Linguistic Preprocessing for Distributional Classification of Words

Viktor PEKAR
CLG, University of Wolverhampton
MB 114, Stafford Road
Wolverhampton, UK, WV1 1SB
v.pekar@wlv.ac.uk

Abstract

The paper is concerned with automatic classification of new lexical items into synonymic sets on the basis of their co-occurrence data obtained from a corpus. Our goal is to examine the impact that different types of linguistic preprocessing of the co-occurrence material have on the classification accuracy. The paper comparatively studies several preprocessing techniques frequently used for this and similar tasks and makes conclusions about their relative merits. We find that a carefully chosen preprocessing procedure achieves a relative effectiveness improvement of up to 88% depending on the classification method in comparison to the window-based context delineation, along with using much smaller feature space.

1 Introduction

With the fast development of text mining technologies, automated management of lexical resources is presently an important research issue. A particular text mining task often requires a lexical database (e.g., a thesaurus, dictionary, or a terminology) with a specific size, topic coverage, and granularity of encoded meaning. That is why a lot of recent NLP and AI research has been focusing on finding ways to speedily build or extend a lexical resource ad hoc for an application.

One attractive idea to address this problem is to elicit the meanings of new words automatically from a corpus relevant to the application domain. To do this, many approaches to lexical acquisition employ the distributional model of word meaning induced from the distribution of the word across various lexical contexts of its occurrence found in the corpus. The approach is now being actively explored for a wide range of semantics-related tasks including automatic construction of thesauri (Lin, 1998; Caraballo, 1999), their enrichment (Alfonseca and Manandhar, 2002; Pekar and Staab, 2002), acquisition of bilingual lexica from non-aligned (Kay and Röscheisen, 1993) and non-parallel corpora (Fung and Yee, 1998), learning of information extraction patterns from un-annotated text (Riloff and Schmelzenbach, 1998).

However, because of irregularities in corpus data, corpus statistics cannot guarantee optimal performance, notably for rare lexical items. In order to improve robustness, recent research has attempted a variety of ways to incorporate external knowledge into the distributional model. In this paper we investigate the impact produced by the introduction of different types of linguistic knowledge into the model.

Linguistic knowledge, i.e., the knowledge about linguistically relevant units of text and relations holding between them, is a particularly convenient way to enhance the distributional model. On the one hand, although describing the "surface" properties of the language, linguistic notions contain conceptual information about the units of text they describe. It is therefore reasonable to expect that the linguistic analysis of the context of a word yields additional evidence about its meaning. On the other hand, linguistic knowledge is relatively easy to obtain: linguistic analyzers (lemmatizers, PoS-taggers, parsers, etc) do not require expensive hand-encoded resources, their application is not restricted to particular domains, and their performance is not dependent on the amount of the textual data. All these characteristics fit very well with the strengths of the distributional approach: while enhancing it with external knowledge, linguistic analyzers do not limit its coverage and portability.

This or that kind of linguistic preprocessing is carried out in many previous applications of the approach. However, these studies seldom motivate the choice of a particular preprocessing procedure, concentrating rather on optimization of other parameters of the methodology. Very few studies exist that analyze and compare different techniques for linguistically motivated extraction of distributional data. The goal of this paper is to exploire in detail a range of variables in the morphological and syntactic processing of the context information and reveal the merits and drawbacks of their particular settings.

The outline of the paper is as follows. Section 2 describes the preprocessing methods under study.

Section 3 describes the settings for their empirical evaluation. Section 4 details the experimental results. Section 5 discusses related work. Section 6 summarizes the results and presents the conclusions from the study.

2 Types of Linguistic Preprocessing

In order to prepare a machine-processable representation of a word from particular instances of its occurrence, one needs to decide on, firstly, what is to be understood by the context of a word's use, and, secondly, which elements of that context will constitute distributional features. A straightforward decision is to take a certain number of words or characters around the target word to be its occurrence context, and all uninterrupted letter sequences within this delineation to be its features. However, one may ask the question if elements of the text most indicative of the target word's meaning can be better identified by looking at the linguistic analysis of the text.

In this paper we empirically study the following types of linguistic preprocessing.

