

Ensemble Methods for Automatic Thesaurus Extraction

James R. Curran

Institute for Communicating and Collaborative Systems

University of Edinburgh

2 Buccleuch Place, Edinburgh EH8 9LW

United Kingdom

jamesc@cogsci.ed.ac.uk

Abstract

Ensemble methods are state of the art for many NLP tasks. Recent work by Banko and Brill (2001) suggests that this would not necessarily be true if very large training corpora were available. However, their results are limited by the simplicity of their evaluation task and individual classifiers.

Our work explores ensemble efficacy for the more complex task of automatic thesaurus extraction on up to 300 million words. We examine our conflicting results in terms of the constraints on, and complexity of, different contextual representations, which contribute to the sparseness and noise-induced bias behaviour of NLP systems on very large corpora.

1 Introduction

Ensemble learning is a machine learning technique that combines the output of several different classifiers with the goal of improving classification performance. The classifiers within the ensemble may differ in several ways, such as the learning algorithm or knowledge representation used, or data they were trained on. Ensemble learning has been successfully applied to numerous NLP tasks, including POS tagging (Brill and Wu, 1998; van Halteren et al., 1998), chunking (Tjong Kim Sang, 2000), word sense disambiguation (Pederson, 2000) and statistical parsing (Henderson and Brill, 1999). Dietterich (2000) presents a good introduction to ensemble methods.

Ensemble methods ameliorate learner bias by amortising individual classifier bias of over different systems. For an ensemble to be more effective than its constituents, the individual classifiers must have better than 50% accuracy and must produce *diverse* erroneous classifications (Dietterich, 2000). Brill and Wu (1998) call this complementary disagreement *complementarity*. Although ensembles are often effective on problems with small training sets, recent work suggests this may not be true as dataset size increases. Banko and Brill (2001) found that for confusion set disambiguation with corpora larger than 100 million words, the best individual classifiers outperformed ensemble methods.

One limitation of their results is the simplicity of the task and methods used to examine the efficacy of ensemble methods. However, both the task and applied methods are constrained by the ambitious use of one billion words of training material. Disambiguation is relatively simple because confusion sets are rarely larger than four elements. The individual methods must be inexpensive because of the computational burden of the massive training set, so they must perform limited processing of the training corpus and can only consider a fairly narrow context surrounding each instance.

We explore the value of ensemble methods for the more complex task of automatic thesaurus extraction, training on corpora of up to 300 million words. The increased complexity leads to results contradicting Banko and Brill (2001), which we explore using ensembles of different contextual complexity. This work emphasises the link between contextual complexity and the problems of representation sparseness and noise as corpus size increases, which in turn impacts on learner bias and ensemble efficacy.

2 Automatic Thesaurus Extraction

The development of large thesauri and semantic resources, such as WordNet (Fellbaum, 1998), has allowed lexical semantic information to be leveraged to solve NLP tasks, including collocation discovery (Pearce, 2001), model estimation (Brown et al., 1992; Clark and Weir, 2001) and text classification (Baker and McCallum, 1998).

Unfortunately, thesauri are expensive and time-consuming to create manually, and tend to suffer from problems of bias, inconsistency, and limited coverage. In addition, thesaurus compilers cannot keep up with constantly evolving language use and cannot afford to build new thesauri for the many sub-domains that NLP techniques are being applied to. There is a clear need for automatic thesaurus extraction methods.

Much of the existing work on thesaurus extraction and word clustering is based on the observations that related terms will appear in *similar* contexts. These systems differ primarily in their definition of “context” and the way they calculate similarity from the contexts each term appears in. Many systems extract co-occurrence and syntactic information from the words surrounding the target term, which is then converted into a vector-space representation of the contexts that each target term appears in (Pereira et al., 1993; Ruge, 1997; Lin, 1998b). Curran and Moens (2002b) evaluate thesaurus extractors based on several different models of context on large corpora. The context models used in our experiments are described in Section 3.

