Lexical Substitution as a Task for WSD Evaluation

Diana McCarthy
School of Cognitive and Computing Sciences,
University of Sussex,
Falmer,
Brighton,
BN1 9QH. UK

Abstract

The paper is intended to promote discussion on potential application oriented methodologies for the next SENSEVAL, and to suggest one possibility for an application-oriented task. Whilst the traditional gold-standard sense-tagging methodology has proved useful in the last two SENSEVALs, the problem of coming up with a satisfactory sense inventory remains, as the choice of the inventory typically creates biases in favour of particular systems. Coupled with the problems that these biases impose is the issue that the inventory, and level of granularity, should reflect the purpose of the application for which the WSD component is intended. In the last SENSEVAL, the Japanese Translation task was a step in this direction. In this paper we will outline some possibilities for a lexical substitution task, and argue that such a task is relevant to several applications to which a WSD system might be applied and would permit participants to select their own inventory.

1 Introduction

Two related problems have plagued everyone involved in SENSEVAL, that is which sense inventory do we use and what is the motivation for this? The methodology for the SENSEVAL evaluation exercise to date, with the exception of the SENSEVAL-2 Japanese translation task, has been for participating systems to identify the correct sense tag, from a given inventory, for a given target in context. The systems' responses are then evaluated against the sense-tags supplied in a gold-standard which has been produced by human annotators. Whilst the model of obtaining a sense-tagged gold-standard using human-taggers is useful for discovering a system's capability on a given inventory, biases are created by the choice of the inventory (Kilgarriff, 1998). Furthermore, the inventory may not be appropriate for some of the tasks that the participant systems are designed for. Whilst we do not advocate that SENSEVAL abandons this method of evaluation altogether, the need for systems to be evaluated on relevant application tasks, and more flexibility in terms of the inventory used are clearly required.

SENSEVAL has gone a long way towards its goal of creating a level playing field for the evaluation of WSD systems and the creation of valuable resources for the WSD community. The first SENSEVAL kicked off with lexical sample tasks for English, Italian and French. At SENSEVAL-2 evaluation was increased to three types of task over twelve languages. The three types of task were all-words, lexical sample and translation.

An all-words task was created for four of the twelve languages, the lexical sample for nine languages and the translation task only for Japanese. The all-words task required sense tagging of almost all content words in a sample of continuous texts. The lexical sample was on a selected sample of content words. Training data was not always available, for example for

the all words English task and the Italian lexical sample task [1]. Both the lexical sample and all words task shared the gold-standard sense tagging methodology. The task of WSD is isolated and no application specific machinery is required. The Dutch all words task additionally required participants to extract the inventory which was implicit in the sense tags supplied with the training data. The problem that faces the organisers of a task using this methodology is which inventory should be used, and why.

The Japanese translation task was different in that participation focussed on a specific and relevant application. The participants were given a mapping between the Japanese test items and possible English equivalents. Thus the inventory was selected for a purpose for which WSD is necessary, and participants were asked to provide the translations, given this inventory. The senses that the systems had to choose between were selected because they corresponded to distinct translations from Japanese into English. Sense distinctions with different forms in different languages are more likely to correspond to coarser grained distinctions than sense distinctions in a monolingual dictionary (Resnik and Yarowsky, 2000). Disambiguating these differences has relevance for an application. In contrast, much WSD work has gone on without a goal-driven inventory (Kilgarriff, 1997). Whilst it is of interest to investigate if a machine can distinguish senses that lexicographers have thought up, and also whether human annotators can themselves distinguish these senses, the questions still remains which inventory should be used to discover these distinctions, since inventories vary considerably in the distinctions that they make, and why. If we are not interested in which human distinctions WSD systems can make, but we want to apply our systems to some task then we need some rationale behind the inventory that we use.

The inventories that we choose bias systems. One rationale for the use of Hector (Atkins, 1993) for the English task at the original SENSEVAL was that because no systems were using it, everyone would be penalised (Kilgarriff and Rosenzweig, 2000). A mapping between WordNet and Hector was provided so that participants with systems built around WordNet could share a common mapping. However, the mapping itself then creates biases (Agirre et al., 2000; Carroll and McCarthy, 2000). In SENSEVAL-2, WordNet was used as the inventory for both English tasks, because WordNet is widely available, and Hector much less so. WordNet does not however link related senses and there is no clear level of coarse sense distinctions. This was probably one of the reasons that the results for the English SENSEVAL-2 were noticeably lower than those of the original SENSEVAL.

There seems to be no easy way out of biasing systems, unless we do not prescribe a specific inventory. There are systems which disambiguate according to the inventories which they themselves detect in the data (Schütze, 1998). An evaluation task which did not assume a particular inventory might be easier for such systems than the traditional gold-standard tasks. If no inventory is supplied then the gold-standard sense-tagging methodology becomes hard if not impossible to adopt because of the work involved for a human annotator recognising a wide variety of sense tags. SENSEVAL participants were at liberty to merge senses from the given inventory, and supply more than one sense tag, but they would then be penalised if the human annotators didn't make the same decisions.

