An evaluation of a lexicographer's workbench incorporating word sense disambiguation

Adam Kilgarriff and Rob Koeling

Information Technology Research Institute (ITRI), Brighton, 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. 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 (WSD) 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.

1 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 (herafter, 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 MT version of the problem is to find the appropriate translation for a word in a given context, where the bilingual dictionary gives several possibilities, and this is just as hard. The lexicographer needs all the help they can get because the analysis of meaning is the second hardest part of their job [1], 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, and, in the late 1970s and 1980s, English language dictionary publishers were rather quicker to pick up on the potential of large corpora than most parts of the HLT world. The pioneering project was COBUILD [2] 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, as in figure 1. 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, the yellow as the evidence for the third, 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. (It contrasts with a conception of word senses as mental objects, which is not useful to HLT.)

As corpus-based HLT took off, in the 1990s, researchers such as [4] 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 (or believed) 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.

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*), *word1* 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
```

A65 1065	On Tuesday we opted for a more strenuous hike from Braithwaite village up the steep sloped of Grisedale	**pike**	.
A6R 13	Only when it was in the net did I realise what size it was and it weighed 26 ob 8 oz,'; added John who went on to bank five other	**pike**	, two of 8 lb 8 oz, and others of 10 lb, 11lb and 14 lb.
A6R 390	I hit the fish and stright away though it was a good one, but my son Tony has never caught a	**pike**	so I handed him the rod,'; said Lee.
A6R 825	Skimmers, roach and small perch from most Liverpool sections but	**pike**	active.
A6R 950		**pike**	to 17 lb 2 oz showing.
A6R 1130	Roach at Bishop Monkton,	**pike**	around the canal mouth.
A7C 1528	Press forward every gallant man With hatchet,	**pike**	and gun!
AA0 92	Cardiff City: Wood; Rodgerson, Daniel, Barnard, Abraham, Gibbins, Morgan, Scott,	**pike**	, Kelly, Chandler.
ABL 650	I met him returning from one of the Penn ponds with the largest	**pike**	of the year swinging by his side and a look of sheer elation on his face.
AL3 19	The fishing habits of the angler banned from the British	**pike**	Championship for allegedly using photographs of the same fish to claim three separate and spectacular catches have landed him in trouble again.
ALU 269	Faulkner's local History mentions trout,	**pike**	, carp, roach, dace, perch, chub, barbel, smelt, flounder, shad, lamprey and eel all being caught in the river off Chelsea and also records nine salmon weighing 171 lbs.
ASN 2834	No Elsie they found nothing in Loch Craig but a huge	**pike**	.
ASW 863	Towards the close of the twelfth century the	**pike**	was used to counter cavalry charges, and remained in use in various forms until as late as the eighteenth century.

Fig. 1. BNC samples containing the noun *pike*

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* [5].

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. They then hit the "set senses" button and the word sketch is updated, 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 [6]), and it is usually immediately evident to the lexicographer which sense is salient, so the task is not unduly taxing. It is not necessary for the lexicographer 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". Then 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 word.

1.2 WASPBENCH and Machine Translation (MT)

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. A description of the same system in MT terms follows.

MT has long needed many rules 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

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

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 [7].

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. These three communities have 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 alternative products to compare it with.
- On the lexicography front: human analysis of meaning is decidedly 'craft' (or even 'art') rather than 'science'. WASPBENCH is, we hope, 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 [8–10]. 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 rules as a distinct function in the way that we do. (Most existing MT systems take a highly domain-based view of word sense ambiguity. In this approach, once the domain is identified, it is assumed that ambiguity goes away, since words tend to only have one meaning and one translation within a given domain. The domain is usually fixed by the user selecting which lexicon they want to use. This strategy has taken MT a long way. It has effectively been the only option available for commercial MT for most words and language pairs, up until developments such as WASPBENCH. It also serves as a useful corrective to the tendency in the WSD world to take the level of ambiguity displayed in paper dictionaries at face value, rather than taking a serious interest in the concept of domain. While clearly the solution for many ambiguity types, the domain-based view fails

for many cases where words have multiple meanings/translations within a single domain, and is also hard to apply in situations where the user cannot realistically be asked to select the domain, such as web-page translation. For further discussion see [11–14])

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 approaches:

- WASPBENCH as a WSD system, within the SENSEVAL evaluation exercise [15]
- the word sketches have been put to the test within a large scale commercial lexicography project; they were used as the main source of corpus evidence for a word's behaviour in the production of the Macmillan English Dictionary for Advanced Learners [16]; [17]
- three expert reports were commissioned from experienced lexicographers
- one set of experiments (with students at the Centre for Translation Studies, Leeds University[2]) explored the performance of WASPBENCH-based translations in comparison with translations produced by commercial MT systems
- a further set of experiments, with a larger group of subjects, explored the extent to which different individuals, working with the same data, produced consistent results.