We define a *context relation* instance as a tuple (w, r, w') where w is a thesaurus term, occurring in a relation of type r , with another word w' in the sentence. We refer to the tuple (r, w') as an *attribute* of w . The relation type may be grammatical or it may label the position of w' in a context window: e.g. the tuple $(\text{dog}, \text{direct-obj}, \text{walk})$ indicates that the term *dog*, was the direct object of the verb *walk*. After the contexts have been extracted from the raw text, they are compiled into attribute vectors describing all of the contexts each term appears in. The thesaurus extractor then uses clustering or nearest-neighbour matching to select similar terms based on a vector similarity measure.

Our experiments use k -nearest-neighbour match-

```
(adjective, good) 2005
(adjective, faintest) 89
(direct-obj, have) 1836
(indirect-obj, toy) 74
(adjective, preconceived) 42
(adjective, foggiest) 15
```

Figure 1: Example attributes of the noun *idea*

ing for thesaurus extraction, which calculates the pairwise similarity of the target term with every potential synonym. Given n terms and up to m attributes for each term, the asymptotic time complexity of k -nearest-neighbour algorithm is $O(n^2m)$. We reduce the number of terms by introducing a minimum occurrence filter that eliminates potential synonyms with a frequency less than five.

3 Individual Methods

The individual methods in these ensemble experiments are based on different extractors of contextual information. All the systems use the JACCARD similarity metric and TTEST weighting function that were found to be most effective for thesaurus extraction by Curran and Moens (2002a).

The simplest and fastest contexts to extract are the word(s) surrounding each thesaurus term up to some fixed distance. These window methods are labelled $W(L_1R_1)$, where L_1R_1 indicates that window extends one word on either side of the target term. Methods marked with an asterisk, e.g. $W(L_1R_1^*)$, do not record the word’s position in the relation type.

The more complex methods extract grammatical relations using shallow statistical tools or a broad coverage parser. We use the grammatical relations extracted from the parse trees of Lin’s broad coverage principle-based parser, MINIPAR (Lin, 1998a) and Abney’s cascaded finite-state parser, CASS (Abney, 1996). Finally, we have implemented our own relation extractor, based on Grefenstette’s SEXTANT (Grefenstette, 1994), which we describe below as an example of the NLP system used to extract relations from the raw text.

Processing begins with POS tagging and NP/VP chunking using a Naïve Bayes classifier trained on the Penn Treebank. Noun phrases separated by prepositions and conjunctions are then concatenated, and the relation attaching algorithm is run on the sentence. This involves four passes over the sen-

Corpus	Sentences	Words
British National Corpus	6.2M	114M
Reuters Corpus Vol 1	8.7M	193M

Table 1: Training Corpora Statistics

tence, associating each noun with the modifiers and verbs from the syntactic contexts they appear in:

1. nouns with pre-modifiers (left to right)
2. nouns with post-modifiers (right to left)
3. verbs with subjects/objects (right to left)
4. verbs with subjects/objects (left to right)

This results in relations representing the contexts:

1. term is the subject of a verb
2. term is the (direct/indirect) object of a verb
3. term is modified by a noun or adjective
4. term is modified by a prepositional phrase

The relation tuple is then converted to root form using the Sussex morphological analyser (Minnen et al., 2000) and the POS tags are stripped. The relations for each term are collected together producing a context vector of attributes and their frequencies in the corpus. Figure 1 shows the most strongly weighted attributes and their frequencies for *idea*.

4 Experiments

Our experiments use a large quantity of text which we have grouped into a range of corpus sizes. The approximately 300 million word corpus is a random conflation of the BNC and the Reuters corpus (respective sizes in Table 1). We then create corpus subsets down to $\frac{1}{128}$ th (2.3 million words) of the original corpus by randomly sentence selection.

