Combining Disambiguation Techniques to Enrich an Ontology

Mark Stevenson[1]

Abstract. This paper presents a method for adding domain information to ontologies using information which is readily available in thesauruses. The approach uses a combination of three existing disambiguation methods from the NLP literature. The approach is implemented on a portion of the WordNet hierarchy with thesaural categories derived from the Cambridge International Dictionary of English. Evaluation is carried out by comparing the system output against a gold-standard derived from SEMCOR, an existing disambiguated corpus, and found to perform well when compared with human judgments.

1 The Tennis Problem

WordNet [5] is a broad coverage thesaurus which is now widely used in Natural Language Processing and Information Retrieval. It consists of sets of lexical items (words and phrases) with similar meanings called synsets which are organised into a hyponomy (IS_A) hierarchy. For example, "doctor" and "physician" are in the same synset which is directly subsumed by "medical practitioner". It has been suggested ([12][9][20]) that WordNet could be a valuable resource for Information Retrieval since the words in the synsets could be used to expand the query thereby allowing the retrieval of documents which mention the same concept but use different words to do so. For example, a query about probability may include the term "stochastic" but not "probabilistic". However it is likely that retrieving documents which include either of these terms would be beneficial for the search.

WordNet has many advantages for such applications; it is a large resource with wide lexical coverage, it is freely available and has been widely studied allowing researchers to easily build upon each others' work. However, the main drawback is that the resource consists mainly of IS_A links and does not include other lexical relations such as discourse information. An example of this often given is that of "player" (a hyponym of person) which is not closely related to "racket", "ball" or "net" (hyponyms of artifact). Motivated by this example, Fellbaum [5] dubbed this the "tennis problem". This problem can be clearly seen in Figure 1 where the synset containing the phrase "tennis ball" is related to the one containing "net" by a relatively small number of steps in the IS_A hierarchy. However, the hierarchical distance between "tennis ball" and "ball boy" is large since the shortest path goes through the root node of the hierarchy. In fact, the WordNet hierarchy

does not indicate that there is any more relation between "tennis ball" and "ball boy" than, for example, "tennis ball" and "orange" or "tennis ball" and "coffee".
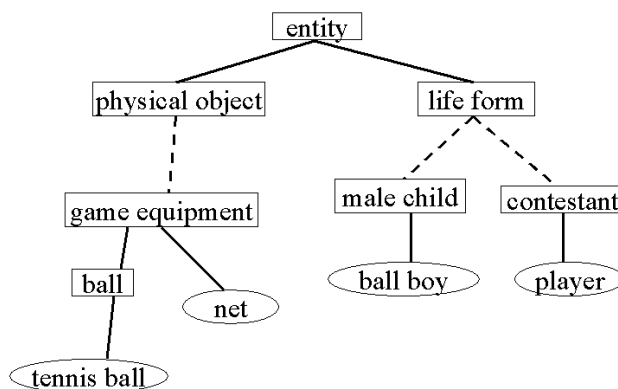

Figure 1. Fragment of existing WordNet hierarchy showing lack of connection between topically related concepts.

WordNet would be a richer resource if information about the concepts likely to occur together in a text on a particular domain were included. Such a resource could be used to identify documents on the internet by concept rather than simple keywords, a central goal of the Semantic Web project [3]. This paper reports a step towards this by automatically adding thesaural relations to the noun taxonomy in WordNet. The aim is to produce groups of noun synsets which are related by topic or domain. Once identified these links can be added to WordNet to denote this new form of lexical information, which is in addition to the existing hyponomy and hypernymy relations. This is shown in Figure 2.

In Section 2 we describe the thesaural classes on which the new links are based. A combination of three disambiguation algorithms is used to identify the relevant senses in WordNet and this process is also described in Section 2. An evaluation of this approach is presented in Section 3. Some related work is described in Section 4. The paper is concluded in Section 5 with suggestions for ways in which this work could be extended and applied.

