An Information Retrieval Approach to Sense Ranking

Mirella Lapata and **Frank Keller**
School of Informatics, University of Edinburgh
2 Buccleuch Place, Edinburgh EH8 9LW, UK
{mlap,keller}@inf.ed.ac.uk

Abstract

In word sense disambiguation, choosing the most frequent sense for an ambiguous word is a powerful heuristic. However, its usefulness is restricted by the availability of sense-annotated data. In this paper, we propose an information retrieval-based method for sense ranking that does not require annotated data. The method queries an information retrieval engine to estimate the degree of association between a word and its sense descriptions. Experiments on the Senseval test materials yield state-of-the-art performance. We also show that the estimated sense frequencies correlate reliably with native speakers' intuitions.

1 Introduction

Word sense disambiguation (WSD), the ability to identify the intended meanings (senses) of words in context, is crucial for accomplishing many NLP tasks that require semantic processing. Examples include paraphrase acquisition, discourse parsing, or metonymy resolution. Applications such as machine translation (Vickrey et al., 2005) and information retrieval (Stokoe, 2005) have also been shown to benefit from WSD.

Given the importance of WSD for basic NLP tasks and multilingual applications, much work has focused on the computational treatment of sense ambiguity, primarily using data-driven methods. Most accurate WSD systems to date are supervised and rely on the availability of training data (see Yarowsky and Florian 2002; Mihalcea and Edmonds 2004 and the references therein). Although supervised methods typically achieve better performance than unsupervised alternatives, their applicability is limited to those words for which sense labeled data exists, and their accuracy is strongly correlated with the amount of labeled data available. Furthermore, current supervised approaches rarely outperform the simple heuristic of choosing the most common or dominant sense in the training data (henceforth "the first sense heuristic"), despite taking local context into account. One reason for this is the highly skewed distribution of word senses (McCarthy et al., 2004a). A large number of frequent content words is often associated with only one dominant sense.

Obtaining the first sense via annotation is obviously costly and time consuming. Sense annotated corpora are not readily available for different languages or indeed sense inventories. Moreover, a word's dominant sense will vary across domains and text genres (the word *court* in legal documents will most likely mean *tribunal* rather than *yard*). It is therefore not surprising that recent work (McCarthy et al., 2004a; Mohammad and Hirst, 2006; Brody et al., 2006) attempts to alleviate the annotation bottleneck by inferring the first sense automatically from raw text. Automatically acquired first senses will undoubtedly be noisy when compared to human annotations. Nevertheless, they can be usefully employed in two important tasks: (a) to create preliminary annotations, thus supporting the "annotate automatically, correct manually" methodology used to provide high volume annotation in the Penn Treebank project; and (b) in combination with supervised WSD methods that take context into account; for instance, such methods could default to the dominant sense for unseen words or words with uninformative contexts.

This paper focuses on a knowledge-lean sense ranking method that exploits a sense inventory like WordNet and corpus data to automatically induce dominant senses. The proposed method infers the associations between words and sense descriptions automatically by querying an IR engine whose index terms have been compiled from the corpus of interest. The approach is inexpensive, language-independent, requires minimal supervision, and uses no additional knowledge other than the word senses proper and morphological query expansions. We

Proceedings of NAACL HLT 2007, pages 348–355,
Rochester, NY, April 2007.

evaluate our method on two tasks. First, we use the acquired dominant senses to disambiguate the meanings of words in the Senseval-2 (Palmer et al., 2001) and Senseval-3 (Snyder and Palmer, 2004) data sets. Second, we simulate native speakers' intuitions about the salience of word meanings and examine whether the estimated sense frequencies correlate with sense production data. In all cases our approach outperforms a naive baseline and yields performances comparable to state of the art.

In the following section, we provide an overview of existing work on sense ranking. In Section 3, we introduce our IR-based method, and describe several sense ranking models. In Section 4, we present our results. Discussion of our results and future work conclude the paper (Section 5).