Ensemble voting methods for this task are quite interesting because the result consists of an ordered set of extracted synonyms rather than a single class label. To test for subtle ranking effects we implemented three different methods of combination:

MEAN mean rank of each term over the ensemble;

HARMONIC the harmonic mean rank of each term;

MIXTURE ranking based on the mean score for each term. The individual extractor scores are not normalised because each extractor uses the same similarity measure and weight function.

We assigned a rank of 201 and similarity score of zero to terms that did not appear in the 200 synonyms returned by the individual extractors. Finally, we build ensembles from all the available extractor methods (e.g. **MEAN**(*)) and the top three performing extractors (e.g. **MEAN**(3)).

To measure the complementary disagreement between ensemble constituents we calculated both the complementarity C and the Spearman rank-order correlation R_s .

$$C(A, B) = \left(1 - \frac{|\text{errors}(A) \cap \text{errors}(B)|}{|\text{errors}(A)|}\right) * 100\% \quad (1)$$

$$R_s(A, B) = \frac{\sum_i (r(A_i) - \bar{r}(A))(r(B_i) - \bar{r}(B))}{\sqrt{\sum_i (r(A_i) - \bar{r}(A))^2} \sqrt{\sum_i (r(B_i) - \bar{r}(B))^2}} \quad (2)$$

where $r(x)$ is the rank of synonym x . The Spearman rank-order correlation coefficient is the linear correlation coefficient between the rankings of elements of A and B . R_s is a useful non-parametric comparison for when the rank order is more relevant than the actual values in the distribution.

5 Evaluation

The evaluation is performed on thesaurus entries extracted for 70 single word noun terms. To avoid sample bias, the words were randomly selected from WordNet such that they covered a range of values for the following word properties:

frequency Penn Treebank and BNC frequencies;

number of senses WordNet and Macquarie senses;

specificity depth in the WordNet hierarchy;

concreteness distribution across WordNet subtrees.

Table 2 shows some of the selected terms with frequency and synonym set information. For each term we extracted a thesaurus entry with 200 potential synonyms and their similarity scores.

The simplest evaluation measure is direct comparison of the extracted thesaurus with a manually-created gold standard (Grefenstette, 1994). However, on smaller corpora direct matching is often too

Word	PTB Rank	PTB #	BNC #	Reuters #	Macquarie #	WordNet #	Min / Max	WordNet subtree roots
company	38	4076	52779	456580	8	9	3 / 6	entity, group, state
interest	138	919	37454	146043	12	12	3 / 8	abs., act, group, poss., state
problem	418	622	56361	63333	4	3	3 / 7	abs., psych., state
change	681	406	35641	55081	8	10	2 / 12	abs., act, entity, event, phenom.
house	896	223	47801	45651	10	12	3 / 6	act, entity, group
idea	1227	134	32754	13527	10	5	3 / 7	entity, psych.
opinion	1947	78	9122	16320	4	6	4 / 8	abs., act, psych.
radio	2278	59	9046	20913	2	3	6 / 8	entity
star	5130	29	8301	6586	11	7	4 / 8	abs., entity
knowledge	5197	19	14580	2813	3	1	1 / 1	psych.
pants	13264	5	429	282	3	2	6 / 9	entity
tightness	30817	1	119	2020	5	3	4 / 5	abs., state

Table 2: Examples of the 70 thesaurus evaluation terms

coarse-grained and thesaurus coverage is a problem. To help overcome limited coverage, our evaluation uses a combination of three electronic thesauri: the topic-ordered Macquarie (Bernard, 1990) and Roget’s (Roget, 1911) thesauri and the head ordered Moby (Ward, 1996) thesaurus. Since the extracted thesaurus does not separate senses we transform Roget’s and Macquarie into head ordered format by collapsing the sense sets containing the term. For the 70 terms we create a gold standard from the union of the synonym lists of the three thesauri, resulting in a total of 23,207 synonyms.

