An Evaluation of a Lexicographer's Workbench: building lexicons For Machine Translation

Rob Koeling
COGS, University of Sussex
robk@cogs.susx.ac.uk

Adam Kilgarriff, David Tugwell, Roger Evans
ITRI, University of Brighton
{adam,david,roger}@itri.bton.ac.uk

Abstract

NLP system developers and corpus lexicographers would both benefit from a tool for finding and organizing the distinctive patterns of use of words in texts. Such a tool would be an asset for both language research and lexicon development, particularly for lexicons for Machine Translation (MT). We have developed the WASPBENCH, a tool that (1) presents a "word sketch", a summary of the corpus evidence for a word, to the lexicographer; (2) supports the lexicographer in analysing the word into its distinct meanings and (3) uses the lexicographer's analysis as the input to a state-of-the-art word sense disambiguation algorithm, the output of which is a "word expert" for the word which can then disambiguate new instances of the word. In this paper we describe a set of evaluation experiments, designed to establish whether WASPBENCH can be used to save time and improve performance in the development of a lexicon for Machine Translation or other NLP application.

Motivations

On the one hand, Human Language Technologies (HLT) need dictionaries, to tell them what words mean and how they behave. On the other hand, the people making dictionaries (hereafter, lexicographers) need HLT, to help them identify how words behave so they can make better dictionaries. This potential for synergy exists across the range of lexical data - in the construction of headword lists, for spelling correction, phonetics, morphology and syntax, but nowhere is it truer than for semantics, and in particular the vexed question of how a word's meaning should be analysed into distinct senses. HLT needs all the help it can get from dictionaries, because it is a very hard problem to identify which meaning of a word applies, and if the dictionary does not provide both a coherent and accurate analysis of what the meanings are, and a good set of clues as to where each meaning applies, then the enterprise is doomed. The lexicographer needs all the help they can get because the analysis of meaning is the second hardest part of their job (Kilgarriff, 1998), it occupies a large share of their working hours, and it is one where, currently, they have very little to go on beyond intuition.

Synergy between HLT and lexicographer becomes a possibility with the advent of the corpus. Lexicographers have long been aware of their great need for evidence about how words behave. The pioneering project was COBUILD (Sinclair, 1987) and its first offering to the world, the Collins COBUILD English Dictionary came out in 1987. The basic working methodology, in those early days, was the 'coloured pens' method. A lexicographer who was to write an entry for a word, say *pike,* was given the corpus evidence for *pike* in the form of a key-word-in-context printout. They then read the corpus lines, identifying different meanings as they went along, assigning a colour to each meaning and marking each corpus line with the appropriate colour. Once they had marked all (or almost all - there are always anomalies) the corpus lines, they could then go back to write a definition

for each sense, using, eg, the red corpus lines as the evidence for the first meaning, the green as the evidence for the second, and so on.

In this scenario, note that a meaning, or word sense, corresponds to a cluster of corpus lines. This is a representation that HLT can work with. As corpus-based HLT took off, in the 1990s, researchers such as (Gale et al., 1993) explored corpus methods for word sense disambiguation (WSD). Here the correspondence between word senses and sets of corpus lines was taken at face value, with a set of corpus lines which were known to belong to a particular sense being used as a training set. A machine-learning algorithm was then able to use the training set to induce a word expert which could decide which sense a new corpus instance belonged to.

So the stage is set for software which both uses HLT to support the corpus lexicographer in developing good meaning analyses, and uses the meaning analysis, realised as corpus evidence, to support accurate WSD. This is what the WASPBENCH aims to do.

1.1 The WASPBENCH system

Behind the current implementation of the English WASPBENCH lies a database of 70M instances of grammatical relations for English. These are 5-tuples:

< *gramrel, word1, word2, particle, pointer* >

gramrel can be any of a set of 27 core grammatical relations for English (including *subject, subject-of, object, object-of, modifier, and/or, PP-comp), wordl* and *word2* are words of English (nouns, verbs or adjectives, lemmatized to give dictionary headword form; *word2* may be null), *particle* is a particle or preposition, so that grammatical relations involving prepositions as well as two fully lexical arguments can be captured. For all relations except *PP-comp* it is null. *Pointer* points into the corpus, so we can identify where the instance occurs and retrieve its context if required. Examples of 5-tuples are

PP-comp,look,picture,at, 1004683
object, sip, beer, -, 1005678

The database was prepared by parsing a lemmatised, part-of-speech-tagged version of the British National Corpus, a 100M word corpus of recent spoken and written British English.[1]

Using this database, WASPBENCH prepares a set of lists for each *word1* in which, for each *gramrel,* the words which occur frequently and with high mutual information as *word2* are identified and sorted according to their lexicographic salience. This set of lists is presented to the lexicographer for whom it is a useful summary of the word's behaviour. This is a *word sketch* (Kilgarriff and Tugwell, 2001b).

