Using Lexical Semantic Knowledge from Machine Readable Dictionaries for Domain Independent Language Modelling

George Demetriou[1], **Eric Atwell**[2], **Clive Souter**[2]

[1]Department of Computer Science
University of Sheffield
211 Portobello Street, Sheffield S1 4DP, United Kingdom
G.Demetriou@dcs.shef.ac.uk

[2]School of Computer Studies
University of Leeds
Woodhouse Lane, Leeds LS2 9JT, United Kingdom
{eric,cs}@scs.leeds.ac.uk

Abstract

Machine Readable Dictionaries (MRDs) have been used in a variety of language processing tasks including word sense disambiguation, text segmentation, information retrieval and information extraction. In this paper we describe the utilization of semantic knowledge acquired from an MRD for language modelling tasks in relation to speech recognition applications. A semantic model of language has been derived using the dictionary definitions in order to compute the semantic association between the words. The model is capable of capturing phenomena of latent semantic dependencies between the words in texts and reducing the language ambiguity by a considerable factor. The results of experiments suggest that the semantic model can improve the word recognition rates in "noisy-channel" applications. This research provides evidence that limited or incomplete knowledge from lexical resources such as MRDs can be useful for domain independent language modelling.

1. Introduction

The use Machine Readable Dictionaries (MRDs) in Natural Language Processing (NLP) has been studied extensively for over a decade in the hope that online dictionaries might provide a way out of the knowledge acquisition bottleneck. Examples of NLP research using MRDs include amongst others sublanguage analysis (Walker and Amsler, 1987), knowledge acquisition and organisation (Alshawi, 1987; Calzolari and Picchi, 1988; Wilks et al, 1989; Kwong, 1998), word sense disambiguation (Lesk, 1986; Veronis and Ide, 1990; Guthrie et al, 1991; Bruce and Wiebe, 1994; Rigau et al, 1997), information retrieval (Krovetz and Croft, 1992), information extraction (Cowie et al, 1993) and text coherence (Kozima and Furugori, 1993).

On-line dictionaries seemed to offer the possibility for enormous savings in time and human and the problem changed from one of how to construct a knowledge resource to that of knowledge utilisation i.e. how to make the available knowledge really useful and efficient for NLP applications. However, research with MRDs so far has not fulfilled the prior expectations and many have criticized the dictionary knowledge as being vague, weak or incomplete while others have wondered whether the research in MRDs has been a 'waste of time' (Veronis and Ide 1994).

The usual paradigm of MRD research in computational linguistics has followed an extract and test approach, that is, extract the semantic information from the dictionary and test it on some set of text data gathered purposefully for the task. But, rarely in works reported in the literature the test data could satisfy the requirements of large scale investigations and, to our knowledge, no tests have been conducted to evaluate the effectiveness of the semantic knowledge from MRDs on large volumes of real texts. From the language engineering point of view and, regardless of the representation of the knowledge in the dictionary, the utility of the resource should be assessed by its ability to provide constraints in order to restrict or rank alternative hypotheses in NLP tasks.

In this paper we investigate whether or not the use of knowledge from MRDs can be a valid and workable method for introducing large-scale natural language constraints. We describe the development of a language model using the lexical semantic knowledge from an online dictionary, the Longman Dictionary of Contemporary English (Procter, 1978), and its application for word prediction tasks. The modelling of the semantic knowledge is based on the association between two words which is computed from the textual representations of their meanings in the dictionary with the use of an appropriate distance metric.

The assessment of the *satisfiability* and *diagnosticity* of the model's language constraints is an important consideration for language modelling. This assessment has been carried out by testing the efficiency of the constraints on large text samples taken from various genres of the British National Corpus (BNC) and estimating the reduction in lexical ambiguity.

We also examine the semantic associations in relation to the distance of the words in the text and determine the contextual range of the semantic constraints of the dictionary.

Finally, we evaluate the efficiency of the semantic model for recognition tasks by applying it to word lattices and sentence hypotheses generated from speech confusion data.