With this gold standard resource in place, it is possible to use precision and recall measures to evaluate the quality of the extracted thesaurus. To help overcome the problems of coarse-grained direct comparisons we use several measures of system performance: direct matches (DIRECT), inverse rank (INVR), and top n synonyms precision ($P(n)$).

INVR is the sum of the inverse rank of each matching synonym, e.g. gold standard matches at ranks 3, 5 and 28 give an inverse rank score of $\frac{1}{3} + \frac{1}{5} + \frac{1}{28} \approx 0.569$. With at most 200 synonyms, the maximum INVR score is 5.878. Top n precision is the percentage of matching synonyms in the top n extracted synonyms. We use $n = 1, 5$ and 10.

6 Results

Figure 2 shows the performance trends for the individual extractors on corpora ranging from 2.3 million up to 300 million words. The best individual context extractors are SEXTANT, MINIPAR and $W(L_1R_1)$, with SEXTANT outperforming MINIPAR beyond approximately 200 million words. These

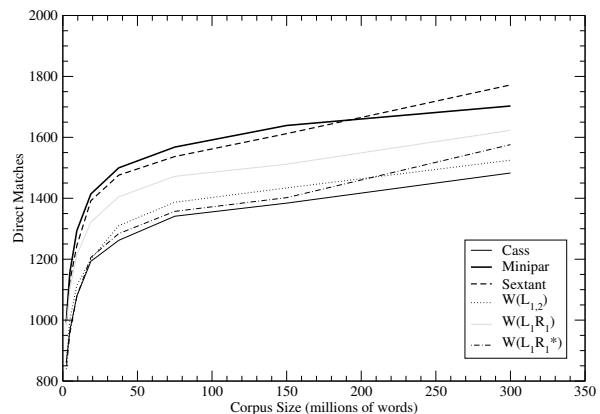


Figure 2: Single extractor performance to 300Mw

three extractors are combined to form the top-three ensemble. CASS and the other window methods perform significantly worse than SEXTANT and MINIPAR. Interestingly, $W(L_1R_1^*)$ performs almost as well as $W(L_1R_1)$ on larger corpora, suggesting that position information is not as useful with large corpora, perhaps because the left and right set of words for each term becomes relatively disjoint.

Table 3 presents the evaluation results for all the individual extractors and the six ensembles on the full corpus. At 300 million words all of the ensemble methods outperform the individual extractors. These results disagree with those Banko and Brill (2001) obtained for confusion set disambiguation. The best performing ensembles, MIXTURE(*) and MEAN(*), combine the results from all of the individual extractors. MIXTURE(*) performs approximately 5% better than SEXTANT, the best individual extractor. Figure 3 compares the performance behaviour over the range of corpus sizes for the best three individ-

System	DIRECT	P(1)	P(5)	P(10)	INVR
CASS	1483	50%	41%	33%	1.58
MINIPAR	1703	59%	48%	40%	1.86
SEXTANT	1772	61%	47%	39%	1.87
W(L _{1,2})	1525	54%	43%	37%	1.68
W(L ₁ R ₁)	1623	57%	46%	38%	1.76
W(L ₁ R ₁ *)	1576	63%	44%	38%	1.78
MEAN(*)	1850	66%	50%	43%	2.00
MEAN(3)	1802	63%	50%	44%	1.98
HARMONIC(*)	1821	64%	51%	43%	2.00
HARMONIC(3)	1796	63%	51%	43%	1.96
MIXTURE(*)	1858	64%	52%	44%	2.03
MIXTURE(3)	1794	63%	51%	44%	1.99