2 Related Work

McCarthy et al. (2004a) were the first to propose a computational model for acquiring dominant senses from text corpora. Key in their approach is the observation that distributionally similar neighbors often provide cues about a word's senses. The model quantifies the degree of similarity between a word's sense descriptions and its closest neighbors, thus delivering a ranking over senses where the most similar sense is intuitively the dominant sense. Their method exploits two notions of similarity, distributional and semantic. Distributionally similar words are acquired from the British National Corpus using an information-theoretic similarity measure (Lin, 1998) operating over dependency relations (e.g., verb-subject, verb-object). The latter are obtained from the output of Briscoe and Carroll's (2002) parser. The semantic similarity between neighbors and senses is measured using a manually crafted taxonomy such as WordNet (see Budanitsky and Hirst 2001 for an overview of WordNet-based similarity measures).

Mohammad and Hirst (2006) propose an algorithm for inferring dominant senses without relying on distributionally similar neighbors. Their approach capitalizes on the collocational nature of semantically related words. Assuming a coarse-grained sense inventory (e.g., the Macquarie Thesaurus), it first creates a matrix whose columns represent *all* categories (senses) $c_1 \ldots c_n$ in the inventory and rows the ambiguous target words $w_1 \ldots w_m$; the matrix cells record the number of times a target word t_i co-occurs with category c_j within a window of size s. Using an appropriate statistical test, they estimate the relative strength of association between an ambiguous word and each of its senses. The sense with the highest association is the predominant sense.

Our work shares with McCarthy et al. (2004a) and Mohammad and Hirst (2006) the objective of inferring dominant senses automatically. We propose a knowledge-lean method that relies on word association and requires no syntactic annotation. The latter may be unavailable when working with languages other than English for which state-of-the-art parsers or taggers have not been developed. Mohammad and Hirst (2006) estimate the co-occurrence frequency of a word and its sense descriptors by considering small window sizes of up to five words. These estimates will be less reliable for moderately frequent words or for sense inventories with many senses. Our approach is more robust to sparse data – we work with document-based frequencies – and thus suitable for both coarse and fine grained sense inventories. Furthermore, it is computationally inexpensive; in contrast to McCarthy et al. (2004a) we do not rely on the structure of the sense inventory for measuring the similarity between synonyms and their senses. Moreover, unlike Mohammad and Hirst (2006), our algorithm only requires co-occurrence frequencies for the target word and its senses, without considering all senses in the inventory and all words in the corpus simultaneously.

3 Method

3.1 Motivation

Central in our approach is the assumption that context provides important cues regarding a word's meaning. The idea dates back at least to Firth (1957) ("You shall know a word by the company it keeps") and underlies most WSD work to date. Another observation that has found wide application in WSD is that words tend to exhibit only one sense in a given discourse or document (Gale et al., 1992). Furthermore, documents are typically written with certain topics in mind which are often indicated by word distributional patterns (Harris, 1982).

For example, documents talking about congressional tenure are likely to contain words such as *term of office* or *incumbency*, whereas documents talking about legal tenure (i.e., the right to hold property)

are likely to include the words *right* or *land*. Now, we could estimate which sense of tenure is most prevalent simply by comparing whether *tenure* co-occurs more often with *term of office* than with *land* provided we knew that both of these terms are semantically related to *tenure*. Fortunately, senses in WordNet (and related taxonomies) are represented by synonym terms. So, all we need to do for estimating a word's sense frequencies is to count how often it co-occurs with its synonyms. We adopt here a fairly broad definition of co-occurrence, two words co-occur if they are attested in the same document. We could obtain such counts from any document collection; however, to facilitate comparisons with prior work (e.g., McCarthy et al. 2004a), all our experiments use the British National Corpus (BNC). In what follows we describe in detail how we retrieve co-occurrence counts from the BNC and how we acquire dominant senses.

3.2 Dominant Sense Acquisition

Throughout the paper we use the term frequency as a shorthand for document frequency, i.e., the number of documents that contain a word or a set of words which may or may not be adjacent. The method we propose here exploits document frequencies of words and their sense definitions. We base our discussion below on the WordNet sense inventory and its representation of senses in terms of synonym sets (synsets). However, our approach is not limited to this particular lexicon; any dictionary with synonym-based sense definitions could serve our purposes.