2. Dictionary Definitions and Semantic Associations

The semantic association between two words in our model can be described as the degree of semantic overlap or linkage between the meanings of the words in the dictionary. The semantic associations are computed by considering the dictionary sense definitions as sets of semantic primitives or concepts that are represented by the words in the definition. To avoid combinatorial phenomena between word senses and, because we intend to use the semantic knowledge for word prediction rather than word sense disambiguation, all different senses of a word are joined into a single definition of "meaning". If $S_1, S_2,..,S_n$ are different senses of a word x, then the "total" meaning of x is defined as

$$X = S_1 \bigcup S_2 ... \bigcup S_n$$

We can now specify an appropriate distance measure in order to quantify the degree of semantic association between two words. The measure used for this quantification is based on the simple matching coefficient (else called Jaccard coefficient; Jaccard, 1908) and is defined as follows.

Let x and y be two words in the dictionary whose meanings are represented by the sets X and Y respectively. The semantic association S between x and y is given by[1]

$$S = \frac{m\{X \bigcap Y\}}{m\{X \bigcup Y\}} \qquad (1)$$

i.e. the semantic association between x and y is the number of semantic primitives the definitions of x and y have in common divided by the total number of distinct semantic primitives between them. This measure takes values from 0 (no association) to 1 (total semantic overlap).

The semantic association measure for two words can be used as the basis for computing the *semantic associativity* of longer word sequences in texts, such as phrases, sentences or paragraphs.

Let W be a word sequence consisting of n elements i.e. $W=(W_1, W_2, ..., Wn)$. The Semantic Associativity (SA) of the words in W is defined as

$$SA(W) = \frac{1}{k} \sum_{i,j} S(W_i, W_j) \qquad i \neq j \qquad (2)$$

where S is the semantic association as defined in (1) and k is a normalization factor. Typical values of k are 1 (i.e. no normalization) or $C_2^n = (n-1)n/2$ (i.e. the semantic associativity of a word string can be interpreted as the mean semantic association of all pairwise word combinations in W).

Before computing the semantic associations some preprocessing was required to filter certain kinds of information in the dictionary. Firstly, all definitions for a lexical entry were merged, and common words or stopwords (such as 'a', 'the', 'of', etc.) in the definitions were removed as they cannot be good indicators of semantic relationships. Secondly, a lemmatisation procedure was used to conflate all remaining words to their root forms and multiple instances of the same semantic primitive were removed from the definitions. The definitions are also supplemented by a set of codes that are provided in the dictionary, i.e. subject codes (such as economics, engineering, etc) and codes indicating semantic selection restrictions or preferences between certain classes of words (verbs, nouns and adjectives).

3. Semantic Associativity of Word Combinatorics

The model derived from the dictionary can be used to provide language constraints for restricting alternative hypotheses in test data. However, no constraint can be useful unless there is some probability that the constraint will be satisfied by the data. On the other hand, for every constraint in the model there is a probability of this constraint being satisfied by randomly chosen hypotheses.

To quantitatively evaluate the effectiveness of the semantic model derived from the dictionary we have conducted experiments to measure the satisfiability and the diagnosticity of its constraints. We use the terms satisfiability and diagnosticity in the same way as in Lea (1980 p. 219). The satisfiability of a constraint measures the expected frequency for a test to yield positive results while its diagnosticity measures the amount of information the constraint adduces. In practice, it is often the case that there is a trade-off between satisfiability and diagnosticity. A highly diagnostic constraint would rule out a large number of competing hypotheses whereas a very general constraint would have low diagnosticity.

3.1. The Satisfiability of Semantic Associations

The estimation of the satisfiability of the semantic constraints is useful because it can answer the question of whether or not words grouped together in natural language texts exhibit stronger semantic associations than expected at random. The answer to such a question may be of no interest to the theoretical linguist or lexicographer. They can afford to assume that the dictionary represents a more or less a standard source for the meanings of the words. After all, dictionaries are meant for human users and they can use tremendous amounts of contextual knowledge to interpret the semantics of the words either for language comprehension or generation.