Table 3: Extractor performance at 300Mw

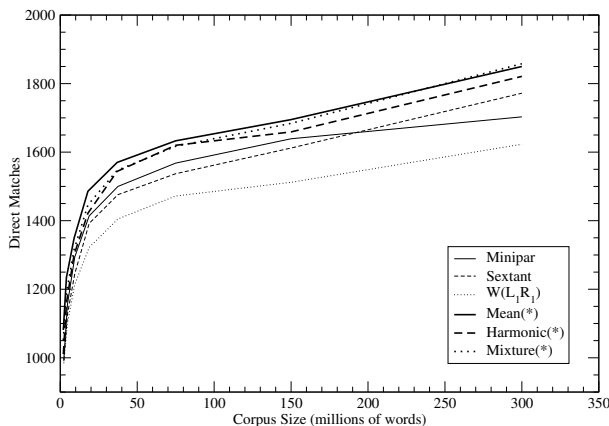


Figure 3: Ensemble performance to 300Mw

ual methods and the full ensembles. SEXTANT is the only competitive individual method as the corpus size increases. Figure 3 shows that ensemble methods are of more value (at least in percentage terms) for smaller training sets. The trend in the graph suggests that the individual extractors will not outperform the ensemble methods, unless the behaviour changes as corpus size is increased further.

From Table 3 we can also see that full ensembles, combining all the individual extractors, outperform ensembles combining only the top three extractors. This seems rather surprising at first, given that the other individual extractors seem to perform significantly worse than the top three. It is interesting to see how the weaker methods still contribute to the ensembles performance.

Firstly, for thesaurus extraction, there is no clear concept of *accuracy greater than 50%* since it is not a simple classification task. So, although most of the evaluation results are significantly less than 50%,

Ensemble	R_s	C
Ensemble(*) on 2.3M words	0.467	69.2%
Ensemble(3) on 2.3M words	0.470	69.8%
Ensemble(*) on 300M words	0.481	54.1%
Ensemble(3) on 300M words	0.466	51.2%

Table 4: Agreement between ensemble members

System	CASS	MINI	SEXT	W(L _{1,2})	W(L ₁ R ₁)	W(L ₁ R ₁ *)
CASS	0%	58%	59%	65%	63%	69%
MINI	57%	0%	47%	57%	54%	60%
SEXT	58%	47%	0%	54%	53%	58%
W(L _{1,2})	65%	58%	55%	0%	40%	43%
W(L ₁ R ₁)	63%	54%	54%	39%	0%	33%
W(L ₁ R ₁ *)	69%	60%	58%	43%	33%	0%

Table 5: Complementarity for extractors

this does not represent a failure of a necessary condition of ensemble improvement. If we constrain the thesaurus extraction to selecting a single synonym classification using the P(1) scores, then all of the methods achieve 50% or greater accuracy. Considering the complementarity and rank-order correlation coefficients for the constituents of the different ensembles proves to be more informative. Table 4 shows these values for the smallest and largest corpora and Table 5 shows the pairwise complementarity for the ensemble constituents.

It turns out that the average Spearman rank-order correlation is not sensitive enough to errors for the purposes of comparing favourable disagreement within ensembles. However, the average complementarity clearly shows the convergence of the ensemble constituents, which partially explains the reduced efficacy of ensemble methods for large corpora. Since the top-three ensembles suffer this to a greater degree, they perform significantly worse at 300 million words. Further, the full ensembles can amortise the individual biases better since they average over a larger number of ensemble methods with different biases.

7 Analysis

Understanding ensemble behaviour on very large corpora is important because ensemble classifiers are state of the art for many NLP tasks. This section explores possible explanations for why our results disagree with Banko and Brill (2001).

Thesaurus extraction and confusion set disam-

biguation are quite different tasks. In thesaurus extraction, contextual information is collected from the entire corpus into a single description of the environments that each term appears in and classification, as such, involves comparing these collections of data. In confusion set disambiguation on the other hand, each instance must be classified individually with only a limited amount of context. The disambiguator has far less information available to determine each classification. This has implications for representation sparseness and noise that a larger corpus helps to overcome, which in turn, affects the performance of ensemble methods against individual classifiers.