It is the last evaluation strategy that we report on here. A report bringing together evidence from all evaluation approaches is in preparation.

The setting

Following a March 2001 workshop designed to set the stage for India-UK collaboration in HLT [18] and interest generated there, the University of Brighton licenced WASPBENCH to Prof. Rajeev Sangal of the Indian Institute for Information Technology (IIIT) Hyderabad. This was the first time WASPBENCH had been used outside its development environment in Brighton, UK. WASPBENCH was installed and was then used in IIIT on a project which is developing an English-Hindi translation system. The goal was this: where an English word[3] had more than one possible Hindi translation, the WASPBENCH provides a computational environment and high-level HLT support for the lexicographer in "telling" the computer when it should be translated the one way, when the other.

In early 2002 we were seeking experimental subjects to evaluate WASPBENCH. We approached IIIT, who were glad to co-operate. We prepared datasets and experimental protocols and sent them to IIIT where the staff, who were already familiar with WASPBENCH, trained a group of students in its use and ran the experiments.

[2] We would like to thank Prof. Tony Hartley for his help in setting this up.

[3] The word would have to be a noun, verb or adjective; WASPBENCH does not address grammatical words or, at the current time, adverbs.

3 Experimental setup

We asked the participants to work with the WASPBENCH to create 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 them to a set of previously unseen test sentences and asked the participants to assess the results.

3.1 The task

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 the word would be translated in 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 or amend the input. After submitting, the word expert is applied to a set of test sentences.

Asssessing 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 certain word in a particular context. There is just no gold standard and thus we cannot evaluate the decisions automatically. Therefore we asked the participants to assess the the word experts' judgements.[4]

The assessment task can best be introduced by looking at a screenshot. In figure 2 we present part of the evaluation screen with the results of applying the word expert made by participant 'one' for the noun *bank* to the set of 45 test sentences. The assesser 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

[4] Similar difficulties were encountered in the Japanese SENSEVAL-2 machine translation task, and a similar strategy was adopted ([19]).

translation. The assesser is asked to judge the correctness of the translation in this particular context in the fourth column. 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, the choice 'other' is offered in the alternatives list and 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.

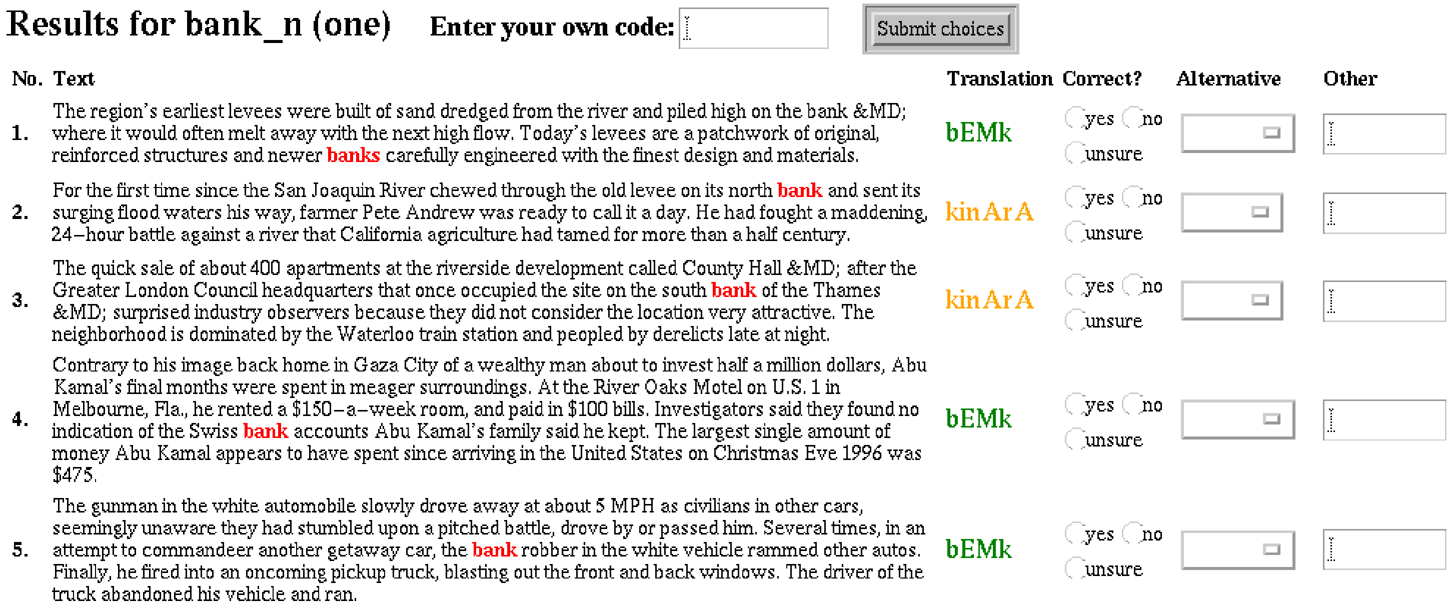
3.2 Instruction and available time

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 four days were allowed for working on the task: about two days for creating word experts and two 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 approximately 30 minutes per word was available. Our first experiment taught us that that was not a reasonable thing to ask. Even though our first experiment showed that the speed at which the subjects created the word experts increased considerably as they became more familiar with the task and the workbench, more time was needed and 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.

3.3 Data

The 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