What is needed is at least one more application oriented task, where the inventory used should be as relevant to the overall goal as possible. Whilst we could still use the traditional methodology for investigating which sense-tags from a given inventory are easier to tag, we could also see the performance of systems in the context of an application, and explore how the choice of inventory affects this performance. The translation methodology is a good one, in that WSD has been shown to improve performance (Brown et al., 1991). It would be fair to allow systems to compete on more than one application platform, since some systems are designed for purposes other than machine trans-

[1] Although systems could use freely available sense tagged data, such as SemCor.

lation, such as information retrieval (Schütze, 1998). We would particularly like to explore an application which is flexible enough to allow systems to respond using their own inventory, though evaluating the system responses will then be harder.

Another criteria we have for an application-oriented task is that it should allow participants to focus on the WSD task within the overall application, rather than trying to evaluate too many different subtasks at the same time. For the machine translation task, effort was focussed just on translation of words, thus focusing on the WSD task. One way to evaluate WSD in the context of an application would be where a core system was provided which required a WSD module and could read in the responses from participants' systems. The difference in performance with and without the WSD module could then be measured. However, if it is the case that organisers are not able to provide such a system for participants, then it would certainly not be fair to expect participants to produce the architecture themselves.

So what applications are there, in addition to the machine translation task, that would be suitable for objective evaluation without requiring a whole host of other activities? WSD systems have been suggested as being of use for many applications, although aside from machine translation the benefits have yet to be proved.

2 Possible Applications for WSD

What applications can WSD be applied to and which of these would be suitable for an evaluation exercise? Kilgarriff (1997) identified machine translation, lexicography and information retrieval as being applications which can benefit from WSD. Machine translation is the clearest case where WSD is required for an NLP application. Whilst WSD has been shown to benefit information retrieval (Schütze, 1998), the effect is diminished when longer queries are supplied. The combination of words in the query go a long way to ranking documents in order of sense relevance. Lexicography is not an NLP task, but one where sense-tagged data is useful to lexicographers developing dictionaries. The traditional sense tagging methodology of SENSEVAL, with the production of gold-standard resources clearly feeds into this. Presumably it is also useful for lexicographers to be aware where inter-tagger agreement is low, and where and why WSD systems fall down on the same items.

There are other applications which might benefit from WSD. One application which we think would benefit from WSD is text simplification, which is just one variation of a more general lexical substitution task where a word is replaced by another word for a particular application. For example, in PSET (Devlin and Tait, 1998) the main enterprise was to simplify newspaper text for aphasic adults. One subtask was to substitute words with more familiar, or more frequent, synonyms, for example *learn* might be used in place of *memorize*. It makes clear sense to substitute with synonyms from the appropriate sense rather than from any of the synonyms for the target word form. Thus if we are going to simplify *scheme* in the sentence:

A recent government study singled out the scheme as an example to others

one would want to use *strategy* rather than *dodge* as a replacement for *scheme* in this context.

Having identified the sense of the target word, lexical choice is required for generating the replacement as a given sense may have more than one synonym. Typically one would want words which were near synonyms, or less specific replacements. For text simplification there is a particular agenda as regards the requirements of the word used for replacement. The word should be easier to understand for the target audience.

There are other possible uses for lexical substitution. Text summarisation might benefit from a module which identifies the sense of a word and is able to suggest a number of alternative expressions (Banko et al., 2000). Information retrieval may also benefit from lexical substitution in term expansion, assuming the user is interested in documents without the exact key words supplied.

3 Lexical Substitution as an Application Oriented WSD Task

The text substitution task is rather like the translation task, in that there is a mapping between the target form and one or more sets of substitutions. Cases with only one set will arise for monosemous words. Whilst we could constrain the systems to select from a given inventory of sets, there is a more appealing option of letting them generate the sets themselves. This would create additional work for participants, although they could use man-made inventories such as WordNet. Crucially, it would allow systems which produce their own inventories into the arena, and permit users of predefined inventories to merge senses and create coarser grained inventories where it makes sense to do so. Possibilities for the inventory and how we might actually evaluate the system answers are outlined in the next two subsections.

3.1 The Inventory

For a lexical substitution task, we can choose whether we restrict users to a given inventory, or allow them to select their own. Whilst not specifying a predefined inventory makes human annotation much harder we contend that this would reduce bias and encourage participation from users who build their own classifications. Systems that deal in semantic space (Schütze, 1998) could then participate, as well as systems that are committed to a given inventory.

