467-479.

Hindle, D. (1990), "Noun Classification from Predicate-Argument Structures". In: *Proceedings of 28th Annual Meeting of ACL*, 268-275.

Hatzivassiloglou, V. and McKeown, K. (1993), "Towards the Automatic Identification of Adjectival Scales: Clustering Adjectives According to Meaning". In: *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*, 172-182.

Chen, K.-H. and Chen, H.-H. (1993), "A Probabilistic Chunker". In: *Proceedings of ROCLING VI*, 99-117.

Chen, K.-H. and Chen, H.-H. (1994), "Extracting Noun Phrases from Large-Scale Texts: A Hybrid Approach and Its Automatic Evaluation". In: *Proceedings of the 32nd Annual Meeting of ACL*, 234-241.

Chen, H.-H. and Lee, Y.-S. (1994), "An Unsupervised Clustering Algorithm for Storage Reduction in Corpus-Based Applications". Submitted to *IEEE Transactions on Knowledge and Data Engineering*.