2 Disambiguating Thesaural Classes

The thesaural links which are added to WordNet are generated from the related words classes found in the CIDE+ ma-

---
[1] Reuters Ltd, 85 Fleet Street, London, UK EC4P 4AJ mark.stevenson@reuters.com

chine readable dictionary [17]. The main problem with using these existing classes is that it is not clear which of the possible synsets in WordNet to which each word refers. Therefore the problem reduces to disambiguating sets of related nouns against the senses in the taxonomy. This problem is addressed by combining existing disambiguation techniques and applying this approach to WordNet and CIDE+.
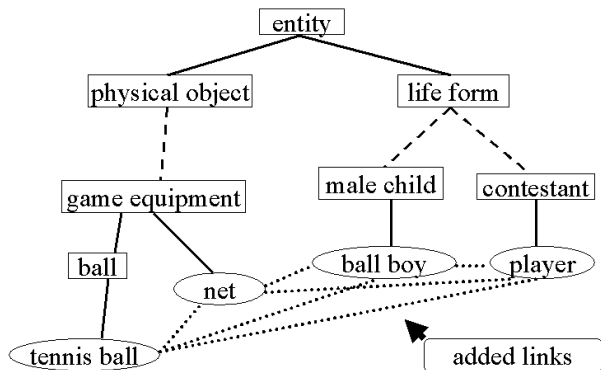

Figure 2. Proposed augmented hierarchy. Newly added links are shown as dotted lines.

The machine readable version of the CIDE+ dictionary contains thesaural information in the form of 1924 related word classes. Example categories are *Board Games* (*dice_1_0, square_1_3, shake_1_12, backgammon_1_0 ...*) and *Colours* (*silver_1_1, redness_1_0, black_1_22, amber_1_0 ...*). Each member of the category relates to a particular CIDE+ sense which allows access to the textual definitions contained in the dictionary. This extra information is the main advantage of using CIDE+ related word categories compared to a traditional thesaurus which consists only of groups of related words untagged for sense.

The first stage is to disambiguate each noun in CIDE+ thesaurus classes against WordNet. Formally, assume that CIDE+ thesaurus class, N, contains the word W with senses $\{s_1, s_2...s_n\}$. We wish to assign a value to each possible WordNet sense for W indicating the likelihood of that sense being relevant given that W is a member of the noun group N. This is achieved using a combination of existing lexical similarity metrics which are now described.

2.1 Information and semantic similarity

Resnik [18] reports an algorithm for disambiguating groups of nouns against their WordNet synsets. This is achieved by assigning each node in WordNet a numerical value which indicates its information content. This value is derived from frequency of occurrence in a training corpus by assuming that each parent of a term is implied by a corpus instance. The actualy information content value is calculated, in the standard information theoretic way, as the negative logarithm of the probability. When these values have been calculated each node in the taxonomy is more informative than its parents, for example, "nurse" and "professor" are each more informative than "professional". Disambiguation is carried out by compar-

ing senses in a pairwise fashion and selecting the combination which yields the highest information value.

Resnik [18] trained his algorithm on the 1 million word Brown Corpus of American English [6]. Our reimplementation used the written portion of the British National Corpus [4] which contains roughly 90 million words. However this additional training data did not seem to make much difference to the algorithm's performance.

More formally, Resnik's algorithm returns a value, $\phi(s_i)$, for each each sense, s_i, associated with a word W which is a member of the CIDE+ noun group N indicating the likelihood that the sense is relevant to the noun group. We normalise these values for each sense thus to create the Information Content (IC) metric:

$$IC_i = \frac{\phi(s_i)}{\sum_{j=1}^{n} \phi(s_j)} \qquad (1)$$

2.2 Conceptual Distance

In Resnik's disambiguation method distance between senses in the taxonomy is determined by their information content in the WordNet hierarchy. Agirre and Rigau [2] presented a method for disambiguating nouns in text where the distance between WordNet senses is determined solely by the structure of the taxonomy, i.e. the corpus frequencies used by Resnik were not considered. This method is known as Conceptual Distance. It prefers senses closely related in the WordNet hierarchy taking into account the depth of the hierarchy, density of senses and length of shortest path between concepts. Like Resnik's method, this approach takes a set of nouns as input and calculated the Conceptual Distance value for each possible WordNet sense relative to the words in the set.

Assuming that the value returned by the conceptual distance algorithm for sense s_i is $\theta(s_i)$ then the value given by the conceptual distance (CD) metric for a sense is given by:

$$CD_i = \frac{\theta(s_i)}{\sum_{j=1}^{n} \theta(s_j)} \qquad (2)$$

The software which implements this disambiguation algorithm has been made publically available[2] and was used for the experiments described in this paper.

2.3 Word Overlap Metric

The Information Content and Conceptual Distance metrics disambiguate via the structure of WordNet's taxonomy and corpus information. Lesk [14] suggested an alternative approach in which similarity is defined as the number of content words shared by the textual definitions of senses.