As an example consider the noun *tenure*, which has the following senses in WordNet:

(1)
```
Sense 1
tenure, term of office, incumbency
=> term
Sense 2
tenure, land tenure
=> legal right
```

The senses are represented by the two synsets {`tenure, term of office, incumbency`} and {`tenure, land tenure`}. (The hypernyms for each sense are also listed; indicated by the arrows.) We can now approximate the frequency with which a word w_1 occurs with the sense s by computing its **synonym frequencies:** for each word $w_2 \in \text{syns}(s)$, the set of synonyms of s, we field a query of the form w_1 `AND` w_2. These synonym frequencies can then be used to determine the most frequent sense of w_1 in a variety of ways (to be detailed below).

The synsets for the two senses in (1) give rise to the queries in (2) and (3). Note that two queries are generated for the first synset, as it contains two synonyms of the target word *tenure*.

(2) a. `"tenure" AND "term of office"`
b. `"tenure" AND "incumbency"`

(3) `"tenure" AND "land tenure"`

For example, query (2-a) will return the number of documents in which *tenure* and *term of office* co-occur. Presumably, *tenure* is mainly used in its dominant sense in these documents. In the same way, query (3) will return documents in which *tenure* is used in the sense of *land tenure*. Note that this way of approximating synonym frequencies as document frequencies crucially relies on the "one sense per discourse" hypothesis (Gale et al., 1992), under the assumption that a document counts as a discourse for word sense disambiguation purposes.

Apart from synonym frequencies, we also generate **hypernym frequencies** by submitting queries of the form w_1 `AND` w_2, for each $w_2 \in \text{hype}(s)$, the set of immediate hypernyms of the sense s. The hypernym queries for the two senses of *tenure* are:

(4) `"tenure" AND "term"`

(5) `"tenure" AND "legal right"`

Hypernym queries are particularly useful for synsets of size one, i.e., where a word in a given sense has no synonyms, and is only differentiated from other senses by its hypernyms.

Before submitting queries such as the ones in (2) and (3) to an IR engine, we perform **query expansion** to make sure that all relevant inflected forms are included. For example the query term `"tenure"` is expanded to `("tenure" OR "tenures")`, i.e., both singular and plural noun forms are generated. Similarly, all inflected verb forms are generated, e.g., `"keep up"` gives rise to the query term `("keep up" OR "keeps up" OR "keeping up" OR "kept up")`. John Carroll's suite of morphological tools (`morpha` and `morphg`) is used to generate inflected forms for verbs and

nouns.[1]

The queries generated this way are then submitted to an IR engine to obtain document counts. Specifically, we indexed the BNC using GLIMPSE (Global Implicit Search) a fast and flexible indexing and query system[2] (Manber and Wu, 1994). GLIMPSE supports approximate and exact matching, Boolean queries, wild cards, regular expressions, and many other options. The text is divided into equal size blocks and an inverted index is created containing the words and the block numbers in which they occur. Given a query, GLIMPSE will retrieve the relevant documents using a two-level search method. It will first locate the query in the inverted index and then use sequential search to find an exact answer.

Once synonym frequencies and hypernym frequencies are in place, we can compute a word's predominant sense in a number of ways. First, we can vary the way the frequency of a given sense is estimated based on synonym frequencies:

- **Sum:** The frequency of a given synset is computed as the sum of the synonym frequencies. For example, the frequency of the dominant sense of *tenure* would be computed by adding up the document frequencies returned by queries (2-a) and (2-b).
- **Average (Avg):** The frequency of a synset is computed by taking the average of synonym frequencies.
- **Highest (High):** The frequency of a synset is determined by the synonym with the highest frequency.

Secondly, we can vary whether or not hypernyms are taken into account:

- **No hypernyms (−Hyp):** Only the synonym frequencies are included when computing the frequency of a synset. For example, only the queries of (2-a) and (2-b) are relevant for estimating the dominant sense of *tenure*.
- **Hypernyms (+Hyp):** Both synonym and hypernym frequencies are taken into account when computing sense frequency. For example, the frequency for the senses of *tenure* would be computed based on the document frequencies returned by queries (2-a), (2-b), and (4) (by summing, averaging, or taking the highest value, as before).