But the answer to the above question is very much of interest to the language engineer who would like to investigate whether or not a systematic relationship between semantic associativity and word co-occurrence in texts exists and, possibly, identify the strength of such a relationship.

To estimate how satisfiable the constraints are, we first needed to derive the distribution of semantic associations expected at random from the dictionary. For about 36,000 distinct entries in the dictionary, the semantic associations for about 650 million word pairs were computed using equation (1). We then extracted text samples from various genres[2] of the British National Corpus (30 samples

[1] The notation m{ } is used to denote the number of elements in the expression within m{ }.

[2] For the written part of the BNC the texts were from the genres: *leisure, social science, world affairs, arts, imaginative, applied science, natural science, commerce/finance, belief/thought.* For

extracted randomly from the BNC with approximately 1 million words in each sample) and computed the semantic associations for all possible word pairs within the sentences in the texts. Function words were excluded from these computations.

The cumulative frequencies of the semantic association values for the two distributions are plotted in figure 1 where the x-axis represents the cumulative values of percentages of the number of word pairs and the y-axis represents the value of the semantic association between the two words.

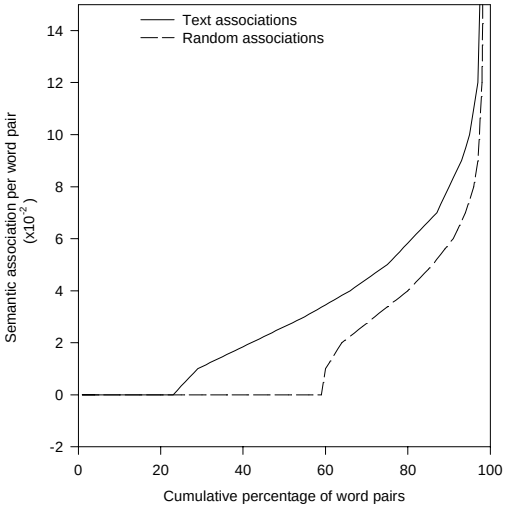

Figure 1: Cumulative frequency distributions of the semantic associations.

It can be seen that, although the two distributions generally have the same shape, there are considerable differences between the frequencies of the expected (random) and the observed (text) association values. For example, while about 60% of word pairs in the random model have no semantic association at all, only about 20% of word pairs in the texts are not semantically associated.

The average value of the semantic associations for word pairs observed in the texts was found to be about three times larger than what would have been expected at random (3.74×10^{-2} vs.1.27×10^{-2}). Parametric and non-parametric statistical tests (t-test, χ^2, log-likelihood, Mann-Whitney) have indicated significant differences between the two distributions at much better than 0.01 significance level.

These findings suggest that the word pairs in the texts exhibit significantly larger semantic associations than what would have been expected by randomly selected hypotheses. Consequently, the semantic model derived from the dictionary seems capable of capturing phenomena of *latent semantic associativity* between words in natural language. Although variations between the values of the semantic associations from genre to genre were observed, the satisfiability of the constraints was found be larger than random for all different genres of the BNC used in the experiments (see figure 2).

the spoken part the texts were from the genres: *leisure, educative/informative, public/institutional, business.*

3.2. Lexical Ambiguity and Semantic Entropy

The diagnosticity of the semantic constraints is estimated from the reduction in entropy when the model is used to discriminate between random and non-random hypotheses. The estimation of the entropy according to information theoretic criteria requires the approximation of the semantic model to a probabilistic model. For this reason, the semantic association values were 'normalized' so that they summed to 1 and the semantic model was treated as a bigram model in which the prediction of the next word depends on the knowledge or occurrence of the previous word in the text. As first order probabilities we used the unigram probabilities of the words in the texts.