The word sketch is a good starting point for the lexicographer to analyse the different meanings (step 1). They study it. All underlying corpus evidence is available at a mouseclick, in case they are unsure what contexts *word1* occurs in *gramrel* with *word2* in. They reach preliminary opinions about the different meanings the word has. They assign a short mnemonic label to each sense, and type the labels into a text-input box provided. Hitting the "set senses" button updates the word sketch, with each collocate now having a pull-down menu through which it can be assigned to one of the senses.

The lexicographer then spends some time - typically some thirty minutes for a moderately complicated word- assigning collocates to senses (step 2). The majority of high-salience < *collocate, gramrel* > pairs relate to one sense of a word only (in accordance with Yarowsky's "one sense per collocation" dictum (Yarowsky, 1993)), and it is usually immediately evident which sense is salient, so the task is not unduly taxing. The lexicographer does not have to assign all, or any particular, collocate, and any collocate which is associated with more than one sense should be left unassigned.

When the lexicographer has assigned a good range of collocates, they press "submit". The WSD algorithm takes over, using the corpus instances where the collocates assigned by the lexicographer apply as the clusters of instances corresponding to a sense, and bootstrapping further evidence about how other corpus instances are assigned (step 3). The algorithm produces a *word expert* which can disambiguate new instances of the

[1] http://info.ox.ac.uk/bnc

word. The algorithm currently in use is a reimplementation of Yarowski's decision list learner (Yarowsky, 1995).

1.2 WASPBENCH and Machine Translation

WASPBENCH is designed particularly with the needs of MT lexicography in mind. In that context, the components of the problem take on a slightly different form, sometimes with different names. MT has long needed many rales of the form,

> *in context* **C**, *translate source language word* **S** *as target language word* **T**

The problem has traditionally been that these rules are hard for humans to identify, and, as there is a large number of possible contexts for most words and a large number of ambiguous words, a very large number of rules is needed. In step (1), the word sketch, WASPBENCH identifies and displays to the user a good set of candidate rules but with the target word **T** unspecified. In step (2), it supports the assignment of target words, by the lexicographer, for a number of the rules. In step (3), it takes this small set of rules and uses a bootstrapping algorithm to automatically identify a very large set of rules, so the word can be appropriately translated wherever it occurs (Kilgarriff and Tugwell, 2001a).

2 Evaluating WASPBENCH

Evaluating how successful we have been in developing the WASPBENCH presents a number of challenges.

- We straddle three communities - the (largely commercial) dictionary-making world, the (largely research) Human Language Technology (and specifically, WSD) world, and the (part commercial, part research) MT world, all with very different ideas about what makes a technology useful.

- There are no precedents. WASPBENCH performs a function - corpus-based disambiguating-lexicon development with human input - which no other technology performs. We believe no other technology provides even a remotely similar combination of inputs (corpus + human) and outputs (meaning analysis + word expert). This leaves us with no other products to compare it with.

- On the lexicography front: human analysis of meaning is decidedly 'craft' (or even 'art') rather than 'science'. WASPBENCH is aiding the practitioners of this craft in doing their job better and faster. But, in the dictionary world, even qualitative analyses of the relative merits of one meaning analysis as against another are rare treats. Quantitative evaluations are unheard of.

- A critical question for commercial MT would be "does it take less time to produce a word expert using WASPBENCH than using traditional methods, for the same quality of output". We are constrained in pursuing this route because we do not have access to MT companies' lexicography budgets, and moreover consider it unlikely that MT companies would view the production of disambiguation rales as a distinct function in the way that we do.