1. *The use of original word forms vs. their stems vs. their lemmas as distributional features*. It is not evident what kind of morphological preprocessing of context words should be performed, if at all. Stemming of context words can be expected to help better abstract from their particular occurrences and to emphasize their invariable meaning. It also relaxes the stochastic dependence between features and reduces the dimensionality of the representations. In addition to these advantages, lemmatization also avoids confusing words with similar stems (e.g., *car* vs. *care*, *ski* vs. *sky*, *aide* vs. *aid*). On the other hand, morphological preprocessing cannot be error-free and it may seem safer to simply use the original word forms and preserve their intended meaning as much as possible. In text categorization, stemming has not been conclusively shown to improve effectiveness in comparison to using original word forms, but it is usually adopted for the sake of shrinking the dimensionality of the feature space (Sebastiani, 2002). Here we will examine both the effectiveness and the dimensionality reduction that stemming and lemmatization of context words bring about.

2. *Morphological decomposition of context words*. A morpheme is the smallest meaningful unit of the language. Therefore decomposing context words into morphemes and using them as features may eventually provide more fine-grained evidence about the target word. Particularly, we hypothesize that using roots of context words rather than their stems or lemmas will highlight lexical similarities between context words belonging to different parts of speech (e.g., *different*, *difference*, *differentiate*) or differing only in affixes (e.g., *build* and *rebuild*).

3. *Different syntactically motivated methods of delimiting the context of the word's use*. The lexical context permitting occurrence of the target word consists of words and phrases whose meanings have something to do with the meaning of the target word. Therefore, given that syntactic dependencies between words presuppose certain semantic relations between them, one can expect syntactic parsing to point to most useful context words. The questions we seek answers to are: Are syntactically related words indeed more revealing about the meaning of the target word than spatially adjacent ones? Which types of syntactic dependencies should be preferred for delimiting the context of a target word's occurrence?

4. *Filtering out rare context words*. The typical practice of preprocessing distributional data is to remove rare word co-occurrences, thus aiming to reduce noise from idiosyncratic word uses and linguistic processing errors and at the same time form more compact word representations (e.g., Grefenstette, 1993; Ciaramita, 2002). On the other hand, even single occurrence word pairs make up a very large portion of the data and many of them are clearly meaningful. We compare the quality of the distributional representations with and without context words that occurred only once with the target word.

3 Evaluation

3.1 Experimental Task

The preprocessing techniques were evaluated on the task of automatic classification of nouns into semantic classes. The evaluation of each preprocessing method consisted in the following. A set of nouns N each belonging to one semantic class $c \hat{I} C$ was randomly split into ten equal parts. Co-occurrence data on the nouns was collected and preprocessed using a particular method under analysis. Then each noun $n \hat{I} N$ was represented as a vector of distributional features: $\vec{n} = (v_{n,1}, v_{n,2}, \ldots v_{n,i})$, where the values of the features are the frequencies of n occurring in the lexical context corresponding to v. At each experimental run, one of the ten subsets of the nouns was used as the test data and the remaining ones as the train data. The reported effectiveness measures are microaveraged precision scores averaged over the ten runs. The statistical significance of differences between performance of particular preprocessing methods reported below was estimated by means of the one-tailed paired t-test.

3.2 Data

The set of nouns each provided with a class label to be used in the experiments was obtained as follows. We first extracted verb-noun dependencies from the British National Corpus, where nouns are either direct or prepositional objects to verbs. Each noun that occurred with more than 20 different verbs was placed into a semantic class corresponding to the WordNet synset of its most frequent sense. The resulting classes with less than 2 nouns were discarded. Thus we were left with 101 classes, each containing 2 or 3 nouns.

3.3 Classification Methods

Two classification algorithms were used in the study: Naïve Bayes and Rocchio, which were previously shown to be quite robust on highly dimensional representations on tasks including word classification (e.g., Tokunaga et al., 1997, Ciaramita, 2002).

The Naïve Bayes algorithm classifies a test instance n by finding a class c that maximizes $p(c|\vec{n})$. Assuming independence between features, the goal of the algorithm can be stated as:

$$\arg\max_i \; p(c_i \mid \vec{n}) \approx \arg\max_i \; p(c_i) \prod_{v \in \vec{n}} p(v \mid c_i)$$

where $p(c_i)$ and $p(v|c_i)$ are estimated during the training process from the corpus data.

The Naïve Bayes classifier was the binary independence model, which estimates $p(v|c_i)$ assuming the binomial distribution of features across classes. In order to introduce the information inherent in the frequencies of features into the model all input probabilities were calculated from the real values of features, as suggested in (Lewis, 1998).