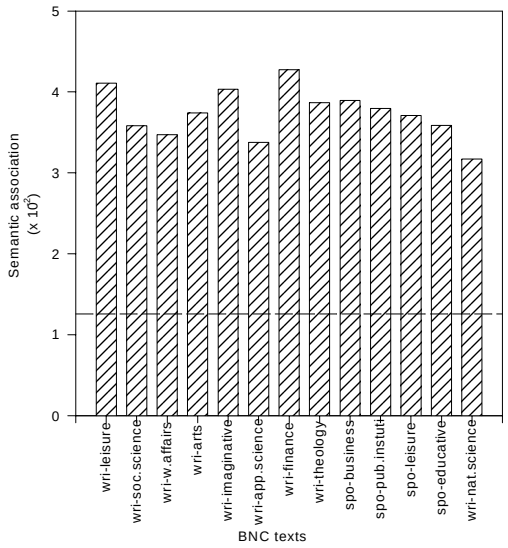

Figure 2: The semantic associations (mean values) for different text genres.

It should be pointed out that the probabilistic analogy would not generally hold in the case of semantic constraints used by humans because humans arguably do not use probabilistic processes for word prediction. Nevertheless, when a statistical model for random sampling behaviour against which to compare observed values is required, the probabilistic approximation has significant practical potential.

For a vocabulary of about 80,000 words, the semantic entropy of the dictionary model was found to be 8.8. The corresponding average branching factor or perplexity (Jelinek, 1990) was 446 whereas the perplexity of the unigram model was more than four times higher. This indicates that the semantic model can be used to reduce the ambiguity in natural language considerably. Although the semantic constraint of the model does not seem to be as strong as in powerful trigram models (where one would expect values of perplexity between 200 to 300), there is clear evidence that the model can be used to eliminate a large proportion of uncertainties in word discrimination tasks.

3.3. The Impact of Context on Semantic Associativity

It would be useful to analyze the relation of the semantic associations with respect to the distance of the words in the text. To carry out this analysis, the semantic

associations between word pairs were computed within a window of 100 words in the text.

The findings suggest that the semantic association between two words is somewhat sensitive to their distance in the text and, generally, the larger the distance, the smaller the association as can be seen from the graph of figure 3. The distance between two words is negatively correlated to their semantic association (as can be seen by the negative slope of the regression line). The value of the coefficient of determination r^2 was 0.758 which indicates that about 75% of the variation of semantic association can be predictable from the variation of distance between the words in the text.

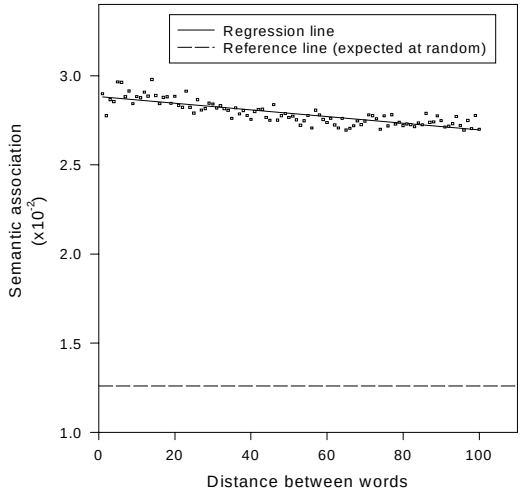

Figure 3: The effect of distance on the semantic associativity.

The semantic associations still have higher values (regression line) than what would have expected at random (reference line) for distances up to 100 words. Assuming that the same trend in the data exists, the maximum distance up to which the semantic associations can be expected to be better than random was estimated to be about 650 words for the texts used in the experiments. This indicates that the semantic model from the dictionary can capture not only short distance but also long distance semantic relationships between words and these relationships may extend beyond sentence boundaries.

The phenomenon of distant semantic relationships can be partly attributed to the thematic codes in the dictionary and partly to the distributional characteristics of domain specific words. For example, words such as "bank", "money" or "business" can occur very frequently in the financial domain and words such as "algorithm", "byte" or "software" in the computer science domain. When the topic of the text is quite specific, the occurrence of such 'keywords' can increase the probability for semantically associated words.