In the light of these issues, we have adopted a 'divide and rule' strategy, setting up different evaluation themes for different perspectives. We have pursued five different evaluation strategies. One of them is the subject of this paper.[2] Of the other strategies, we only mention the application of word sketches within a large scale commercial lexicography project here (the production of Macmillan English Dictionary for Advanced Learners) (Kilgarriff and Rundell, 2002). The set of experiments that we report on in this paper explored the performance of WASPBENCH-based translations in comparison with translations produced by commercial MT systems.

3 Experimental setup

A group of twelve people were involved in the experiment. All were students in translation studies at the University of Leeds. None of them had a

[2] A report bringing together evidence from all evaluation approaches is in preparation.

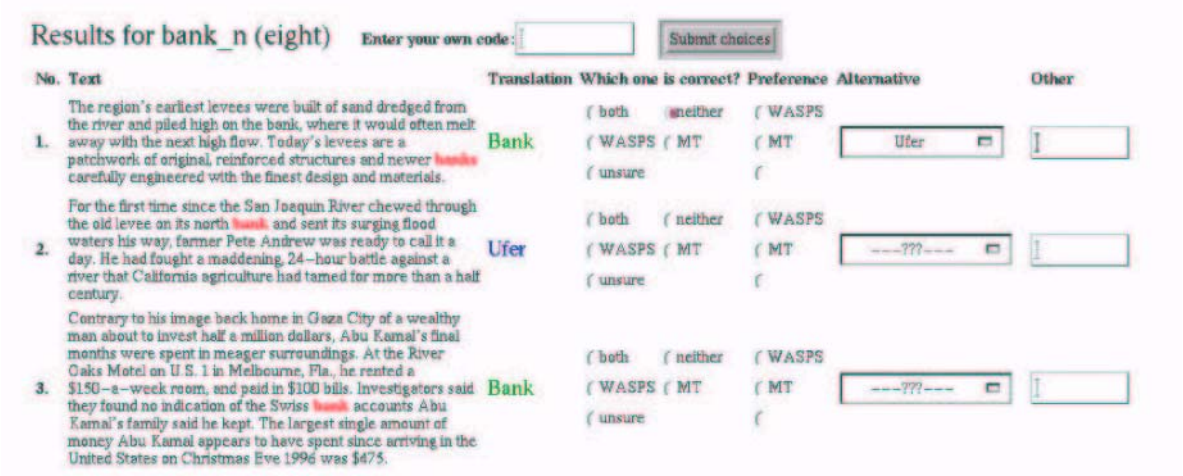

Figure 1 : Snapshot of the evaluation screen

specific background in lexicography. They were all native or near-native speakers of both English and the language they worked with for the experiment. The students worked with Chinese (4), French (3),German (2) and Italian (1).[3]

We asked the participants to work with the WASPBENCH; creating word experts for the selected words. This task gave us information about how the users experienced using the workbench, either explicitly, by giving us feedback, or implicitly by supplying us with data. This part of the experiment created the word experts. The other task was to evaluate the word experts. We applied their word-experts to a set of previously unseen test sentences and compared the output of the WASPBENCH with the output of a commercial MT system.

Creating the word experts The main task for the participants was to use the WASPBENCH to create word experts for a list of selected ambiguous English words. The evaluation task focussed on translation. The user was asked to use the WASPBENCH in order to find out how the word was used in English (i.e. as represented by the BNC) and how the different uses of this word would be translated into a target language of the participant's choice. After the user has chosen the translations for the word and selected the clues giving evidence for when the word should receive a particular translation, the user submits the data and the WASPBENCH infers further rules to complete the word expert. The user is presented the rule set and can manually inspect it. If they are happy with the set, they can decide to submit the word expert and continue with the next word. If they are not happy with the rule set, they can return to the wordsketch definition form and add to or amend their input. After submitting, the word expert is applied to a set of test sentences.

Assessing the results Evaluating a word expert is like evaluating the work of a translator. The work of a translator can be judged by someone else, who can disagree on certain decisions made by the translator. The disagreement can be a matter of personal style. The assessment task here involves the same kind of problem. In this experimental paradigm we do not define beforehand what the desired translation is. Every subject may identify a different set of target translations for each word and even if they work with the same set, people might disagree on the preferred translation of a word in a particular context. There is no gold standard and thus we cannot evaluate the decisions automatically. Therefore we asked the participants to assess the word experts' judgements.