The Rocchio classifier builds a vector for each class $c \hat{I} C$ from the vectors of training instances. The value of jth feature in this vector is computed as:

$$v_{c,j} = b \cdot \frac{\sum_{i \in c} v_{i,j}}{|c|} - g \cdot \frac{\sum_{i \in \bar{c}} v_{i,j}}{|\bar{c}|}$$

where the first part of the equation is the average value of the feature in the positive training examples of the class, and the second part is its average value in the negative examples. The parameters b and g control the influence of the positive and negative examples on the computed value, usually set to 16 and 4, correspondingly. Once vectors for all classes are built, a test instance is classified by measuring the similarity between its vector and the vector of each class and assigning it to the class with the greatest similarity. In this study, all features of the nouns were modified by the TFIDF weight before the training.

4 Results

4.1 Syntactic Contexts

The context of the target word's occurrence can be delimited syntactically. In this view, each context word is a word that enters in a syntactic dependency relation with the target word, being either the head or the modifier in the dependency. For example, in the sentence *She bought a nice hat* context words for *hat* are *bought* (the head of the predicate-object relation) and *nice* (the attributive modifier).

We group typical syntactic relations of a noun together based on general semantic relations they indicate. We define five semantic types of distributional features of nouns that can be extracted by looking at the dependencies they participate in.

A. verbs in the active form, to which the target nouns are subjects (e.g., the **committee** *discussed* (the issue), the **passengers** *got on* (a bus), etc);
B. active verbs, to which the target nouns are direct or prepositional objects (e.g., *hold* a **meeting**; *depend on* a **friend**); passive verbs to which the nouns are subjects (e.g., the **meeting** is *held*);
C. adjectives and nouns used as attributes or predicatives to the target nouns (e.g., a *tall* **building**, the **building** is *tall*; *amateur* **actor**, the **actor** is an *amateur*);
D. prepositional phrases, where the target nouns are heads (e.g., the **box** *in* the *room*); we consider three possibilities to construct distributional features from such a dependency: with the preposition (*in_room*, D_1), without it (*room*, D_2), and creating to separate features for the preposition and the noun (*in* and *room*, D_3).
E. prepositional phrases, where the target nouns are modifiers (the *ball in* the **box**); as with type D, three subtypes are identified: E_1 (*ball_in*), E_2 (*ball*), and E_3 (*ball* and *in*);

We compare these feature types to each other and to features extracted by means of the window-based context delineation. The latter were collected by going over occurrences of each noun with a window of three words around it. This particular size of the context window was chosen following

findings of a number of studies indicating that small context windows, i.e. 2-3 words, best capture the semantic similarity between words (e.g., Levy et al., 1998; Ciaramita, 2002). Thereby, a common stoplist was used to remove too general context words. All the context words experimented with at this stage were lemmatized; those, which co-occurred with the target noun only once, were removed.

We first present the results of evaluation of different types of features formed from prepositional phrases involving target nouns (see Table 1).

	Naïve Bayes	Rocchio	#dim
D_1	**23.405**	**16.574**	11271
D_2	18.571	13.879	5876
D_3	19.095	13.879	5911
E_1	**28.166**	**17.619**	7642
E_2	25.31	13.067	3433
E_3	26.714	13.067	3469

Table 1. Different kinds of features derived from prepositional phrases involving target nouns.

On both classifiers and for both types D and E, the performance is noticeably higher when the collocation of the noun with the preposition is used as one single feature (D_1 and E_1). Using only the nouns as separate features decreases classification accuracy. Adding the prepositions to them as individual features improves the performance very slightly on Naïve Bayes, but has no influence on the performance of Rocchio. Comparing types D_1 and E_1, we see that D_1 is clearly more effective, particularly on Naïve Bayes, and uses around 30% less features than E_1.

	NB	Rocchio	#dim
A	21.052	15.075	1533
B	34.88	**29.889**	4039
C	36.357	**28.242**	4607
D_1	23.405	16.574	11271
E_1	28.166	17.619	7642
Window	**38.261**	18.767	35902

Table 2. Syntactically-defined types of features.