The third option relates to whether the sense frequencies are used in raw or in normalized form:

- **Non-normalized (−Norm)**: The raw synonym frequencies are used as estimates of sense frequencies.
- **Normalized (+Norm)**: Sense frequencies are computed by dividing the word-synonym frequency by the frequency of the synonym in isolation. For example, the normalized frequency for (2-a) is computed by dividing the document frequency for `"tenure" AND "term of office"` by the document frequency for `"term of office"`. Normalizing takes into account the fact that the members of the synset of a sense may differ in frequency.

The combination of the above parameters yields 12 sense ranking models. We explore the parameter space exhaustively on the Senseval-2 benchmark data set. The best performing model on this data set is then used in all our subsequent experiments. We use Senseval-2 as a development set, but we also demonstrate that a far smaller manually annotated sample is sufficient for selecting the best model.

4 Experiments

Our experiments were driven by three questions: (1) Is WSD feasible at all with a model that does not employ any syntactic or semantic knowledge? Recall that McCarthy et al. (2004a) propose a model that crucially relies on a robust parser for estimating dominant senses. (2) What is the best parameter setting for our model? (3) Do the acquired dominant senses correlate with human judgments? If our sense frequencies exhibit no such correlation, it is unlikely that they will be useful in practical applications.

To address the first two questions we use the induced first senses to perform WSD on the Senseval-2 and Senseval-3 data sets. For our third question we compare native speakers' semantic intuitions against the BNC sense frequencies.

[1]The tools can be downloaded from `http://www.informatics.susx.ac.uk/research/nlp/carroll/morph.html`.

[2]The software can be downloaded from `http://webglimpse.net/download.php`

	−Norm				+Norm			
	+Hyp		−Hyp		+Hyp		−Hyp	
	P	*R*	*P*	*R*	*P*	*R*	*P*	*R*
Sum	42.3	40.8	46.3	44.6	45.9	44.3	48.6	46.8
High	51.6	49.8	51.1	49.3	57.2	55.1	**59.7**	**57.6**
Avg	44.1	42.6	48.5	46.8	49.6	47.8	51.5	49.6

Table 1: Results for Senseval-2 data by model instantiation

	BaseR		BaseS		Model		
	P	*R*	*P*	*R*	*P*	*R*	*N*
Noun	26.8	25.4	45.8	43.4	$53.1^{*\#}$	$50.2^{*\#}$	1,063
Verb	11.2	11.1	19.9	19.5	$48.2^{*\#}$	$47.3^{*\#}$	569
Adj	22.1	21.4	56.5	56.0	56.7^{*}	56.2^{*}	451
Adv	48.0	45.9	66.4	62.9	$86.4^{*\#}$	$81.8^{*\#}$	301
All	26.3	25.4	42.2	40.7	$59.7^{*\#}$	$57.6^{*\#}$	2,384

Table 2: Results of best model (High, +Norm, −Hyp) for Senseval-2 data by part of speech (*: sig. diff. from BaseR, $^{\#}$: sig. diff. from BaseS; $p < 0.01$ using χ^2 test)

4.1 Model Selection

The goal of our first experiment is to establish which model configuration (see Section 3.2) is best suited for the WSD task. We thus varied how the overall frequency is computed (Sum, High, Avg), whether hyponyms are included (±Hyp), and whether the frequencies are normalized (±Norm). To explore the parameter space, we used the Senseval-2 all-words test data as our development set. This data set consists of three documents from the Wall Street Journal containing approximately 2,400 content words. Following McCarthy et al. (2004a), we first use our method to find the dominant sense for all word types in the corpus and then use that sense to disambiguate tokens without taking contextual information into account. We used WordNet 1.7.1 (Fellbaum, 1998) senses.[3]

We compared our results to a baseline that selects for each word type a random sense, assumes it is the dominant one, and uses it to disambiguate all instances of the target word (McCarthy et al., 2004a). We also report the WSD performance of a more competitive baseline that always chooses the sense with the largest synset as the dominant sense. Consider again the word *tenure* from Section 3.2. According to this baseline, the dominant sense for *tenure* is the first one since it is represented by the largest synset (three members).