The assessment task can best be introduced by looking at a screenshot. In figure 1 we present part of the evaluation screen with the results of ap-

[3] Two more students worked with Japanese, but at the time of the experiment we did not have the MT translations for Japanese available. Their word experts were evaluated in a different way. We do not discuss these results in this paper.

plying the word expert made by participant 'one' for the noun *bank* to the set of 45 test sentences. The assessor is asked to enter their own number for identification purposes. The second column gives the test sentences with the word we are interested in (here *bank)* highlighted. The third column presents the word expert's translation. The assesser is asked to judge the correctness of the translation in this particular context in the fourth column. It was our intention to either include the whole translated sentence as generated by the MT system on the screen (with the target word highlighted) or just the translated target word. However, last minute technical problems made this impossible and we had to provide the MT system output on paper. The assesser was asked to decide which translation was correct in the given context. The options given were 'WASPS', 'MT', 'both', 'neither', 'unsure' and combinations like 'both correct, but WASPS preferable'.

In case they disagree with the translation offered, they can pick their preferred translation from the pulldown menu in the fifth column **(Alternative).** This pulldown menu offers all the other suggested target translations for *bank* as defined by participant 'one'. In case the assesser thinks the proper target translation is not available, their choice can be entered in the last column **(Other).**

After judging all 45 test sentences, the assesser is asked to submit the form by pressing the button in the right upper corner.

3.1 Instruction and Available Time

Most participants had not worked with the WASPBENCH before. They were given a theoretical introduction and the opportunity afterwards to explore the user interface and its functionality by creating a word expert. The participants were allowed plenty of time to create the word expert and play with the WASPBENCH. They then applied the word expert to a set of test sentences and inspected the results, to conclude the introduction.

After the instruction session, approximately 4 days were allowed for working on the task: about two days for creating word experts and two days for assessment. The participants were instructed to take their time to create the word experts, but to keep in mind that we did not expect perfection. In order to finish all 33 words in two working days, only aproximately 30 minutes per word was available. We did not expect them to complete the full list. To ensure that every word on the list would be covered by equally many subjects, everyone was asked to start at a different position in the list of words.

3.2 The Data

Words For the experiment we chose a set of words that are clearly ambiguous in English. We only selected words that were fairly, but not extremely, common (i.e. with 1,500 - 20,000 instances in the BNC). A total of 33 words were selected: 16 nouns, 10 verbs and 7 adjectives. Some of the words have just two clearly distinct meanings in English, others have more. There may of course also be further, more subtle meaning distinctions. All of the words were checked to confirm that the 'clearly distinct meanings' receive different translations in at least one of the languages at our disposal (Dutch, German and French). While we had identified a set of meanings for the words in the course of this process, this set was never shown to the participants. They were asked to create their own word expert with its own inventory of meanings/translations. This might result in different sets of target translation for different languages. In some languages two distinct different meanings might be translated with the same word, while subtle meaning differences might produce different translations in the target language. It is, of course, possible that, where more than one participant was working on the same language, they disagreed on the one set of target translations.

Test Data In order to test the performance of the word experts, we selected for every word between 40 and 50 text fragments containing the target word. These fragments consisted of the complete sentence in which the word occurred plus one or two surrounding sentences. The test sentences were selected from the North American News Text Corpus.[4] Random samples were taken from the corpus and inspected for suitability. This was done

[4]Available from the Linguistic Data Consortium.

Language	Wasps	MT	both	neither	unsure
German	0.60% (0.41)	0.28% (0.09)	0.19%	0.26%	0.05%
French	0.61 (0.24)	0.45 (0.07)	0.37	0.28	0.04
Chinese	0.68 (0.32)	0.42 (0.05)	0.37	0.23	0.03
Italian	0.67 (0.44)	0.29 (0.06)	0.23	0.22	0.05
All	0.64 (0.35)	0.36 (0.07)	0.29	0.25	0.04

Figure 2: WASPBENCH results compared with MT per language

to make sure that the samples were usable (some samples, like words from headlines, did not have much surrounding text) and to ensure that for every identified distinct meaning there were at least some test sentences available. If we had chosen a large set of test sentences from the corpus, we could have relied on pure random selection to take care of the proper meaning distribution, but a considerably larger sample than the 45 test sentences taken here would be necessary to rely on that.