Table 2 describes the results of the evaluation of all the five feature types described above. On Naïve Bayes, each of the syntactically-defined types yields performance inferior to that of the window-based features. On Rocchio, window-based is much worse than B and C, but is comparable to A, D_1 and E_1. Looking at the dimensionality of the feature space each method produces, we see that the window-based features are much more numerous than any of the syntactically-defined ones, although collected from the same corpus. The much larger feature however space does not yield a proportional increase in classification accuracy. For example, there are around seven times less type C features than window-based ones, but they are only 1.9% less effective on Naïve Bayes and significantly more effective on Rocchio.

Among the syntactically-defined features, types B and C perform equally well, no statistical significance between their performances was found on either NB or Rocchio. In fact, the ranking of the feature types wrt their performance is the same for both classifiers: types B and C trail E_1 by a large margin, which is followed by D_1, type A being the worst performer. The results so far suggest that adjectives and verbs near which target nouns are used as objects provide the best evidence about the target nouns' meaning.

We further tried collapsing different types of features together. In doing so, we appended a tag to each feature describing its type so as to avoid confusing context words linked by different syntactic relations to the target noun (see Table 3). The best result was achieved by combining all the five syntactic feature types, clearly outperforming the window-based context delineation on both Naïve Bayes (26% improvement, $p<0.05$) and Rocchio (88% improvement, $p<0.001$) and still using 20% smaller feature space. The combination of B and C produced only slightly worse results (the differences not significant for either classifiers), but using over 3 times smaller feature space.

	NB	Rocchio	#dim
B+C	43.071	35.426	8646
B+C+D_1+E_1	47.357	36.469	27559
A+B+C+D_1+E_1	**48.309**	**36.829**	29092
D_1+E_1	30.095	22.26	18913
Window	38.261	18.767	35902

Table 3. Combinations of syntactically-defined feature types.

4.2 Original word forms vs. stems vs. lemmas

We next looked at the performance resulting from stemming and lemmatization of context words. Since morphological preprocessing is likely to differently affect nouns, verbs, and adjectives, we study them on data of types B (verbs), C (adjectives), and the combination of D_1 and E_1 (nouns) from the previous experiment. Stemming

was carried out using the Porter stemmer. Lemmatization was performed using a pattern-matching algorithm which operates on PoS-tagged text and consults the WordNet database for exceptions. As before, context words that occurred only once with a target noun were discarded. Table 4 describes the results of these experiments.

	NB	Rocchio	#dim
Verbs			
Original	35.333	31.648	9906
Stem	35.357	27.665	7506
Lemma	34.88	29.889	4039
Adjectives			
Original	37.309	28.911	4765
Stem	36.833	29.168	4390
Lemma	36.357	28.242	4607
Nouns			
Original	28.69	23.076	19628
Stem	29.19	22.176	19141
Lemma	20.976	22.26	15642

Table 4. Morphological preprocessing of verbs, adjectives, and nouns.

There is very little difference in effectiveness between these three methods (except for lemmatized nouns on NB). As a rule, the difference between them is never greater than 1%. In terms of the size of feature space, lemmatization is most advisable for verbs (32% reduction of feature space compared with the original verb forms), which is not surprising since the verb is the most infected part of speech in English. The feature space reduction for nouns was around 25%. Least reduction of feature space occurs when applying lemmatization to adjectives, which inflect only for degrees of comparison.

4.3 Morphological decomposition

We further tried constructing features for a target noun on the basis of morphological analysis of words occurring in its context. As in the experiments with stemming and lemmatization, in order to take into account morphological differences between parts of speech, the effects of morphological decomposition of context words was studied on the distributional data of types B (verbs), C (adjectives), and D_1+E_1 (nouns).

The decomposition of words into morphemes was carried out as follows. From "*Merriam-Webster's Dictionary of Prefixes, Suffixes, and Combining Forms*"[1] we extracted a list of 12 verbal, 59 adjectival and 138 nounal suffixes, as well as 80 prefixes, ignoring affixes consisting of only one character. All suffixes for a particular part-of-speech and all prefixes were sorted according to their character length. First, all context words were lemmatized. Then, examining the part-of-speech of the context word, presence of each affix with it was checked by simple string matching, starting from the top of the corresponding array of affixes. For each word, only one prefix and only one suffix was matched. In this way, every word was broken down into maximum three morphemes: the root, a prefix and a suffix.