Results for bank_n (one) Enter your own code: Submit choices

No.	Text	Translation	Correct?	Alternative	Other
1.	The region's earliest levees were built of sand dredged from the river and piled high on the bank &MD; where it would often melt away with the next high flow. Today's levees are a patchwork of original, reinforced structures and newer **banks** carefully engineered with the finest design and materials.	bEMk	○yes ○no ○unsure		
2.	For the first time since the San Joaquin River chewed through the old levee on its north **bank** and sent its surging flood waters his way, farmer Pete Andrew was ready to call it a day. He had fought a maddening, 24–hour battle against a river that California agriculture had tamed for more than a half century.	kinArA	○yes ○no ○unsure		
3.	The quick sale of about 400 apartments at the riverside development called County Hall &MD; after the Greater London Council headquarters that once occupied the site on the south **bank** of the Thames &MD; surprised industry observers because they did not consider the location very attractive. The neighborhood is dominated by the Waterloo train station and peopled by derelicts late at night.	kinArA	○yes ○no ○unsure		
4.	Contrary to his image back home in Gaza City of a wealthy man about to invest half a million dollars, Abu Kamal's final months were spent in meager surroundings. At the River Oaks Motel on U.S. 1 in Melbourne, Fla., he rented a $150–a–week room, and paid in $100 bills. Investigators said they found no indication of the Swiss **bank** accounts Abu Kamal's family said he kept. The largest single amount of money Abu Kamal appears to have spent since arriving in the United States on Christmas Eve 1996 was $475.	bEMk	○yes ○no ○unsure		
5.	The gunman in the white automobile slowly drove away at about 5 MPH as civilians in other cars, seemingly unaware they had stumbled upon a pitched battle, drove by or passed him. Several times, in an attempt to commandeer another getaway car, the **bank** robber in the white vehicle rammed other autos. Finally, he fired into an oncoming pickup truck, blasting out the front and back windows. The driver of the truck abandoned his vehicle and ran.	bEMk	○yes ○no ○unsure		

Fig. 2. Snapshot of the evaluation screen

subtle meaning differences might produce different translations in the target language. It is, of course, possible that, whenever more than participant was working on the same language, they disagreed on the one set of target translations.

The 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.[5] Random samples were taken from the corpus and inspected for suitability. This was done 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 40 or 50 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.

3.4 The participants

A group of eighteen people were involved in the experiment. None of them had a specific lexicography or translation background, but all of them were postgraduates in linguistics or a closely related discipline (e.g. natural language processing). One of our goals for this experiment was to obtain data from several participants on the same words for the same target language. In the Leeds evaluation we worked with several people working on different languages. In order to minimize the effects of personal preferences we wanted to average the results from several (at least five) people working on the same word and target language. Most people worked with Hindi as target language (sixteen in total). Six of them were native speakers, the others were all fluent speakers of Hindi. Two subjects worked on other languages: Russian and Telugu. This was the mother tongue for both of them. All subjects had an excellent command of English, but were not necessarily fluent.