4. Semantic Associativity and Word Prediction Tasks

To test the efficiency of the semantic model for word prediction tasks, large vocabulary sentence hypotheses and word lattices were generated using phoneme confusion data acquired from a speech recognition front-end and a pronunciation lexicon. For the generation of sentence hypotheses we used 50 sentences for each of 13 genres of the BNC (650 sentences of about 15,000 words, 650,000 sentence hypotheses in total). Examples of sentence hypotheses are shown in figure 4.

```
how glider accidents happen
how glider accidents happen
how glide de accidents happen
how glider wrack Sid 'un happen
how guyed adder accident happen
how glider row Sid 'un happen
...
```

Figure 4: Examples of sentence hypotheses.

For each sentence hypothesis, the recognition accuracy (percentage of correct words in the hypothesis) was calculated. The semantic associativity score for the hypothesis was computed using equation (2) and the hypothesis with the best semantic score was assumed to be the most likely one.

The results of this evaluation suggest a considerable improvement in terms of word recognition accuracy for a wide range of baseline values (table 1).

Test set	Baseline (% correct)	Semantic model (% correct)	Increase (%)
1	47.5	72.8	25.4
2	52.5	75.1	22.6
3	63.9	79.1	15.1
4	76.2	84.2	7.9
5	83.8	88.6	4.8
6	92.2	93.7	1.5

Table 1: Recognition improvement on sentence hypotheses.

There was variation in the improvement of recognition performance from genre to genre but not as much to suggest that the model is biased towards a particular genre. With respect to the baseline accuracy the error reduction is decreased as the quality of sentence hypotheses in the test set increases. This is understandable because the "more correct" the sentence hypotheses, the more difficult for the model to discriminate between them. Nevertheless, even when the baseline accuracy is quite high (92% words correct), the improvement by using the semantic model can range from 0.6% (written leisure text) to 2.9% words correct (written belief/thought text).

In figure 5 (next page) we have plotted the semantic associativity scores against the percentage of the correct words in sentence hypotheses. The data suggest that there is an underlying trend towards better word prediction for larger values of semantic scores. The Pearson correlation coefficient indicates high positive correlation between the semantic and recognition scores at better than the required level of significance of 0.01. We can therefore conclude that, in the large-scale case, the probability of errors by using the model for the ranking or filtering of sentence hypotheses should be quite small.

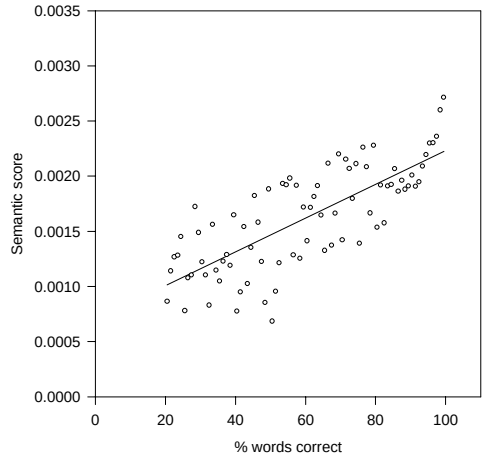

Figure 5: Semantic scores vs. word accuracies in sentence hypotheses.

It would be reasonable to assume that the longer the input sentence, the higher the probability of having semantically related words in the same sentence. Could this imply that the performance of the model is better for long utterances rather than short ones? To answer such a question the word error reduction rates were grouped for length intervals of 5 words. The results are shown in figure 6 where it can be seen that, generally, the longer the input utterance, the higher the reduction in error rate and the better the recognition performance. However, for utterances with 5 words or less we have a decrease of about 30% in recognition performance from the baseline. This can be explained by the fact that there are only a few

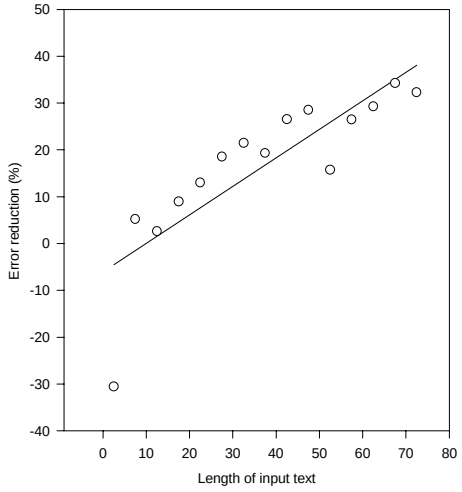

Figure 6: Error reduction with respect to input length.