This method is used to identify the WordNet synset which is most similar to a CIDE+ sense by comparing the textual definition of that sense against the textual definition of all potential WordNet synsets. We call this metric the Dictionary Overlap (DO) measure and, like the previous pair, the value it returns is normalised against the total sum of word matches across all senses. So if we define α_i as the number of overlapping words in CIDE+ definition and gloss of the sense s_i then DO is calculated in our implementation thus:

---

[2] http://www.ji.si.ehu.es/users/eneko/00-release-v01.tar

$$DO_i = \frac{\alpha(s_i)}{\sum_{j=1}^{n} \alpha(s_j)} \tag{3}$$

This metric can be implemented more easily than the previous pair. Our implementation preprocesses the definitions in WordNet and CIDE+ by removing stop words and empty heads.[3] The remaining words are then reduced to their morphological roots.

2.4 Combining Metrics

So far we have described three metrics which could be used to disambiguate the sense in CIDE+ noun groups against Word-Net synsets. An obvious next step would be to combine these metrics. This is done by computing the sum and product of various combinations shown in Table 1. The combinations of metrics are computed from the value of the relevant combination for a particular sense normalised by the total value for that combination across all possible senses. Thus, for example, the value of the $IC + CD$ metric combination for sense i is given by:

$$(IC + CD)_i = \frac{\phi(s_i) + \theta(s_i)}{\sum_{j=1}^{n} (\phi(s_j) + \theta(s_j))} \tag{4}$$

Single metric	**1** DO
	2 IC
	3 CD
Sum	**4** $DO + IC$
	5 $DO + CD$
	6 $IC + CD$
	7 $DO + IC + CD$
Product	**8** $DO * IC$
	9 $DO * CD$
	10 $IC * CD$
	11 $DO * IC * CD$

Table 1. Combination of metrics

3 Evaluation

To evaluate the disambiguation method we require some resource which lists the appropriate WordNet sense for each of the members of the CIDE+ categories. To our knowledge no such resource exists and we are forced to adapt an existing resource. The most widely used and reliable text tagged with WordNet senses is SEMCOR [13], a 200,000 word portion of the Brown corpus semantically tagged as part of the WordNet project.[4] This corpus consists of 103 files on a wide variety of topics, each concerned with a particular subject. Content words are tagged with WordNet synset numbers but there is

no direct way of telling which of these are related to CIDE+ senses. However, when semantically related nouns occur together they do so with the related meanings. For example, if "monitor", "drive", "zip" and "screen" occur in the same text it is highly likely that they are all used with their senses related to computer hardware. In addition nouns appearing in a text on a given topic are likely to be related to that topic. So the word "Java" is more likely to mean 'programming language' than 'coffee' or 'island' in a text about software development. These two observations about the behavior of noun meanings can be used to automatically derive disambiguated noun groups suitable for our evaluation from the SEMCOR texts. It is likely that in SEMCOR texts which contain a large portion of the types from a particular thesaurus class those words will be used with the WordNet sense appropriate to that class.

To match SEMCOR texts against CIDE+ categories we compared each text against each category and attached a score to their relation based on the percentage of words in the category which appeared at least once in the document. This allowed us to produce a ranked list of document-category pairs based on this simple measure of relatedness. Any documents in which a CIDE+ category's words were used with less than 5 tokens were discarded. We then extracted the 10 highest ranking pairs, ignoring any categories which have already appeared. In effect this is a naive information retrieval (IR) system in which SEMCOR is the document collection and the words in the CIDE+ related word categories are the queries. We did not use a full IR system as the results from this simple method appeared adequate for the evaluation.

Table 2 shows the ten CIDE+ categories used in our evaluation and the document associated with each. The first column shows the CIDE+ category number and description, the second the percentage of types in that category followed by the number of types and tokens from the category which appear in the SEMCOR document listed in the next column. The final column contains a short description of the topic for the SEMCOR file. A first observation is that there is a reasonable semantic relation between each of the CIDE+ categories and the texts to which they were mapped.

A distinct advantage of this approach is that it allows more than one sense in WordNet to be associated with a CIDE+ sense. This is necessary since the lexicographers for each resource may have made different decisions about how rough or fine grained the sense distinctions should be. Table 3 lists the words used with more than one sense in each of the remaining four SEMCOR files. A first observation is that all senses appear consistent with the CIDE+ category related to that file (listed in Table 2). The two senses of "surface" are extremely similar while "football" and "church" exhibit clear regular polysemy. This analysis is consistent with the, at first sight conflicting, results reported by both Gale et. al. [7] and Krovetz [11]. The first claimed that most words are used with the same broad meaning (or homograph) in a given discourse while Krovetz claimed that closely related senses are often observed in the same discourse.