4 Evaluation of the results

4.1 Summary of the data

A total of 370 word experts were produced for the 33 words. This means that an average of 11.2 word experts per word are available. The minimum number

[5] Available from the Linguistic Data Consortium (www.ldc.upenn.edu).

Total (241 word experts)				
	Correct	Incorrect	Unsure	correct
All	6316	4011	485	58%
Nouns	3505	2014	236	61%
Verbs	1839	1238	101	58%
Adjectives	972	759	148	52%
Hindi (214 word experts)				
All	5712	3472	435	59%
Nouns	3179	1786	216	59%
Verbs	1683	1065	91	59%
Adjectives	850	621	128	53%
Hindi by native speakers (103 word experts)				
All	2721	1750	196	58%
Nouns	1608	928	94	61%
Verbs	762	571	56	55%
Adjectives	351	251	46	54%
Russian (22 word experts)				
All	523	430	33	53%
Nouns	326	228	20	57%
Verbs	98	109	2	47%
Adjectives	99	93	11	49%

Fig. 3. Summary of the India evaluation data

of word experts per word was 9 and the maximum 13. As explained below, not all the results of applying the word experts to the test-data could be assessed. The results of a total number of 241 word experts was evaluated. This gives an average of 7.3 per word, with a minimum of 6 evaluated word experts for a word and a maximum of 10[6]. We are planning to evaluate the remaining 129 word experts at a later stage.

In figure 3 a summary of the results is presented. In 58% of the test sentences, the evaluator judged the word expert's prediction to be correct. In 33% the prediction was thought to be incorrect and in the remaining 5% they were not sure.

It is difficult to work out whether these results are good or bad. We would like to establish a "baseline" to compare WASPBENCH performance with.[7] With an average of 4.2 target translations per word (see figure 4) the WASPBENCH

[6] For some of the words, one of the word experts was made for the target language Russian. This means that in a few cases we have a minimum of 5 different evaluated word experts that can be compared.

[7] In our report on the results of the Leeds experiment we can compare with the machine translation results and we can conclude that the WASPBENCH outperforms those results.

performs significantly better then the naive baseline that distributes the possible target translations evenly over the test sentences.A better baseline could arguably be set by assigning the most frequent occurring target translation to every sentence in the test set. However, this cannot be done once for all the participants, but needs to be done for every single word expert, due to the fact that different participants will often give different sense labels/translations for the same concept or take incompatible views of the words ambiguity. As mentioned above, test sentences were not a random sample of corpus instances containing the word, but were a subset of a random sample, chosen manually, to ensure that a range of senses were covered. While this was necessary for experimental design, it complicates the issue of producing a baseline. A single random sample might well have produced 40 instances, all of the same meaning, implying a baseline of 100%, of little use for evaluating WASPBENCH. The opposite position of selecting test instances so that all senses were equally represented was considered, but rejected on the grounds that it was too far removed from the typically Zipfian facts of word frequency distribution. The approach adopted was a compromise.

4.2 Discussion

Considering the fact that the word experts were produced by inexperienced users in a relatively short amount of time (an average of 20.5 word experts in two days), we think that the overall results of the WASPBENCH are promising.

We expected a significantly better result for the nouns. It is often easier to determine the set of target translations for a noun than, for example, for a verb. Verbs often occur in constructions that are translated completely differently in the target language. This intuition is confirmed when compared to the results for the adjectives, but even though nouns do score overall better than verbs, the differences are small.

We did not find evidence for a difference in performance in the word experts between those that were produced by the native speakers of Hindi and by those that were non-native. Both the performance and the time needed for creating them were nearly identical.

Three of the participants volunteered to do the assessment task for their own word expert as well as for someone else's. The data from these three participants assessing their own word experts did not suggest any significant differences.

We expected decreasing success rates with increasing numbers of target translations. Although we do not have the space to give full results for every word, we have selected a few words in figure 5. The results for, for example, the nouns *party* and *policy* versus the noun *line* confirm this intuition. The verbs *move* versus *pray* and the adjectives *flat* versus *funny* are more evidence for this trend.