content words in short sentences (usually not more than two in a five word sentence) and the fact that function words cannot be used for computing the semantic associations. As function words are often acoustically confused with short content words (e.g. “red” with “and”, “Ann” with “an”, etc.) and because there is not enough context to disambiguate them, occasional errors can occur. It should be noted however, that about one third of the small length sentences in the test sets included headings, titles or subtitles. The effect of such idiosyncratic text on the semantic associativity between words would require further investigation. The general conclusion from this experiment is that the more the context used, the more reliable the semantic associativity model for word prediction tasks.

The semantic model was also tested with word lattices (directed acyclic graphs of word hypotheses). Because the number of possible paths through a word lattice can be very high and cannot be searched exhaustively (the average number of paths was 1.38×10^{20} for our lattices), an efficient search strategy is required.

The algorithm we developed for parsing lattices incorporates a meaning-driven look-ahead search procedure for extending a partial path through the lattice. The algorithm combines characteristics of the A* algorithm (Nilsson, 1971) with a look-ahead evaluation function. The function evaluates the “promise” of a partial path with respect to the semantics of the word hypotheses that can be encountered next in the lattice in order to prune the list of all possible paths at each search step. The results of the application of this algorithm on word lattices from different text genres are given in table 2.

Text sample	Recognition rate (% correct words)	
	Baseline	Sem. algorithm
wri_leisure	12.5	44.2
wri_soc_science	15.6	60.1
wri_world_affairs	16.3	60.5
wri_arts	16.8	61.7
wri_imaginative	16.3	52.7
wri_app_science	14.3	49.7
wri_commerce/finan.	14.9	48.9
wri_theology	14.8	43.9
Spo_business	13.1	37.5
Spo_public/institut.	15.6	50.6
Spo_leisure	15.5	54.3
Spo_educative/infor.	17.6	62.4
wri_nat_science	14.0	52.5
ALL	15.3	53.8

Table 2: Recognition improvement for word lattices.

The results show that the application of the algorithm can improve the recognition rate by about 37% on the average over all texts. A large percentage of the errors made by the algorithm were due to short content words competing with function words for the same position in the utterance. Because function words do not contribute to the semantic associations there is an increased probability that erroneous content words will be (mis)recognized instead. Many of these errors could be eliminated with the addition of syntactic or bigram constraints used in conjunction with the semantic model in order to process combinations of word classes such as <function word> <content word>, <content word> <function word>, <function word> <function word>, etc.

5. Conclusions

The vast quantities of textual information and the need for large vocabulary habitable systems require the investigation into the utility of linguistic knowledge from available resources. The research described in this paper provides evidence that the lexical semantic knowledge from MRDs can be modelled and used efficiently for large vocabulary NLP tasks related to recognition applications.

In practice, it would be impossible to model by hand or even by corpus training all semantic dependencies between all words language in all probable contexts. The dictionary knowledge, however weak or incomplete it may be, can provide semantic constraints about most words in the language in a way that is economical (in terms of computer processing) and easy to use.

In the area of language modelling for "noisy-channel" applications, such as speech or text recognition, this work distinguishes itself from other more established approaches, such as n-grams, in that it uses information which can be fully acquired from reusable language resources (MRDs) without the need for hand-coding or training procedures. N-gram models are usually dependent on the particular domain of the training texts whereas the dictionary model is generally domain-independent having shown signs of robustness across various genres. In fact, the findings suggest that the semantic model derived from the dictionary can be used in a way to complement rather than compete with other methods of language modelling. This is because, in contrast to n-grams that can provide local constraints for the words, the dictionary model can capture word dependencies at longer text distances.

As an extension to this work, we are currently experimenting with strategies that make use of these two different kinds of constraints, probabilistic from text corpora and semantic from MRDs within a unified mixture model using an Expectation-Maximization strategy. Further work will also concentrate on providing semantic constraints from dictionaries within a cache-based framework (Kuhn and De Mori, 1990; Lau et al, 1993).