The complexity of the contextual representation and the strength of the correlation between target term and the context also plays a significant role. Curran and Moens (2002b) have demonstrated that more complex and constrained contexts can yield superior performance, since the correlation between context and target term is stronger than simple window methods. Further, structural and grammatical relation methods can encode extra syntactic and semantic information in the relation type. Although the contextual representation is less susceptible to noise, it is often sparse because fewer context relations are extracted from each sentence.

The less complex window methods exhibit the opposite behaviour. Depending on the window parameters, the context relations can be poorly correlated with the target term, and so we find a very large number of irrelevant relations with low and unstable frequency counts, that is, a noisy contextual representation. Since confusion set disambiguation uses limited contexts from single occurrences, it is likely to suffer the same problems as the window thesaurus extractors.

To evaluate an ensemble’s ability to reduce the data sparseness and noise problems suffered by different context models, we constructed ensembles based on context extractors with different levels of complexity and constraints.

Table 6 shows the performance on the full corpus for the three syntactic extractors, the top three performing extractors and their corresponding mean rank ensembles. For these more complex and constrained context extractors, the ensembles continue to outperform individual learners, since the context representation are still reasonably sparse. The aver-

System	DIRECT	P(1)	P(5)	P(10)	INVR
CASS	1483	50%	41%	33%	1.58
MINIPAR	1703	59%	48%	40%	1.86
SEXTANT	1772	61%	47%	39%	1.87
MEAN(P)	1803	60%	48%	42%	1.89
$W(L_1R_1)$	1623	57%	46%	38%	1.76
MINIPAR	1703	59%	48%	40%	1.86
SEXTANT	1772	61%	47%	39%	1.87
MEAN(3)	1802	63%	50%	44%	1.98

Table 6: Complex ensembles better than individuals

System	DIRECT	P(1)	P(5)	P(10)	INVR
$W(L_1)$	1566	59%	42%	35%	1.70
$W(L_2)$	1235	44%	36%	31%	1.38
$W(R_1)$	1198	44%	28%	24%	1.19
$W(R_2)$	1200	49%	30%	24%	1.25
MEAN($D_{1 2}$)	1447	54%	46%	37%	1.74
$W(L_{1,2})$	1525	54%	43%	37%	1.68
$W(L_1R_1)$	1623	57%	46%	38%	1.76
$W(R_{1,2})$	1348	53%	32%	29%	1.40
MEAN($D_{1,2}$)	1550	63%	46%	39%	1.81
$W(L_{1,2}^*)$	1500	50%	41%	36%	1.60
$W(L_1R_1^*)$	1576	63%	44%	38%	1.78
$W(R_{1,2}^*)$	1270	46%	29%	27%	1.28
MEAN($D_{1,2}^*$)	1499	64%	46%	39%	1.82

Table 7: Simple ensembles worse than individuals

age complementarity is greater than 50%.

Table 7 shows the performance on the full corpus for a wide range of window-based extractors and corresponding mean rank ensembles. Most of the individual learners perform poorly because the extracted contexts are only weakly correlated with the target terms. Although the ensemble performs better than most individuals, they fail to outperform the best individual on direct match evaluation. Since the average complementarity for these ensembles is similar to the methods above, we must conclude that it is a result of the individual methods themselves. In this case, the most correlated context extractor, e.g. $W(L_1R_1)$, extracts a relatively noise free representation which performs better than amortising the bias of the other noisy ensemble constituents.

Finally, confusion set disambiguation yields a single classification from a small set of classes, whereas thesaurus extraction yields an ordered set containing every potential synonym. The more flexible set of ranked results allow ensemble methods to exhibit more subtle variations in rank than simply selecting a single class.

We can contrast the two tasks using the single syn-

onym, P(1), and rank sensitive, INVR, evaluation measures. The results for P(1) do not appear to form any trend, although the results show that ensemble methods do not always improve single class selection. However, if we consider the INVR measure, all of the ensemble methods outperform their constituent methods, and we see a significant improvement of approximately 10% with the MEAN(3) ensemble.