Two kinds of features were experimented with: one where features corresponded to the roots of the context words and one where all morphemes of the context word (i.e., the root, prefix and suffix) formed separate features. When combining features created from context words belonging to different parts-of-speech, no tags were used in order to map roots of cognate words to the same feature. The results of these experiments are shown in Table 5.

	roots	roots+ affixes	lemmas
Naïve bayes			
B	**37.261**	35.833	34.88
C	38.738	**39.214**	36.357
D_1+E_1	29.119	25.785	**30.095**
B+C	**43.976**	42.071	43.547
B+C+D_1+E_1	46.88	45.452	**48.309**
Rocchio			
B	24.241	24.061	**29.889**
C	27.803	27.901	**28.242**
D_1+E_1	13.267	12.87	**22.26**
B+C	28.747	28.019	**35.426**
B+C+D_1+E_1	28.863	30.752	**36.469**

Table 5. Distributional features derived from the morphological analysis of context words.

On Naïve Bayes, using only roots increases the classification accuracy for B, C, and B+C compared to the use of lemmas. The improvement, however, is not significant. Inclusion of affixes does not produce any perceptible effect on the performance. In all other cases and when the Rocchio classifier is used, decomposition of words into morphemes consistently decreases performance compared to the use of their lemmas.

These results seem to suggest that the union of the root with the affixes constitutes the most

[1] Available at www.spellingbee.com/pre_suf_comb.pdf

optimal "container" for distributional information. Decomposition of words into morphemes often causes loss of a part of this information. It seems there are few affixes with the meaning so abstract that they can be safely discarded.

4.4 Filtering out rare context words

To study the effect of removing singleton context words, we compared the quality of classifications with and without them. The results are shown in Table 6.

	NB	Rocchio	#dim
Without singletons			
B	34.88	**29.889**	4039
C	36.357	28.242	4607
B+C	43.547	**35.426**	8646
A+B+C+D_1+E_1	**48.309**	**36.829**	29092
Window	38.261	18.767	35902
With singletons			
B	**38.361**	25.164	14024
C	**39.261**	**28.387**	9898
B+C	**45.**	29.535	23922
A+B+C+D_1+E_1	44.	25.31	98703
Window	**41.142**	**19.037**	94606

Table 4: The effect of removing rare context words.

The results do not permit making any conclusions as to the enhanced effectiveness resulting from discarding rare co-occurrences. Discarding singletons, however, does considerably reduce the feature space. The dimensionality reduction is especially large for the datasets involving types B, D_1 and E_1, where each feature is a free collocation of a noun or a verb with a preposition, whose multiple occurrences are much less likely than multiple occurrences of an individual context word.

5 Related work

A number of previous studies compared different kinds of morphological and syntactic preprocessing performed before inducing a co-occurrence model of word meaning.

Grefenstette (1993) studied two context delineation methods of English nouns: the window-based and the syntactic, whereby all the different types of syntactic dependencies of the nouns were used in the same feature space. He found that the syntactic technique produced better results for frequent nouns, while less frequent nouns were more effectively modeled by the windowing technique. He explained these results by the fact that the syntactic technique extracts much fewer albeit more useful features and the small number of features extracted for rare nouns is not sufficient for representing their distributional behavior.

Alfonseca and Manandhar (2002) compared different types of syntactic dependencies of a noun as well as its "topic signature", i.e. the features collected by taking the entire sentence as the context of its occurrence, in terms of their usefulness for the construction of its distributional representation. They found that the best effectiveness is achieved when using a combination of the topic signature with the "object signature" (a list of verbs and prepositions to which the target noun is used as an argument) and the "subject signature" (a list of verbs to which the noun is used as a subject). The "modifier signature" containing co-occurring adjectives and determiners produced the worst results.

Pado and Lapata (2003) investigated different possibilities to delimit the context of a target word by considering the syntactic parse of the sentence. They examined the informativeness of features arising from using the window-based context delineation, considering the sum of dependencies the target word is involved in, and considering the entire argument structure of a verb as the context of the target word, so that, e.g. an object can be a feature for a subject of that verb. Their study discovered that indirect syntactic relations within an argument structure of a verb generally yield better results than using only direct syntactic dependencies or the windowing technique.

Ciaramita (2002) looked at how the performance of automatic classifiers on the word classification task is affected by the decomposition of target words into morphologically relevant features. He found that the use of suffixes and prefixes of target nouns is indeed more advantageous, but this was true only when classifying words into large word classes. These classes are formed on the basis of quite general semantic distinctions, which are often reflected in the meanings of their affixes. In addition to that, the classification method used involved feature selection, which ensured that useless features resulting from semantically empty affixes and errors of the morphological decomposition did not harm the classification accuracy.