Our results on Senseval-2 are summarized in Table 1. We observe that models that do not include hypernyms yield consistently better precision and recall than models that include them. On the one hand, hypernyms render the estimated sense distributions less sparse. On the other hand, they introduce considerable noise; the resulting sense frequencies are often similar – the same hypernyms can be shared among several senses – and selecting one predominant sense over the other can be due to very small frequency differences. We also find that models with normalized document counts outperform models without normalization. This is not surprising, there is ample evidence in the literature (Mohammad and Hirst, 2006; Turney, 2001) that association measures (e.g., conditional probability, mutual information) are better indicators of lexical similarity than raw frequency. Finally, selecting the synonym with the highest frequency (and defaulting to its sense) achieves better results in comparison to averaging or summing over all synsets.

In sum, the best performing model is High, +Norm, −Hyp, achieving a precision of 59.7% and a recall of 57.9%. The results for this model are broken down by part of speech in Table 2. Here, we also include a comparison with the random baseline (BaseR) and a baseline that selects the dominant sense by synset size (BaseS). We observe that the optimal model significantly outperforms both baselines on the complete data set (see row All in Table 2) and on most individual parts of speech (performances are comparable for our model and BaseS on adjectives). BaseS is far better than BaseR and generally harder to beat. Defaulting to synset size in the absence of any other information is a good heuristic; large synsets often describe frequent senses. Variants of our model that select a dominant sense by summing over synset members are closest to this baseline. Note that our best performing model does not rely on synset size; it simply selects the synonym with the highest frequency, despite the fact that it might belong to a large or small synset. We conjecture that its superior performance is due to the collocational nature of semantic similarity (Turney,

[3]Senseval-2 is annotated with WordNet 1.7 senses which we converted to 1.7.1 using a publicly available mapping (see `http://www.cs.unt.edu/~rada/downloads.html`).

	−Norm				+Norm			
	+Hyp		−Hyp		+Hyp		−Hyp	
	P	*R*	*P*	*R*	*P*	*R*	*P*	*R*
Sum	42.3	40.8	46.3	44.6	45.2	44.7	44.6	44.0
High	51.6	49.8	51.1	49.3	55.0	54.3	**61.3**	**60.5**
Avg	44.1	42.6	48.5	46.8	51.5	50.8	50.4	49.8

Table 3: Results for 10% of Senseval-2 data by model instantiation

2001).

In order to establish that High, +Norm, −Hyp is the optimal model, we utilized the whole Senseval-2 data set. Using such a large dataset is more likely to yield a stable parameter setting, but it also raises the question whether parameter optimization could take place on a smaller dataset which is less costly to produce. Table 3 explores the parameter space on a sample randomly drawn from Senseval-2 that contains only 240 tokens (i.e., one tenth of the original data set). The behavior of our models on this smaller sample is comparable to that on the entire Senseval-2 data. Importantly, both sets yield the same best model, i.e., High, +Norm, −Hyp. In the remainder of this paper we will use this model for further experiments without additional parameter tuning.

4.2 Application to Senseval-3 Data

We next evaluate our best model the on the Senseval-3 English all-words data set. Senseval-3 consists of two Wall Street Journal articles and one excerpt from the Brown corpus (approximately 5,000 content words in total). Similarly to the experiments reported in the previous section, we used WordNet 1.7.1. We calculate recall and precision with the Senseval-3 scorer.