8 Conclusion

This paper demonstrates the effectiveness of ensemble methods for thesaurus extraction and investigates the performance of ensemble extractors on corpora ranging up to 300 million words in size. Contrary to work reported by Banko and Brill (2001), the ensemble methods continue to outperform the best individual systems for very large corpora. The trend in Figure 3 suggests that this may continue for corpora even larger than we have experimented with.

Further, this paper examines the differences between thesaurus extraction and confusion set disambiguation, and links ensemble efficacy to the nature of each task and the problems of representation sparseness and noise. This is done by evaluating ensembles with varying levels of contextual complexity and constraints.

The poorly constrained window methods, where contextual correlation is often low, outperformed the ensembles, which parallels results from (Banko and Brill, 2001). This suggests that large training sets ameliorate the predominantly noise-induced bias of the best individual learner better than amortising the bias over many similar ensemble constituents. Noise is reduced as occurrence counts stabilise with larger corpora, improving individual classifier performance, which in turn causes ensemble constituents to converge, reducing complementarity. This reduces the efficacy of classifier combination and contributes to individual classifiers outperforming the ensemble methods.

For more complex, constrained methods the same principles apply. Since the correlation between context and target is much stronger, there is less noise in the representation. However, the added constraints reduce the number of contextual relations extracted from each sentence, leading to data

sparseness. These factors combine so that ensemble methods continued to outperform the best individual methods.

Finally, corpus size must be considered with respect to the parameters of the contextual representation extracted from the corpus. The value of larger corpora is partly dependent on how much information is extracted from each sentence of training material. We fully expect individual thesaurus extractors to eventually outperform ensemble methods as sparseness and complementarity are reduced, but this is not true for 100 or 300 million words since the best performing representations extract very few contexts per sentence.

We would like to further investigate the relationship between contextual complexity, data sparseness, noise and learner bias on very large corpora. This includes extending these experiments to an even larger corpus with the hope of establishing the cross over point for thesaurus extraction. Finally, although wider machine learning research uses large ensembles, many NLP ensembles use only a handful of classifiers. It would be very interesting to experiment with a large number of classifiers using bagging and boosting techniques on very large corpora.

Acknowledgements

We would like to thank Miles Osborne for initial discussions which led to this work, and Marc Moens, Steve Finch and Tara Murphy for their feedback on drafts of this paper. This research is supported by a Commonwealth scholarship and a Sydney University Travelling scholarship.

References

- Steve Abney. 1996. Partial parsing via finite-state cascades. *Journal of Natural Language Engineering*, 2(4):337–344, December.
- L. Douglas Baker and Andrew McCallum. 1998. Distributional clustering of words for text classification. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and Development in Information Retrieval*, pages 96–103, Melbourne, Australia, 24–28 August.
- Michele Banko and Eric Brill. 2001. Scaling to very very large corpora for natural language disambiguation. In *Proceedings of the 39th annual meeting of the*