The fact that we used an American news corpus for the test sentences and that the WASPBENCH currently uses the BNC for creating the word experts caused another problem: some words are used differently in British and American English, for example *lot* which has the 'parking space' meaning in American but not British English.

MT translation The MT translations were produced with BabelFish from Systran.[5] The individual fragments (i.e. the sentence wit the ambiguous word in it plus 1 or 2 surrounding sentences) were submitted as separate paragraphs to the translation engine.

4 Evaluation of the Results

A total of 240 word experts were produced for 32 words.[6] This means that an average of 7.5 word experts per word are available. There were at least 5 different word experts for any word, the maximum number of word experts for one word is 10.

The results for the different words depend very much on the perceived ambiguity of the word and how closely related the different meanings for that word are. For example, a noun like *bank* with two clear and distinct meanings ('financial institution' and 'river bank') gave very good results, while the results for very ambiguous words like the noun *line* were quite poor. The table in figure 3 gives an overview of the results of applying the word experts to the test sentences and comparing the translation of the target word with the translation for that word given by the MT system. The data is presented here per language. The figures in bold face give the overall percentage of cases were the WASPBENCH or the MT system was considered to be right. This number is the sum of the percentage of cases were only WASPBENCH /MT was right (percentage in brackets after the bold face) and those cases where both were considered to have given the right translation.

The table in figure 3 presents the data per PoS tag. This table shows that the WASPBENCH performs slightly better on nouns (which is consistent with the comments we got from the participants, who thought that the nouns were less problematic than the verbs and adjectives).

The data shows that the WASPBENCH results consistently outperform the MT results by a considerable margin. We do have to take into acccount that the sample sentences in the test sets we used here were not taken from one particular domain, but a sample of general text. The gains for translating domain specific text might be less dramatic.

5 User Experience with the Workbench

The evaluation task did not only provide data; it also gave us feedback on working with the workbench. Many comments were given on the pre-

[5] Available over the web via Altavista: http://babelfish.altavista.com/

[6] We experienced problems with one of the nouns. The data for this word *(film')* was discarded.

PoS	Wasps	MT	both	neither	unsure
Noun	**0.69** (0.34)	**0.40** (0.06)	0.35	0.24	0.02
Verb	**0.61** (0.29)	**0.38** (0.05)	0.32	0.27	0.06
Adjective	**0.63** (0.32)	**0.41** (0.10)	0.31	0.24	0.04

Figure 3: WASPBENCH results compared with MT per Part of Speech

sentation of the data, missing navigation abilities, buttons and correction facilities and other user-interface issues. We will incorporate suggestions into future releases of the workbench.

An important issue is that people have difficulties with many of the grammatical relations, and instead, focus on example sentences. This is time-consuming and it would be better if we could clarify the grammatical relations, either on the same screen, or on demand (for example by making help available).

A source of confusion and irritation is PoS tagger errors and errors made in predicting the grammatical relations. It is clear that these components are critical for the usability of the workbench.

6 Conclusions and Further Research

We have already mentioned that the evaluation experiment have provided us with valuable feedback on how people experience working with the WASPBENCH, giving us the opportunity to further develop the workbench. Several changes in the user interface will be made and will improve the usability of the tool. The main objective for this particular experiment, however, was to investigate how well the word-experts created with the WASPBENCH help to disambiguate words in a translation task. These experiments show that with the WASPBENCH it is possible to create word sense disambiguation rules that help translation of ambiguous words enormously without spending a whole lot of time in creating these rales. The results show that people, with no prior experience using the workbench, are able to create disambiguation rules that outperformed a well-established MT system by a great length, even though they had limited time to spend on creating the rules and did not have the opportunity to improve on their efforts.

While thinking about the WASPBENCH as a tool for improving WSD for MT systems, one of the questions we asked ourselves was: "does it take less time to produce a word expert using WASPBENCH than using traditional methods, for the same quality of output". Even though we can't answer this question, we do know now that we can improve substantially upon the quality of the output. We can also estimate the cost (in time or money) to create disambiguation rules for all the words and estimate the improvement in quality it will give us.