Our results are given in Table 4. Besides the two baselines (BaseR and BaseS), we also compare our model to McCarthy et al. (2004b)[4] and the best unsupervised (IRST-DDD) and supervised (GAMBLE) systems that participated in Senseval-3. IRST-DDD was developed by Strapparava et al. (2004) and performs domain driven disambiguation. Specifically, the approach compares the domain of the context surrounding the target word with the domains of its senses and uses a version of WordNet augmented with domain labels (e.g., economy, geography). GAMBL (Decadt et al., 2004) is a supervised system: a classifier is trained for each ambiguous word using memory-based learning. We also report the performance achieved by defaulting to the first WordNet entry for a given word and part of speech. Entries in WordNet are ranked according to the sense frequency estimates obtained from the manually annotated SemCor corpus. First senses obtained from SemCor will be naturally less noisy than those computed by our method which does not make use of manual annotation in any way. We therefore consider the WSD performance achieved with SemCor first senses as an upper bound for automatically acquired first senses.

Our model significantly outperforms the two baselines and McCarthy et al. (2004b). Its precision and recall according to individual parts of speech is shown in Table 5. The model performs comparably to IRST-DDD and significantly worse than GAMBLE. This is not entirely surprising given that GAM-

[4]Comparison against Mohammad and Hirst (2006) was not possible since they use a sense inventory other than WordNet (i.e., Roget's thesaurus) and evaluate their model on artificially generated sense-tagged data.

	P	*R*
BaseR	$23.1^{\#\dagger\$\ddagger}$	$22.7^{\#\dagger\$\ddagger}$
BaseS	$36.6^{*\dagger\$\ddagger}$	$35.9^{*\dagger\$\ddagger}$
McCarthy	$49.0^{*\#\$\ddagger}$	$43.0^{*\#\$\ddagger}$
IR-Model	$58.0^{*\#\dagger\ddagger}$	$57.0^{*\#\dagger\ddagger}$
IRST-DDD	$58.3^{*\#\dagger\ddagger}$	$58.2^{*\#\dagger\ddagger}$
Semcor	$62.4^{*\#\dagger\$}$	$62.4^{*\#\dagger\$}$
GAMBLE	$65.1^{*\#\dagger\$\ddagger}$	$65.2^{*\#\dagger\$\ddagger}$

Table 4: Comparison of results on Senseval-3 data (*: sig. diff. from BaseR, $^\#$: sig. diff. from BaseS, †: sig. diff. from McCarthy, $^\$$: sig. diff. from IR-Model, ‡: sig. diff. from SemCor; $p < 0.01$ using χ^2 test)

	BaseR		BaseS		Model		
	P	*R*	*P*	*R*	*P*	*R*	*N*
Noun	27.8	12.2	41.1	41.0	$58.1^{*\#}$	$58.0^{*\#}$	900
Verb	12.8	4.6	20.0	19.9	$61.0^{*\#}$	$60.8^{*\#}$	732
Adj	29.2	5.2	56.5	56.5	50.3^{*}	50.3^{*}	363
Adv	100.0	0.6	100.0	81.2	100.0	81.2	16
All	23.1	22.7	36.6	35.9	$58.0^{*\#}$	$57.0^{*\#}$	2,011

Table 5: Results of best model (High, +Norm, −Hyp) for Senseval-3 data by part of speech (*: sig. diff. from BaseR, $^\#$: sig. diff. from BaseS; $p < 0.01$ using χ^2 test)

BLE is a supervised system trained on a variety of manually annotated resources including SemCor, data from previous Senseval workshops and the example sentences in WordNet 1.7.1. GAMBLE is the only system that significantly outperforms the SemCor upper bound. Finally, note that our model is conceptually simpler than McCarthy et al. (2004b) and IRST-DDD. It neither requires a parser (for obtaining distributionally similar neighbors) nor any knowledge other than WordNet (e.g., domain labels). This makes our method portable to languages for which syntactic analysis tools and elaborate semantic resources are not available.

4.3 Modeling Human Data

Research in psycholinguistics has shown that the meanings of ambiguous words are not perceived as equally salient in the absence of a biasing context (Durkin and Manning, 1989; Twilley et al., 1994). Rather, language users often ascribe dominant and subordinate meanings to polysemous words. Previous studies have elicited intuitions with regard to word senses using a free association task. For example, Durkin and Manning (1989) collected association norms from native speakers for 175 ambiguous words. They asked subjects to read each word and write down the first meaning that came to mind. The words were presented out of context. From the subjects' responses, they computed sense frequencies, which revealed that most words were attributed a particular meaning with a markedly higher frequency than other meanings.