3.1 Evaluation Metric

Some measure is required to compare the system output with the senses found in the SEMCOR files. Our system does not

---

[3] Empty heads are short phrases which occur at the start of dictionary definitions and are part of the information indicating the sense's genus term. For example "any of various" is an empty head in the WordNet definition of "salmon": "any of various large food and game fishes ...".

[4] The semantic tags in SEMCOR refer to WordNet version 1.5 although the release now contains a mapping for nouns between WordNet versions 1.5 and 1.6 which was used to adapt the corpus to use the lexicon in these experiments.

CIDE+ Category	Overlap	Types/ tokens	File	Description of text topic
268 *Names of months*	0.77	8/11	`br-j56`	history of utilities in US town (mentions months several times)
325 *Planets*	0.73	6/44	`br-j01`	astronomy
1528 *Bays and gulfs*	0.67	4/7	`br-k16`	portion of novel (frequent mentions of geographical features)
1253 *Atoms, molecules and sub-atomic particles*	0.67	3/17	`br-j04`	sub-atomic chemistry
373 *Secondary education*	0.54	2/7	`br-a02`	US senate debates on education
486 *Poultry*	0.53	6/25	`br-k27`	portion of novel (describes character's hens)
147 *Extrasensory perception, Telepathy, psychics*	0.5	3/6	`br-f03`	text on psychoanalysis
1252 *Energy, force and power*	0.43	4/16	`br-j07`	text on engineering/mechanics
22 *American football*	0.41	6/18	`br-a12`	text on American football
748 *Churches, buildings and organizations*	0.41	7/24	`br-d03`	text on history of English church

Table 2. Mapping between CIDE+ categories and SEMCOR documents

File	Word	Senses
`br-f03`	mind	head/brain (3), recall (1)
`br-j07`	surface	outer boundary (1), extended 2D boundary (1)
`br-a12`	football	game (4), object (1)
`br-d03`	church	building (13), organisation (4), service (1)

Table 3. Senses from CIDE+ categories with more than sense in the identified SEMCOR file with number of occurrences in brackets

return a single sense, instead it assigns a score to each which can be viewed as a probability distribution describing the likelihood of each synset belonging to the class in question. We also know that some of the SEMCOR files contain words which are used in more than one sense and consequently their sense taggings can also be viewed as a probability distribution. Resnik and Yarowsky [19] proposed the cross entropy metric for comparing a probability distribution produced by a disambiguation technique with disambiguated text. It is calculated according to the following formula:

$$CE(t(x), s(x)) = - \sum_{x \in S} t(x) \log_2 s(x) \qquad (5)$$

where S is the set of senses for the word in question, $t(x)$ is the probability distribution for those senses observed in the SEMCOR files and $s(x)$ the probability distribution obtained from our system. A perfect match between the distributions returns their entropy (which will, of course, be equal). Higher values indicate lower agreement between the distributions.

To compare the output of our system for all words in a thesaurus class against the distribution found in a SEMCOR document we devised the Average Cross Entropy (ACE) metric:

$$\text{ACE} = \frac{\sum_{w \in W} CE(t(w), s(w))}{|W|} \qquad (6)$$

where W is the set of word types occurring in the document and $t(w)$ and $s(w)$ are the relevant probability density functions. This compares the distribution of each word appearing in the text with that output by the system and adds extra weight for words which appear frequently in the SEMCOR text.

The theoretical minimum for this measure, calculated by assuming a perfect match between the two distributions was found to be 0.07 across all 10 SEMCOR texts. The average entropy for the texts other than those mentioned in Table 3 was 0 since, in these cases, the distribution was such that each type has a single sense with probability of 1 and all others zero probability.

3.2 Results

In order to compare the various implemented methods with a naive approach a baseline was implemented. This randomly chose a sense from the set of possibilities for each word and assigned a probability of 1 to it and 0 to all other senses. The baseline was run 10 times and it was found that the mean of the average cross entropy scores over these 10 runs was 15.96 with a standard deviation of 1.83. The CIDE+ noun groups were also tagged manually with the annotator being asked to choose a single WordNet sense for each word in a group by considering the entire set of nouns it contains as evidence. Like the automatic baseline calculation this was a forced choice task in which the annotator was asked to choose exactly one WordNet sense for each CIDE+ sense.