Another important aspect of the evaluation results is the fact that the results for the different languages are very similar. We feel that consistency is important for a disambiguation tool. Even though the word experts created by the participants will always be different, they should ideally behave similarly. In another experiment (Koeling and Kilgarriff, 2002) we looked explicitly at the consistency of results by comparing word experts (same word, same target language) made by several people. In that experiment we found more evidence for our consistency claim.

Even though we feel that these experiments show that the WASPBENCH successfully meets many of the goals we had in mind when we designed the workbench, there are still ways to improve the current system. The fronts on which we would like to develop the WASPBENCH include:

exploring alternative WSD algorithms (Yarowsky and Florian, 2002) show that "winner-take-all" algorithms, are sometimes preferable, but sometimes cumulative algorithms, where evidence from different clues is summed, perform better. We would like to explore how we might match the algorithm-type to the data instance.

interactivity Currently there is only minimal support for a 'second round' of the lexicogra-

pher revising their meaning analysis according to the feedback provided by the WSD algorithm. We would like the system to enter a dialogue with the lexicographer, whereby it identified anomalies and facilitated revisions to the meaning analysis.

multiwords Although some fuctionality for multiwords is already supported, for phrasal verbs and subcategorising nouns and adjectives, through the three-argument *prep_n,* relation, we would like to extend system functionality by permitting the user to input multiwords, for which collocations would be found.

thesaurus We have already produced a thesaurus from the database (see http://wasps.itri.bton.ac.uk), using Lin's similarity measure (Lin, 1998). We would like to use the thesaural classes in the word sketches and elsewhere, so that evidence from words in the same thesaural class could be pooled, and inferences drawn where two words were not encountered together, but their thesaural classes had high mutual information.

other languages Developments for a number of languages other than English are under way. Once we have two databases of grammatical relations, based on comparable corpora, for different languages, the potential for mapping tuples between the databases (using a bilingual dictionary) arises.

new corpora there's no data like more data, and both wordsketch production and the WSD learning algorithm work better, the more they are fed. Using the BNC, we have insufficient data to say much about words beyond the commonest 20,000 in the language, and miss many patterns. We are exploring using the web (suitably filtered) as the input corpus.

Acknowledgements

This work was supported by the UK EPSRC, under the WASPS project, grant GR/M54971. We would like to thank Prof. Tony Hartley from Leeds University for organising the experiments.

References

COBUILD, 1987. *The Collins COBUILD English Language Dictionary.* Edited by John McH. Sinclair *et al.* London.

William Gale, Kenneth Church, and David Yarowsky. 1993. A method for disambiguating word senses in a large corpus. *Computers and the Humanities,* 26(1-2):415-539.

Adam Kilgarriff and Michael Rundell. 2002. Lexical profiling software and its lexicographical applications - a case study. In *EURALEX 02,* Copenhagen.

Adam Kilgarriff and David Tugwell. 2001a. Waspbench: an MT lexicographer's workstation supporting state-of-the-art lexical disambiguation. In *Proc. MT Summit VIII,* pages 187-190, Santiago de Compostela, Spain, September.

Adam Kilgarriff and David Tugwell. 2001b. Word sketch: Extraction and display of significant collocations tor lexicography. In *Proc. Collocations workshop, ACL 2001,* pages 32-38, Toulouse, France.

Adam Kilgarriff. 1998. The hard parts of lexicography. *International Journal of Lexicography, 11* (l):51-54.

Rob Koeling and Adam Kilgarriff. 2002. Evaluating the waspbench, a lexicography tool incorporating word sense disambiguation. In *Proceedings of ICON 2002,* Mumbai, India, December.

Dekang Lin. 1998. Automatic retrieval and clustering of similar words. In *Proceedings of ACL,* Montreal.

John M. Sinclair, editor. 1987. *Looking Up: An Account of the COBUILD Project in Lexical Computing.* Collins, London.

David Yarowsky and Radu Florian. 2002. Evaluating sense disambiguation performance across diverse parameter spaces. *Journal of Natural Language Engineering,* page In press. Special Issue on Evaluating Word Sense Disambiguation Systems.

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

David Yarowsky. 1995. Unsupervised word sense disambiguation rivaling supervised methods. In *Proc. of ACL 1995,* pages 189-196, Cambridge, MA.