WASP-Bench: a Lexicographic Tool Supporting Word Sense Disambiguation

David Tugwell & Adam Kilgarriff
ITRI, University of Brighton
Lewes Road, Brighton BN2 4GJ, UK
{David.Tugwell,Adam.Kilgarriff}@itri.bton.ac.uk

Abstract

We present WASP-Bench: a novel approach to Word Sense Disambiguation, also providing a semi-automatic environment for a lexicographer to compose dictionary entries based on corpus evidence. For WSD, involving lexicographers tackles the twin obstacles to high accuracy: paucity of training data and insufficiently explicit dictionaries. For lexicographers, the computational environment fills the need for a corpus workbench which supports WSD. Results under simulated lexicographic use on the English lexical-sample task show precision comparable with supervised systems[1], without using the laboriously-prepared training data.

1 Introduction

WASP-Bench[2] is a web-based tool supporting both corpus-based lexicography and Word Sense Disambiguation. The central premise behind the initiative is that deciding what the senses for a word are, and developing a WSD program for it, should be tightly coupled. In the course of the corpus analysis, the lexicographer explores the textual clues that indicate a word is being used in one sense or another; given an appropriate computational environment, these clues can be gathered and used to seed a bootstrapping WSD program.

This strategy clearly requires human input for each word to be disambiguated, which may raise the objection that the lexicon is far too large for any word-by-word work to be viable. However, the amount of human interaction needed is far less than that involved in preparing training data[3] and lexicographers are already in the position of having to inspect every word in the vocabulary. If they use a interactive tool such as the WASP-Bench to help them in this, then total coverage becomes a feasible proposition.

2 WASP-Bench Methodology

The workbench is implemented in perl and uses cgi-scripts and a browser for user interaction.

2.1 Grammatical relations database

The central resource is a collection of all grammatical relations holding between words in the corpus. The corpus currently used in WASP-Bench is the British National Corpus[4] (BNC): . Using finite-state techniques operating over part-of-speech tags, we process the whole corpus finding quintuples of the form: {**Rel**, **W1**, **W2**, **Prep**, **Position**}, where **Rel** is a relation, **W1** is the lemma of the word for which **Rel** holds, **W2** is the lemma of the other open-class word involved, **Prep** is the preposition or particle involved and **Position** is the position of **W1** in the corpus. Relations may have null values for **W2** and **Prep**. The database contains 70 million quintuples.

The current inventory of relations is shown in Table 1. All inverse relations, ie. **subject-of** etc, found by taking W2 as the head word instead of W1 are explicitly represented, to give a total of twenty-six distinct relations. These provide a flexible resource to be used as the basis of the computations of the workbench. Keeping

[1]It should be noted that the lower figure for recall reflects solely the fact that not all words were attempted due to time constraints.

[2]The system has been developed under EPSRC project M54971. A demo is available at *http://wasps.itri.bton.ac.uk*. The second author was also a co-ordinator for the SENSEVAL-2 evaluation exercise–to limit any conflict of interest only the first author was involved applying the system to the SENSEVAL-2 task, and had no prior knowledge of the format of the task.

[3]See results section for details.

[4]100 million words of contemporary British English. see http://info.ox.ac.uk/bnc

relation	example
bare-noun	the angle of **bank**^{1}
possessive	my **bank**^{1}
plural	the **banks**^{1}
passive	was **seen**^{1}
reflexive	**see**^{1} herself
ing-comp	**love**^{1} eating fish
finite-comp	**know**^{1} he came
inf-comp	**decision**^{1} to eat fish
wh-comp	**know**^{1} why he came
subject	the **bank**^{2} **refused**^{1}
object	**climb**^{1} the **bank**^{2}
adj-comp	**grow**^{1} **certain**^{2}
noun-modifier	**merchant**^{2} **bank**^{1}
modifier	a **big**^{2} **bank**^{1}
and-or	**banks**^{1} and **mounds**^{2}
predicate	**banks**^{1} are **barriers**^{2}
particle	**grow**^{1} **up**^{p}
Prep+gerund	**tired**^{1} **of**^{p} eating fish
PP-comp/mod	**banks**^{1} **of**^{p} the **river**^{2}

Table 1: Grammatical Relations

the position numbers of examples allows us to find associations between relations and to display examples.

2.2 Word Sketches

The user enters the word and using the grammatical relations database, the system composes a **word sketch** for this word. This is a page of data such as Table 2, which shows, for the word in question (W1), ordered lists of high-salience grammatical relations, relation-W2 pairs, and relation-W2-Prep triples for the word.

The number of patterns shown is set by the user, but will typically be over 200. These are listed for each relation in order of salience, with the count of corpus instances. The instances can be instantly retrieved and shown in a concordance window. Producing a word sketch for a medium-to-high frequency word takes in the order of ten seconds.