Table 4 shows the evaluation results for various metrics. It can be seen that all metrics perform better (lower ACE) than the random baseline. The best performance is observed from the product of all three metrics where the ACE obtained

(5.76) is close to that obtained from the human annotator, representing a 94% reduction in error rate. This result is consistent with earlier work such as Stevenson and Wilks [21] and McRoy [16] which showed that word sense disambiguation is a task which benefits from a combination of multiple classifiers. Although this result should be considered in the context of the fact that a perfect match with the test data would return a score of 0.07 and the manual annotation task is a forced choice of a single sense while the various metrics assign probabilities to senses.

There is a noticeable difference between the performance of the CD metric compared with the other two. This difference is not statistically significant according to a two-tail paired t-test although this may be due to the small amount of sample data. The pairwise correlations of the performance of each method on each of the 10 SEMCOR files was also quite high (between 0.79 and 0.89) indicating that some of the SEMCOR files were more difficult to disambiguate than others.

It is interesting to note that there does not appear to be much difference between the additive combinations of metrics (**4 - 7**) and the single metrics (**1 - 3**). However, the products of metrics (**8 - 11**) perform noticeably better. This may be because the multiplication combination is more conservative since all metrics must agree that there is some evidence for a particular sense. If any metrics assigns a zero probability to a sense then the product will be zero.

4   Related Work

Mandala et. al. [15] combined three thesauruses to expand queries for an IR system. It was found that the combination of all three produced better results than no query expansion or when a single resource was used. The nature of their application meant that there was no need to produce an explicit mapping between the senses of the three resources.

Agirre et. al. [1] constructed topic signatures constructed from web searches to add extra information to WordNet. A test set of 20 nouns which occur at least 100 times in SEM-COR was chosen. For each possible WordNet sense a query was sent to the AltaVista search engine[5] and the results stored. These documents were used to construct a topic signature for each concept which were evaluated within a sense disambiguation algorithm and found to outperform information extracted directly from WordNet. They produced further improvements when used to cluster senses.

Knight and Luk [10] provided a mapping between WordNet and LDOCE by combining textual definitions with information about the hierarchical structure of the resources reporting 96% mapping accuracy. Green et. al. [8] use a combination of similarity metrics, including Resnik's, to map entries in a verb database onto WordNet senses, reporting 72% precision and 58% recall. This suggests that the approach described here may be useful for other grammatical categories.

5   Conclusion and Future Work

We have presented a method for overcoming the "tennis problem" in taxonomies such as WordNet by adding new relations to the hierarchy obtained by disambiguating the noun

---

[5] http://www.altavista.com

B Baseline	15.96	**6** $IC + CD$	10.30
H Human	5.16	**7** $DO + IC + CD$	10.98
1 DO	11.93	**8** $DO * IC$	7.93
2 IC	11.28	**9** $DO * CD$	5.80
3 CD	9.31	**10** $IC * CD$	9.16
4 $DO + IC$	11.86	**11** $DO * IC * CD$	5.76
5 $DO + CD$	10.78		

Table 4. Results from various combinations of metrics

groups found in existing thesaural classes. It was found that this can be achieved using a combination of existing disambiguation techniques. The techniques were evaluated using gold standard taggings derived automatically from SEMCOR and found to perform well when compared with manual annotation.

Future work is planned on improving the quality of the mapping further by making use of machine learning algorithms. The method for combining the results of the various disambiguation metrics reported here is extremely naive - a simple product of normalised scores. It is possible that the results could be improved upon by making use of some learning algorithm which might optimise this result further. It is likely that machine learning methods may be helpful since there is still a reasonable difference between the result achieved by the best scoring combination of methods and the theoretical minimum which would be achieved if the data were modelled precisely.

Another source of future work would be to experiment with the input thesauruses to this process. It is possible that combining the input of several thesauruses will lead to further enrichment of WordNet and since several are readily available on the internet this should be straightforward to carry out. It would also be of interest to follow the approach of Agirre et. al. [1] in using queries on internet search engines to generate sets of documents related to a topic which could then be analysed to generate further thesaural categories.