Some participants reported difficulties with loan words. Even though they experienced problems with particular sense of these words, the performance appeared to be better than average (see the figures for *film* and *charge* in figure 5). The other problematic cases reported were lexical gaps. The two words named explicitly proved to be very problematic. The results for the words *float* and *moody* were among the worst of the set.

Word	**# Meanings**	**# Target translations**	**Word**	**# Meanings**	**# Target translations**
bank	2	2.6	charge	3	4.7
chest	3	2.8	float	3	5.2
coat	3	2.6	move	3	6.3
film	3	2.7	observe	3	3.4
fit	3	4.9	offend	2	4.2
line	6	7.5	post	4	5.7
lot	4	3.6	pray	2	2.4
mass	3	6.4	ring	4	4.6
paper	3	4.3	toast	2	3
party	3	3.1	undermine	2	2.8
policy	3	2.2			
record	3	4.6	bright	4	4
seal	3	3.9	flat	4	7.4
step	2	4.4	free	5	3.8
term	3	5.4	funny	3	3.2
volume	3	4.9	hot	3	3.6
			moody	2	3.3
			strong	4	6.3

Fig. 4. Number of anticipated meanings and (average) number of target translations per word

One of our goals in this particular experiment was to find out how consistent the results are when several people work on the same data. We found that for most words the several word experts gave very similar results on the test data. The fluctuation in the results were strongly correlated with the number of target translations identified by the creator of the word sketch. Whenever the number of target translation identified by the participants was close to the average, the results for that word were close to the average.

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 presentation of the data, missing navigation abilities, buttons and correction facilities and other user-interface issues. We will not go into details here, but will incorporate suggestions into future releases of the workbench.

An important issue (also mentioned in the Leeds evaluation) 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).

Word	Correct	Incorrect	Unsure
film	74%	25%	1%
charge	65%	33%	2%
float	41%	48%	11%
moody	40%	52%	8%
party	72%	25%	3%
line	37%	54%	9%
policy	69%	29%	2%
move	29%	70%	1%
pray	86%	12%	2%
flat	43%	45%	12%
funny	66%	27%	7%

Fig. 5. The results for some individual words

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

The participants also gave feedback on the evaluation task. Some of the issues raised had an impact on the number of word experts they could produce, others could influence the performance of their word experts. The most important remarks were about the assessment task. In the Leeds experiment, most of the subjects were native or near-native speakers of English. There was very little difference in time needed for creating the word experts between the Leeds group and the India group. However, most of the subjects in the Leeds group needed much less time for the assessment task than the India group. We underestimated the fact that for non-native speakers of English this task is much harder. For the native speakers it does not seem to be necessary to read the test sentences thoroughly. It is often enough just to look at the direct context of the ambiguous word to understand what the correct meaning of the word in this sentence is. It is much harder for the non-native speakers. They often want to understand the sentences properly before deciding on the correctness of the suggested translation. The lengthy test sentences (see the screenshot in figure 2) slowed down the progress of the assessment task considerably. As this had not been anticipated, not all the word experts could be evaluated.

As mentioned above, some participants reported that 'loan words' were problematic in cross evaluation cases. Although words like the noun *film* and the verb *charge* are used in the English form in Hindi for some of the senses, other senses are translated with a Hindi word. There are differences for several Indian languages with respect to which senses are translated. Some of the subjects experienced problems with assessing the results of a word expert made by someone whose mother tongue is different from the assesser.

6 Conclusions and further research

The evaluation experiment presented in this paper has given us a rich source of data. In this paper we have looked at this data from a few angles. The experiments taught us that the WASPBENCH is capable of organizing data in such a way that the users are able to create word experts in a consistent way.

Certain words are clearly causing problems. Identifying them beforehand, so special care can be taken for those, might improve the overall performance considerably. The case of lexical gaps, for example, needs extra attention. When words are significantly more ambiguous, it is probably worthwhile spending more time on creating the word expert. But it is probably not only the creator of the word expert who can improve on these words. It might be necessary to combine evidence from multiple sources, to decide which sense (or target translation) is the most suitable in a certain context. WASPBENCH currently uses a 'winner takes all' strategy for deciding which rule is applied for disambiguation; Sometimes an approach which accumulates evidence from different rules is better [20].