6 Conclusion

In this study we examined the impact which linguistic preprocessing of distributional data produce on the effectiveness and efficiency of semantic classification of nouns.

Our study extends previous work along the following lines. First, we have compared different

types of syntactic dependencies of the target noun in terms of the informativeness of the distributional features constructed from them. We find that the most useful dependencies are the adjectives and nouns used as attributes to the target nouns and the verbs near which the target nouns are used as direct or prepositional objects. The most effective representation overall is obtained when using all the syntactic dependencies of the noun. We find that it is clearly more advantageous than the windowing technique both in terms of effectiveness and efficiency. The combination of the attribute and object dependencies also produces very good classification accuracy, which is only insignificantly worse than that of the combination of all the dependency types, while using several times more compact feature space.

We further looked at the influence of stemming and lemmatization of context words on the performance. The study did not reveal any considerable differences in effectiveness obtained by stemming or lemmatization of context words versus the use of their original forms. Lemmatization, however, allows to achieve the greatest reduction of the feature space. Similarly, the removal of rare word co-occurrences from the training data could not be shown to consistently improve effectiveness, but was very beneficial in terms of dimensionality reduction, notably for features corresponding to word collocations.

Finally, we examined whether morphological decomposition of context words helps to obtain more informative features, but found that indiscriminative decomposition of all context words into morphemes and using them as separate features actually more often decreases performance rather than increases it. These results seem to indicate that morphological analysis of context words should be accompanied by some feature selection procedure, which would identify those affixes which are too general and can be safely stripped off and those which are sufficiently specific and whose unity with the root best captures relevant context information.

7 Acknowledgements

The research was supported by the Russian Foundation Basic Research grant #03-06-80008. We thank our colleauges Steffen Staab and Andreas Hotho for fruitful discussions during the work on this paper.

References

E. Alfonseca and S. Manandhar. 2002. Extending a lexical ontology by a combination of distributional semantics signatures. In *Proceedings of 13th International Conference on Knowledge Engineering and Knowledge Management*, pages 1-7.

S. Caraballo. 1999. Automatic construction of a hypernym-labeled noun hierarchy from text. In *Proceedings of ACL'99*, pages 120-126.

M. Ciaramita. 2003. Boosting automatic lexical acquisition with morphological information. In *Proceedings of the ACL'02 Workshop on Unsupervised Lexical Acquisition*, pages 17-25.

P. Fung and L.Y. Yee. An IR approach for translating new words from nonparallel, comparable texts In *Proceedings of COLING-ACL'98*, pages 414-420.

G. Grefenstette. 1993. Evaluation techniques for automatic semantic extraction: comparing syntactic and window based approaches. In *Proceedings of the SIGLEX Workshop on Acquisition of Lexical Knowledge from Text*, Columbus, Ohio.

M. Kay and M. Röscheisen. 1993. Text-translation alignment. *Computational Linguistics.* 19(1):121-142.

J. Levy, J. Bullinaria, and M. Patel. 1998. Explorations in the derivation of word co-occurrence statistics. *South Pacific Journal of Psychology*, 10(1), 99-111.

D. Lewis. 1998. Naive (Bayes) at forty: The independence assumption in information retrieval. In *Proceedings of ECML'98*, pages 4-15.

D. Lin. 1998. Automatic retrieval and clustering of similar words. In *Proceedings of the COLING-ACL'98*, pages 768-773.

S. Pado and M. Lapata. 2003. Constructing semantic space models from parsed corpora. In *Proceedings of ACL'03*, pages 128-135.

V. Pekar and S. Staab. 2002. Factoring the structure of a taxonomy into a semantic classification decision. In: *Proceedings of COLING'02*, pages 786-792.

E. Riloff and M. Schmelzenbach. 1998. An empirical approach to conceptual case frame acquisition. In: *Proceedings of the 6th Workshop on Very Large Corpora*.

F. Sebastiani. 2002. Machine learning in automated text categorization. *ACM Computing Surveys*, 34(1): 1-47.

T. Tokunaga, A. Fujii, M. Iwayama, N. Sakurai, and H. Tanaka. 1997. Extending a thesaurus by classifying words. In *Proceedings of the ACL-EACL Workshop on Automatic Information Extraction and Building of Lexical Semantic Resources*, pages 16-21.