A further source of future research will be to carry out task-based experiments by making use of the enriched lexicon for information retrieval. This evaluation would invert the previous suggestion in which the internet was used as a resource through which WordNet could be enriched. In this case queries on internet search engines would be used to benchmark the resource through analysis of retrieved documents. If these stages proved to be successful then it may be possible to iteratively improve the resource using a cycle of internet searches, each of which improves upon the accuracy of the last by adapting WordNet.

ACKNOWLEDGEMENTS

REFERENCES

[1] E. Agirre, O. Ansa, E. Hovy, and D. Martinez, 'Enriching very large ontologies using the WWW', in *Proceedings of the ECAI Ontology Learning Workshop*, pp. 73–77, Berlin, Germany, (2000).

[2] E. Agirre and G. Rigau, 'Word sense disambiguation using conceptual density', in *Proceedings of COLING '96*, pp. 16–22, Copenhagen, Denmark, (1996).

[3] T. Berners-Lee, J. Hendler, and O. Lassila, 'The semantic web', *Scientific American*, **28**(5), 34–43, (2001).

[4] L. Burnard, *Users Reference Guide for the British National Corpus*, Oxford University Computing Services, 1995.

[5] C. Fellbaum, *WordNet: An Electronic Lexical Database and some of its Applications*, MIT Press, Cambridge, MA, 1998.

[6] W. Francis and H. Kučera, *Frequency Analysis of English Usage*, Hougton Mufflin Co., New York, 1982.

[7] W. Gale, K. Church, and D. Yarowsky, 'One sense per discourse', in *Proceedings of the DARPA Speech and Natural Language Workshop*, pp. 233–237, Harriman, NY, (1992).

[8] R. Green, L. Pearl, B. Dorr, and P. Resnik, 'Mapping Lexical Entries in a Verb Database to WordNet Senses', in *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*, pp. 244–251, Toulouse, France, (2001).

[9] H. Jing and E. Tzoukermann, 'Information retrieval based on context distance and morphology', in *Proceedings of the 22nd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '99)*, pp. 90–96, University of Berkley, CA, (1999).

[10] K. Knight and S. Luk, 'Building a large knowledge base for machine translation', in *Proceedings of the American Association for Artificial Intelligence Conference (AAAI-94)*, pp. 185–109, Seattle, WA., (1994).

[11] R. Krovetz, 'More than one sense per discourse', in *Proceedings of SENSEVAL Workshop*, Herstmonceux Castle, UK, (1998).

[12] R. Krovetz and B. Croft, 'Lexical ambiguity and information retrieval', *ACM Transactions on Information Systems*, **10**(2), 115–141, (1992).

[13] S. Landes, C. Leacock, and R. Tengi, 'Building a semantic concordance of English', in *WordNet: An electronic lexical database and some applications*, ed., C. Fellbaum, 199–217, MIT Press, Cambridge, MA., (1998).

[14] M. Lesk, 'Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone', in *Proceedings of ACM SIGDOC Conference*, pp. 24–26, Toronto, Canada, (1986).

[15] R. Mandala, T. Tokunaga, and H. Tanaka, 'Combining general hand-made and automatically constructed thesauri for information retrieval', in *Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence (IJCAI-99)*, pp. 920–924, Stockholm, Sweden, (1999).

[16] S. McRoy, 'Using multiple knowledge sources for word sense

disambiguation', *Computational Linguistics*, **18**(1), 1–30, (1992).

[17] *Cambridge International Dictionary of English*, ed., P. Procter, Cambridge University Press, Cambridge, 1995.

[18] P. Resnik, 'Disambiguating Noun Groupings with Respect to WordNet senses', in *Natural Language Processing using Very Large Corpora*, eds., S. Armstrong, K. Church, P. Isabelle, S. Manzi, E. Tzoukermann, and D. Yarowsky, 77–98, Kluwer Academic Press, (1999).

[19] P. Resnik and D. Yarowsky, 'A perspective on word sense disambiguation techniques and their evaluation', in *Proceedings of the SIGLEX Workshop "Tagging Text with Lexical Semantics: What, why and how?"*, pp. 79–86, Washington, D.C., (1997).

[20] H. Schütze and J. Pedersen, 'Information retrieval based on word senses', in *Symposium on Document Analysis and Information Retrieval (SDAIR)*, pp. 161–175, Las Vegas, NV, (1995).

[21] M. Stevenson and Y. Wilks, 'Combining weak knowledge sources for sense disambiguation', in *Proceedings of the Sixteenth International Conference on Artificial Intelligence (IJCAI-99)*, pp. 135–141, Stockholm, Sweden, (1999).