In this experiment, we examine whether our model agrees with human intuitions regarding the prevalence of word senses. We inferred the dominant meanings for the polysemous words used in Durkin and Manning (1989). These exhibit a relatively high degree of ambiguity (the average number of senses per word is three) and cover a wide variety of parts of speech (for the full set of words and elicited sense frequencies see their Appendix A, pp. 501–609). One stumbling block to using this data are the meanings associated with the ambiguous words. These were provided by native English speakers and may not necessarily correspond to senses described by trained lexicographers. Fortunately, we were able to map most of them (except for six which we discarded) on WordNet synsets (version 1.6); two annotators performed the mapping by comparing the sense descriptions provided by Durkin and Manning to WordNet synsets. The annotators agreed in their assignments 81% of the time. Disagreements were resolved through mediation.

act	Freq
pretense/performance	37
to perform	30
to take action	16
division	12
a deed	3

answer	Freq
response	81
solution	18

Table 6: Meaning frequencies for *act* and *answer*; normative data from Durkin and Manning (1989)

Examples of Durkin and Manning's (1989) normative data are given in Table 6. The sense `response` for *answer* was mapped to the WordNet synset {`answer, reply, response`} (Sense 1), the sense `solution` was mapped to the synset {`solution, answer, result, resolution, solvent`} (Sense 2), etc. Durkin and Manning did not take part of speech ambiguity into account, as Table 6 shows, subjects came up with meanings relating to the verb and noun part of speech of *act*.

We explored the relationship between the sense frequencies provided by human subjects and those estimated by our model by computing the Spearman rank correlation coefficient ρ. We obtained sense frequencies from the BNC using the best model from Section 4.1 (High, +Norm, −Hyp). We found that the resulting sense frequencies were significantly correlated with the human sense frequencies ($\rho = 0.384$, $p < 0.01$). We performed the same experiment using McCarthy et al.'s (2004a) model, which also achieved a significant correlation ($\rho = 0.316$, $p < 0.01$). This result provides an additional validation of our model as it demonstrates that the sense frequencies it generates can capture the sense preferences of naive human subjects (rather than trained lexicographers).

5 Discussion

In this paper we proposed an IR-based approach for inducing dominant senses automatically. Our method estimates the degree of association between words and their sense descriptions (represented by synsets in WordNet) simply by querying an IR engine. Evaluation on the Senseval data sets showed that our model significantly outperformed a naive random sense baseline and a more competitive one

based on synset size. Our method was significantly better than McCarthy et al. (2004b) on Senseval-2 and Senseval-3. On the latter data set, its performance was comparable to that of the best unsupervised system (Strapparava et al., 2004).

An important future direction lies in evaluating the disambiguation potential of our models across domains and languages. Furthermore, our experiments have relied on WordNet for providing the appropriate sense descriptions. Future work must assess whether the models presented in this paper can be extended to alternative sense inventories (e.g., dictionary definitions) that may differ in granularity and structure. We will also experiment with a wider range of lexical association measures for quantifying the similarity of a word and its synonyms. Examples include odds ratio (Mohammad and Hirst, 2006) and Turney's (2001) IR-based pointwise mutual information (PMI-IR).

Our experiments revealed that the IR-based model is particularly good at disambiguating certain parts of speech (e.g., verbs, see Tables 2 and 5). A promising direction is the combination of different ranking models (Brody et al., 2006) and the integration of dominant sense models with supervised WSD.

Acknowledgments We are grateful to Diana McCarthy for her help with this work. The authors acknowledge the support of EPSRC (grant EP/C538447/1).

References

Briscoe, Ted and John Carroll. 2002. Robust accurate statistical annotation of general text. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation*. Las Palmas, Gran Canaria, pages 1499–1504.

Brody, Samuel, Roberto Navigli, and Mirella Lapata. 2006. Ensemble methods for unsupervised WSD. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*. Sydney, Australia, pages 97–104.