A nice aspect of the data we have gathered in this experiment is the reusability of the data. Modifications of the WSD engine in the WASPBENCH in the future can be evaluated by testing again with this data (although we are aware of the danger of overspecialising a system for a particular set of test data).

The feedback of the participants in both this experiment and in the Leeds experiment are very valuable for future developments of the WASPBENCH. Taking the workbench out of the laboratory and into the field is an important step in the development of a tool.

Acknowledgements

This work was supported by the UK EPSRC, under the WASPS project, grant GR/M54971. We would like to thank Amba Kulkarni and her colleagues at the IIIT for organising and supervising the experiments.

References

1. Kilgarriff, A.: The hard parts of lexicography. International Journal of Lexicography **11** (1998) 51–54
2. Sinclair, J.M., ed.: Looking Up: An Account of the COBUILD Project in Lexical Computing. Collins, London (1987)
3. COBUILD: The Collins COBUILD English Language Dictionary. *Edited by John McH. Sinclair* et al., London. (1987)
4. Gale, W., Church, K., Yarowsky, D.: A method for disambiguating word senses in a large corpus. Computers and the Humanities **26** (1993) 415–539

5. Kilgarriff, A., Tugwell, D.: Word sketch: Extraction and display of significant collocations for lexicography. In: Proc. Collocations workshop, ACL 2001, Toulouse, France (2001) 32–38
6. Yarowsky, D.: One sense per collocation. In: Proc. ARPA Human Language Technology Workshop, Princeton (1993)
7. Kilgarriff, A., Tugwell, D.: Wasp-bench: an MT lexicographer's workstation supporting state-of-the-art lexical disambiguation. In: Proc. MT Summit VIII, Santiago de Compostela, Spain (2001) 187–190
8. Fillmore, C.J.: Two dictionaries. International Journal of Lexicography **2** (1989) 57–83
9. Atkins, B.T.S., Levin, B.: Admitting impediments. In Zernik, U., ed.: Lexical Acquisition: Exploiting On-Line Resources to Build a Lexicon. Lawrence Erlbaum, Hillsdale, New Jersey (1991) 233–262
10. Atkins, B.T.S.: Then and now: Competence and performance in 35 years of lexicography. In: 10th EURALEX, Proceedings, Copenhagen (2002) 1–28
11. Edmonds, P., Kilgarriff, A.: Introduction to the special issue on evaluating word sense disambiguation systems. Natural Language Engineering (2002, forthcoming)
12. Magnini, B., Strapparava, C., Pezzulo, G., Gliozzo, A.: Using domain information for wsd. In: Proc. SENSEVAL-2: Second International Workshop on Evaluating WSD Systems, Toulouse, ACL (2001) 111–114
13. Vossen, P.: Extending, trimming and fusing wordnet for technical documents. In: Proceedings of the NAACL 2001 Workshop on WordNet and Other Lexical Resources, Pittsburgh (2001) http://www.seas.smu.edu/ rada/mwnw/papers/WNW-NAACL-105.pdf.
14. Buitelaar, P., Sacaleanu, B.: Ranking and selecting synsets by domain relevance. In: Proceedings of the NAACL 2001 Workshop on WordNet and Other Lexical Resources, Pittsburgh (2001)
15. Tugwell, D., Kilgarriff, A.: WASPBENCH: a lexicographic tool supporting wsd. In: Proc. SENSEVAL-2: Second International Workshop on Evaluating WSD Systems, Toulouse, ACL (2001) 151–154
16. Rundell, M., ed.: Macmillan English Dictionary for Advanced Learners. Macmillan, London (2002)
17. Kilgarriff, A., Rundell, M.: Lexical profiling software and its lexicographical applications - a case study. In: EURALEX 02, Copenhagen (2002)
18. McEnery, T., ed.: Language Engineering for South Asian Languages: workshop proceedings, University of Lancaster (2001) http://www.emille.lancs.ac.uk/lesal.htm.
19. Kurohashi, S.: Senseval-2 japanese translation task. In: Proceedings of Second International Workshop of Evaluating Word Sense Disambiguation Systems (SENSEVAL-2), Toulouse (2001) 37–40
20. Yarowsky, D., Florian, R.: Evaluating sense disambiguation performance across diverse parameter spaces. Journal of Natural Language Engineering (2002) In press Special Issue on Evaluating Word Sense Disambiguation Systems.