Salience is calculated as the product of Mutual Information I (Church and Hanks, 1989) and log frequency. I for a word $W1$ in a grammatical relation Rel[5] with a second word $W2$ is calculated as:

$$I(W1, Rel, W2) = log(\frac{\|*, Rel, *\| \times \|W1, Rel, W2\|}{\|W1, Rel, *\| \times \|*, Rel, W2\|})$$

The notation here is adopted from (Lin, 1998) (who also spells out the derivation from the definition of I). $\|W1, Rel, W2\|$ denotes the frequency count of the triple $\{W1, Rel, W2\}$[6] in the grammatical relations database. Where $W1$, Rel or $W2$ is the wild card ($*$), the frequency is of all the dependency triples that match the remainder of the pattern.

The word sketches are presented to the user as a list of relations, with items in each list ordered according to salience. Our experience of working lexicographers' use of Mutual Information or log-likelihood lists shows that, for lexicographic purposes, these over-emphasise low frequency items, and that multiplying by log frequency is an appropriate adjustment.

2.3 Matching patterns with senses

The next task is to enter a preliminary list of senses for the word, possibly in the form of some arbitrary mnemonics: for example, MONEY, CLOUD and RIVER for three senses of *bank*.[7] This inventory may be drawn from the user's knowledge, from a perusal of the word sketch, or from a pre-existing dictionary entry.

As Table 2 shows, and in keeping with "one sense per collocation" (Yarowsky, 1993) in most cases, high-salience patterns or **clues** indicate just one of the word's senses. The user then has the task of associating, by selecting from a pop-up menu, the required sense for unambiguous clues. The number of relations marked will depend on the time available, as well as the complexity of the sense division to be made. The act of assigning senses to patterns may very well lead the user to discover fresh, unconsidered senses usages of the word.

The pattern-sense associations are then submitted to the next stage: automatic disambiguation.

2.4 The Disambiguation Algorithm

The workbench currently uses Yarowsky's decision list approach to WSD (Yarowsky, 1995). This is a bootstrapping algorithm that, given

[5] {Grammatical-relation, preposition} pairs are treated as atomic relations in calculating MI.

[6] Or, strictly, of the quintuple $\{W1, Rel - part - 1, W2, Rel - part - 2, ANY\}$.

[7] WASP-Bench can also be used for Machine Translation lexicography, where arbitrary mnemonics would be replaced by target language translations.

subj-of	num	sal	obj-of	num	sal	modifier	num	sal	n-mod	num	sal
lend	95	21.2	burst	27	16.4	central	755	25.5	merchant	213	29.4
issue	60	11.8	rob	31	15.3	Swiss	87	18.7	clearing	127	27.0
charge	29	9.5	overflow	7	10.2	commercial	231	18.6	river	217	25.4
operate	45	8.9	line	13	8.4	grassy	42	18.5	creditor	52	22.8
modifies			**PP**			**inv-PP**			**and-or**		
holiday	404	32.6	of England	988	37.5	governor of	108	26.2	society	287	24.6
account	503	32.0	of Scotland	242	26.9	balance at	25	20.2	bank	107	17.7
loan	108	27.5	of river	111	22.1	borrow from	42	19.1	institution	82	16.0
lending	68	26.1	of Thames	41	20.1	account with	30	18.4	Lloyds	11	14.1

Table 2: Extract of word sketch for *bank*

some initial seeding, iteratively divides the corpus examples into the different senses. Yarowsky notes that the most effective initial seeding option he considered was labelling salient corpus collocates with different senses. The user's first interaction with the workbench is just this.

At the user-input stage, only clues involving grammatical relations are used. At the WSD algorithm stage, some "bag-of-words" and n-gram clues are also considered. Any content word (lemmatised) occurring within a k-word window of the nodeword is a bag-of-words clue.[8] N-gram clues capture local context which may not be covered by any grammatical relation. The n-gram clues are all bigrams and trigrams including the nodeword. N-grams and context-word clues frequently duplicate the grammatical relations already found, but the merit of the decision list approach is that probabilities are not combined, so such dependencies are not a problem.

2.5 Sense Profiles

The output of the algorithm is both a sense disambiguated corpus, and a decision list. The decision list can be viewed as a lexical entry or as a WSD program. It will contain {Rel, W2} pairs (as in the original word sketch), bag-of-words words, and n-grams. The components of the decision list which assign to a particular sense can be displayed as "sense profiles", in a manner comparable to the original word sketch. They will contain new clues, not originally seen in the word sketch and may point to new senses or usages needing addition to the lexical entry. Users can then re-run the WSD algorithm, iterating until they are satisfied with the sense inventory, and with the accuracy of the disambiguation performed.