- Association for Computational Linguistics*, pages 26–33, Toulouse, France, 9–11 July.
- John R. L. Bernard, editor. 1990. *The Macquarie Encyclopedic Thesaurus*. The Macquarie Library, Sydney, Australia.
- Eric Brill and Jun Wu. 1998. Classifier combination for improved lexical disambiguation. In *Proceedings of the 17th International Conference on Computational Linguistics and of the 36th Annual Meeting of the Association for Computational Linguistics*, pages 191–195, Montréal, Québec, Canada, 10–14 August.
- Peter F. Brown, Vincent J. Della Pietra, Peter V. deSouza, Jennifer C. Lai, and Robert L. Mercer. 1992. Class-based n-gram models of natural language. *Computational Linguistics*, 18(4):467–479, December.
- Stephen Clark and David Weir. 2001. Class-based probability estimation using a semantic hierarchy. In *Proceedings of the Second Meeting of the North American Chapter of the Association for Computational Linguistics*, pages 95–102, Pittsburgh, PA USA, 2–7 June.
- James R. Curran and Marc Moens. 2002a. Improvements in automatic thesaurus extraction. In *ACL-SIGLEX Workshop on Unsupervised Lexical Acquisition*, Philadelphia, PA USA, 12 July. (to appear).
- James R. Curran and Marc Moens. 2002b. Scaling context space. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, Philadelphia, PA USA, 7–12 July. (to appear).
- Thomas G. Dietterich. 2000. Ensemble methods in machine learning. In *Proceedings of the First International Workshop on Multiple Classifier Systems (LNCS 1857)*, pages 1–15. Springer-Verlag, Cagliari, Sardinia, Italy.
- Cristiane Fellbaum, editor. 1998. *WordNet: an electronic lexical database*. The MIT Press, Cambridge, MA USA.
- Gregory Grefenstette. 1994. *Explorations in Automatic Thesaurus Discovery*. Kluwer Academic Publishers, Boston, USA.
- John C. Henderson and Eric Brill. 1999. Exploiting diversity in natural language processing: Combining parsers. In *Proceedings of the Fourth Conference on Empirical Methods in Natural Language Processing (EMNLP-99)*, pages 187–194, College Park, Maryland, USA.
- Dekang Lin. 1998a. Dependency-based evaluation of MINIPAR. In *Workshop on the Evaluation of Parsing Systems, Proceedings of the First International Conference on Language Resources and Evaluation*, pages 234–241, Granada, Spain, 28–30 May.
- Dekang Lin. 1998b. An information-theoretic definition of similarity. In *Proceedings of the Fifteen International Conference on Machine Learning*, pages 296–304, Madison, WI USA, 24–27 July.
- Guido Minnen, John Carroll, and Darren Pearce. 2000. Robust applied morphological generation. In *In Proceedings of the First International Natural Language Generation Conference*, pages 201–208, 12–16 June.
- Darren Pearce. 2001. Synonymy in collocation extraction. In *Workshop on WordNet and Other Lexical Resources: Applications, Extensions and Customizations, (NAACL 2001)*, pages 41–46, Pittsburgh, PA USA, 2–7 June.
- Ted Pederson. 2000. A simple approach to building ensembles of naive bayesian classifiers for word sense disambiguation. In *Proceedings of the 6th Applied Natural Language Processing Conference and the 1st Meeting of the North American Chapter of the Association of Computational Linguistics*, pages 63–69, Seattle, WA USA, 29 April–4 May.
- Fernando Pereira, Naftali Tishby, and Lillian Lee. 1993. Distributional clustering of English words. In *Proceedings of the 31st annual meeting of the Association for Computational Linguistics*, pages 183–190, Columbus, Ohio USA, 22–26 June.
- Peter Roget. 1911. *Thesaurus of English words and phrases*. Longmans, Green and Co., London, UK.
- Gerda Ruge. 1997. Automatic detection of thesaurus relations for information retrieval applications. In *Foundations of Computer Science: Potential - Theory - Cognition, Lecture Notes in Computer Science*, volume LNCS 1337, pages 499–506. Springer Verlag, Berlin, Germany.
- Erik F. Tjong Kim Sang. 2000. Noun phrase recognition by system combination. In *Proceedings of the Language Technology Joint Conference ANLP-NAACL2000*, pages 50–55, Seattle, Washington, USA, 29 April–4 May.
- Hans van Halteren, Jakob Zavrel, and Walter Daelemans. 1998. Improving data driven wordclass tagging by system combination. In *Proceedings of the 17th International Conference on Computational Linguistics and of the 36th Annual Meeting of the Association for Computational Linguistics*, pages 491–497, Montréal, Québec, Canada, 10–14 August.
- Grady Ward. 1996. *Moby Thesaurus*. Moby Project.