6. References

Alshawi, H. 1987. Processing dictionary definitions with phrasal pattern hierarchies. In *Special Issue of Computational Linguistics on the Lexicon*, 13(3-4).

Bruce, R. and Wiebe, J. 1994. Word-sense Disambiguation Using Decomposable Models. In *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics (ACL-94).*

Calzolari, N. and Picchi, E. 1988. Acquisition of semantic information from an on-line dictionary. *In Proceedings of COLING-88*, pp. 87-91.

Cowie J., Wakao, T., Guthrie, L. and Jin, W. 1993. The Diderot Information Extraction System. *In Proceedings of the Pacific Association for Computational Linguistics Conference (PACLING-93).*

Demetriou, G., Atwell, E. and Souter, C. 1997. Large-scale Lexical Semantics for Speech Recognition Support. *In Proceedings of the 5th European Conference on Speech Communication and Technology EUROSPEECH'97).*.

Guthrie, J., Guthrie, L, Wilks, Y. and Aidinejad, H. 1991. Subject Dependent Co-occurrence and Word Sense Disambiguation. In *Proceedings of the 6th European Conference of the Association for Computational Linguistics.*

Ide, N. and Veronis, J. (1994), Have we wasted our time? In *Cambridge Language Reference News*, 4.

Jaccard, P. 1908. Nouvelles recherches sur la distribution florale. In *Bulletin de la Société de Vaud des Sciences Naturelles*, 44.

Jelinek, F. 1990. Self-organised language modeling for speech recognition. In A. Waibel and Kai-Fu Lee, (eds) *Readings in Speech Recognition*, Morgan Kaufmann.

Kozima, H. and Furugori, T. 1993. Similarity between Words Computed by Spreading Activation on an English Dictionary. In *Proceedings of the 32nd Annual Meeting of the Association of Computational Linguistics (EACL'93).*

Krovetz, R. and Croft, B. 1992. Lexical Ambiguity and Information Retrieval. In *ACM Transactions on Information Systems*, 102.

Kuhn, R. and De Mori, R. 1990. A Cache-Based Natural Language Model for Speech Recognition. In *IEEE Trans. PAMI* 12(6), pp. 570-583.

Kwong, Oi-Yee 1998. Bridging the Gap between Dictionary and Thesaurus. In *Proceedings of the COLING-ACL'98 Joint Conference of the 17th International Conference on Computational Linguistics and 36th Annual Meeting of the Association for Computational Linguistics.*

Lau, R., Rosenfeld, R. and Roukos, S. 1993. Adaptive Language Modeling Using the Maximum Entropy Principle. In *Proceedings of the ARPA-93 Human Language Technology Workshop*, pp.108-113

Lesk, M. 1986. Automatic sense disambiguation using MRDs: how to tell a pine cone from an ice cream cone. In *Proceedings of the 1986 SIGDOC*, pp. 24-26.

Nilsson, N. 1971. *Problem-Solving Methods in Artificial Intelligence*, McGraw-Hill.

Procter, P. 1978. *The Longman Dictionary of Contemporary English*, Longman.

Rigau, G., Atserias, J. and Agirre, E. 1997. Combining Unsupervised Lexical Knowledge Methods for Word Sense Disambiguation. *In Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics (ACL) and the 8th Conference of the European Chapter of the Association for Computational Linguistics (EACL).*

Veronis, J. and Ide, N. 1990. Word sense disambiguation with very large neural networks extracted from MRDs. In *Proceedings of the 13th International Conference on Computational Linguistics*, vol. 2.

Walker, D. and Amsler, R. 1986. The Use of Machine Readable Dictionaries in Sublanguage Analysis. In R. Grishman & R. Kittredge (eds*), Analyzing Language in Restricted Domains*, LEA.

Wilks,Y., Fass, D., Guo, C. M., McDonald, J., Plate, T. and Slator, B. 1989. A tractable machine dictionary as a resource for computational semantics. In , B. Boguraev and T. Briscoe (eds), *Computational Lexicography for Natural Language Processing*, Longman.