[8]The user can set the value of k. The default is currently 30.

3 Evaluating the workbench

3.1 Lexicographic evaluation

For the last two years, a set of 6000 word sketches has been used in a large dictionary project (Rundell, 2002), with a team of thirty professional lexicographers covering every medium-to-high frequency noun, verb and adjective of English. The feedback received is that they are hugely useful, and transform the way the lexicographer uses the corpus. They radically reduce the amount of time the lexicographers need to spend reading individual instances, and give the dictionary improved claims to completeness, as common patterns are far less likely to be missed.

3.2 Results for senseval-2

Performance as a WSD system was evaluated on the SENSEVAL-2 English lexical sample exercise.

The words to be tested were divided between the first author and one paid volunteer, who had no previous experience of WASP-Bench. We carried out the procedure as above, with the difference that instead of having to establish a sense inventory, the inventory was already given as that of WordNet. After assigning sufficient clues to cover the various senses, these assignments were submitted as seeds to the disambiguation algorithm. Using the example sentences from the BNC this gave us a decision list of clues, which could then be used to disambiguate the test sentences.

The marking of senses took anywhere from 3 to 35 minutes, depending upon the subtlety of the sense divisions to be made. The average time was around 15 minutes per word. A substantial part of this was taken up by reading and understanding the dictionary entry even before patterns were marked. Crucially we made no use of the training data,[9] although this would certainly have been of use as a reference in clarifying the sense distinctions to be made. Unfortunately, due to severe time constraints, it only proved possible to carry out analysis for the 29 nouns and 15 adjectives in the lexical sample, and there was no time to carry out the analysis of the verbs.[10]

Results on the task were 66.1% for coarse-grained precision and 58.1% for fine-grained.[11] This was significantly higher than other systems which did not use the training data (the best scores being 51.8% for coarse-grained and 40.2% for fine-grained precision), demonstrating that the relatively small amount of human interaction is very beneficial. Indeed, the system's performance was similar to the majority of systems which had used the training data.

3.2.1 Significant problems

The most pervasive problem was the difficulty of getting a clear conception of the sense distinctions made in the inventory, here WordNet. Without this, assigning putative senses to clues could be an exasperating and painful task.

To illustrate, for the adjective *simple* there were no less than 13 sense distinctions to be made, the first two of which were particularly hard to distinguish:

1. simple (vs. complex) – (not complex or complicated or involved): *a simple problem*

2. elementary, simple, uncomplicated, unproblematic – (not involved or complicated): *an elementary problem in statistics*

[9]In fact, we had to download the data to find out the words to be tested, but made no other use of it.

[10]Also no results were returned for the noun *day*, as processing the 93,000+ examples in the BNC led to an processing delay that could not be fixed in time.

[11]Due to the limited number of words attempted the figures for recall were 36.3% and 31.9%. It should be understood that there was no precision/recall tradeoff here—the system returned an answer for all sentences in the words it covered.

Unsurprisingly, the system fared particularly badly here with 37.9% precision, while inter-annotator agreement was also low at 67.8%.

3.2.2 Previous results

We previously measured the performance of the system on the dataset from the SENSEVAL-1 exercise (Kilgarriff and Palmer, 2000) under similar conditions of use. Results for the WASP-Bench here were significantly higher at 74.9% precision which was very close to the best supervised system (within 1%). This was undoubtedly due to the clearer sense distinctions and greater number of examples to be found in the sense inventory used for this task in SENSEVAL-1, which made it possible to assign senses to clues with more confidence.

4 Summary

The results for the WASP-Bench show that high-quality disambiguation can be achieved with much less human interaction than is needed for preparing a training corpus. Furthermore, this interaction can be motivated since it has been shown to be of proven benefit for the users of the system: lexicographers. Establishing this synergy may prove to be of great importance for both camps.

References

Kenneth Church and Patrick Hanks. 1989. Word association norms, mutual information and lexicography. In *ACL Proceedings, 27th Annual Meeting*, pages 76–83, Vancouver.

Adam Kilgarriff and Martha Palmer. 2000. Introduction, Special Issue on SENSEVAL: Evaluating Word Sense Disambiguation Programs. *Computers and the Humanities*, 34(1–2):1–13.

Dekang Lin. 1998. Automatic retrieval and clustering of similar words. In *COLING-ACL*, pages 768–774, Montreal.

Michael Rundell. 2002. *Macmillan English Dictionary for Advanced Learners.* Macmillan.

David Yarowsky. 1993. One sense per collocation. In *Proc. ARPA Human Language Technology Workshop*, Princeton.

David Yarowsky. 1995. Unsupervised word sense disambiguation rivalling supervised methods. In *ACL 95*, pages 189–196, MIT.