Budanitsky, Alexander and Graeme Hirst. 2001. Semantic distance in WordNet: An experimental, application-oriented evaluation of five measures. In *Proceedings of the NAACL Workshop on WordNet and Other Lexical Resources*. Pittsburgh, PA.

Decadt, Bart, Véronique Hoste, Walter Daelemans, and Antal van den Bosch. 2004. GAMBL, genetic algorithm optimization of memory-based WSD. In Mihalcea and Edmonds (2004), pages 108–112.

Durkin, Kevin and Jocelyn Manning. 1989. Polysemy and the subjective lexicon: Semantic relatedness and the salience of intraword senses. *Journal of Psycholinguistic Research* 18(6):577–612.

Fellbaum, Christiane, editor. 1998. *WordNet: An Electronic Database*. MIT Press, Cambridge, MA.

Firth, J. R. 1957. *A Synopsis of Linguistic Theory 1930-1955*. Oxford: Philological Society.

Gale, William A., Kenneth W. Church, and David Yarowsky. 1992. A method for disambiguating word senses in a large corpus. *Computers and the Humanities* 26(5–6):415–439.

Harris, Zellig. 1982. Discourse and sublanguage. In R. Kittredge and J. Lehrberger, editors, *Language in Restricted Semantic Domains*, Walter de Gruyter, Berlin; New York, pages 231–236.

Lin, Dekang. 1998. An information-theoretic definition of similarity. In *Proceedings of the 15th International Conference on Machine Learning*. Madison, WI, pages 296–304.

Manber, Udi and Sun Wu. 1994. GLIMPSE: a tool to search through entire file systems. In *Proceedings of USENIX Winter 1994 Technical Conference*. San Francisco, CA, pages 23–32.

McCarthy, Diana, Rob Koeling, Julie Weeds, and John Carroll. 2004a. Finding predominant senses in untagged text. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics*. Barcelona, pages 279–286.

McCarthy, Diana, Rob Koeling, Julie Weeds, and John Carroll. 2004b. Using automatically acquired predominant senses for word sense disambiguation. In Mihalcea and Edmonds (2004), pages 151–154.

Mihalcea, Rada and Phil Edmonds, editors. 2004. *Proceedings of Senseval-3: The 3rd International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*. Barcelona.

Mohammad, Saif and Graeme Hirst. 2006. Determining word sense dominance using a thesaurus. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics*. Trento, Italy, pages 121–128.

Palmer, Martha, Christiane Fellbaum, Scott Cotton, Lauren Delfs, and Hoa Trang Dang. 2001. English tasks: All words and verb lexical sample. In *Proceedings of Senseval-2: The 3rd International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*. Toulouse.

Snyder, Benjamin and Martha Palmer. 2004. The English all-words task. In Mihalcea and Edmonds (2004).

Stokoe, Christopher. 2005. Differentiating homonymy and polysemy in information retrieval. In *Proceedings of the Human Language Technology Conference and the Conference on Empirical Methods in Natural Language Processing*. Vancouver, pages 403–410.

Strapparava, Carlo, Alfio Gliozzo, and Claudio Giuliano. 2004. Word-sense disambiguation for machine translation. In Mihalcea and Edmonds (2004), pages 229–234.

Turney, Peter D. 2001. Mining the web for synonyms: PMI-IR versus LSA on TOEFL. In *Proceedings of the 12th European Conference on Machine Learning*. Freiburg, Germany, pages 491–502.

Twilley, L. C., P. Dixon, D. Taylor, and K. Clark. 1994. University of Alberta norms of relative meaning frequency for 566 homographs. *Memory and Cognition* 22(1):111–126.

Vickrey, David, Luke Biewald, Marc Teyssier, and Daphne Koller. 2005. Word-sense disambiguation for machine translation. In *Proceedings of the Human Language Technology Conference and the Conference on Empirical Methods in Natural Language Processing*. Vancouver, pages 771–778.

Yarowsky, David and Radu Florian. 2002. Evaluating sense disambiguation across diverse parameter spaces. *Natural Language Engineering* 9(4):293–310.