The substitutions will depend on the type of inventory used. As regards man-made inventories, a thesaurus like WordNet lends itself more easily to a task like this than a dictionary like Hector, since it is organised on the basis of semantic relationships, rather than alphabetically, though useful replacements might well be found in dictionary definitions. WordNet provides synonyms, or near synonyms together in synsets, and these synsets are related to other synsets with relationships such as hyponymy. For verbs and nouns one could use synonyms, or words in hypernym classes. For adjectives one could use the "similar to" relation. For many word senses in WordNet there are no synonyms supplied.

	hypernyms		
word	sense1	sense2	sense3
cascade	*arrange*	*descend*	-
discordant	*discrepant*	*dissonant*	-
church	*service*	*building*	*faith*

Table 1: Hypernyms for different senses of target words

Instead hypernyms might provide adequate replacements. For example, table 1 shows alternative hypernyms for some target words from the SENSEVAL-2 data, depending on the sense in which they are used. [2]

3.2 Evaluating the Responses

So how do we evaluate responses to a lexical substitution task? We would need to provide either a gold-standard of possible substitutions or a task-based evaluation. Possibilities for task-based evaluation might be readability of the output, as determined by human judges, or performance on an information retrieval task. We discuss possibilities for producing a gold-standard and leave open for discussion the question of whether it would be appropriate, and indeed possible to supply an application for evaluation.

Providing a gold-standard for evaluating a wide variety of possible lexical replacements is harder than specifying an exact match criterion for senses, at least in terms of scoring. However, for the human annotators at least it may be that selecting replacement words might be easier than identifying senses. Many issues remain for evaluation:

1. could annotators be asked to supply a gold-standard and training data in advance of the evaluation?
2. do we have a binary response to whether something is a suitable replacement, or can we rank lexical choice?
3. should participants choose more than one replacement, and should these be seen disjunctively, or conjunctively?

[2] We use the prerelease 1.7 version of WordNet used for SENSEVAL-2 in this paper.

Once we remove the restriction to use a given inventory then we need to allow for a wide-range of responses. We could provide annotators with potential replacements and get them to select from these in advance, permitting them to add their own, however we should expect to have to check systems' responses which not are in the set of potential replacements after submission.

The criteria that are used to judge responses is critical. It may well be that a good substitute in terms of the senses of the word may not fit syntactically. For example, one sense of *service*, in WordNet is listed with hypernyms *assist help assistance aid* and a gloss: *an act of help or assistance; "he did them a service"*. Whilst the hypernyms might bear a strong semantic resemblance to this sense of *service*, they would be syntactically anomalous in a phrase such as *did them a service*. There are also collocational constraints to deal with, *strong* would be a better substitute for *potent* than *powerful* in the phrase *potent tea*. We could either require our systems to meet syntactic and collocational criteria, or instruct our annotators to ignore these constraints if we want to put as little non WSD burden on participants. If we opt for a task-based evaluation, rather than a gold-standard, then the participants would need to consider these constraints. The choices for lexical substitution within a full application would depend on the goals of that application, for example whether the text is to be summarized, simplified or expanded.

For a gold-standard evaluation the criteria we might ask the annotators to use is semantic cohesion of the target and the replacement. We could avoid grading responses and count any valid substitution as correct where valid substitutions are those which are semantically close, not antonyms such as *cold* in place of *hot*, not more specific than the target e.g. not *alsation* for *dog*, and not too general so as to be ambiguous e.g. *thing* for *chicken*.

There are several issues as to whether and how participants should be allowed to supply multiple choices for a given test item. In the previous SENSEVALs, participants were allowed to supply more than one sense tag per item, and allowed to specify a probability distribution with their choices or accept a default uniform distribution. If more than one sense tag was correct, then the scoring was performed for all tags, since there was no way for participants to specify whether the answers were expected to be conjuncts (both apply) or disjuncts (one of the disjuncts applies, but the system cannot discriminate between them). We could likewise allow more than one replacement. Again the issue arises, are these answers to be seen as conjuncts or disjuncts? It makes perfect sense that a system might want to supply both. We advocate that the ambiguity be resolved by asking participants to supply brackets around conjuncts, and attach a probability score to the disjuncts. Thus, for example *door* was marked in the gold-standard as having two senses in the English all words task.

```
d00 d00.s04.t15  door%1:06:00:: door%1:06:01::
```

in

The parishioners of St. Michael and All Angels stop to chat at the church door,...

In SENSEVAL-2 A participant might have responded:

```
d00 d00.s04.t15 door%1:06:00::
or
d00 d00.s04.t15 door%1:06:01::
or
d00 d00.s04.t15 door%1:06:00:: door%1:06:01::
or
d00 d00.s04.t15 door%1:06:00:: 0.6 door%1:06:01:: 0.4
```

We suggest that participants could bracket the choices to indicate a conjunct or leave the choices unbracketed as before to indicate a disjunct, with attached probability distribution if the default is not required.

Lexical substitution responses using related senses in the WordNet inventory pictured in figure 1 might then look like :

```
d00 d00.s04.t15 barrier doorway
or
d00 d00.s04.t15 barrier 0.6 doorway  0.4
or
d00 d00.s04.t15 (barrier doorway)
or
d00 d00.s04.t15 (barrier doorway threshold room_access  entrance)
```

If more than one replacement for a given target is offered, we should divide the credit for the test item (i), by the number of replacements offered. The probability distribution supplied with disjuncts could be used to weight this, or the default uniform distribution:

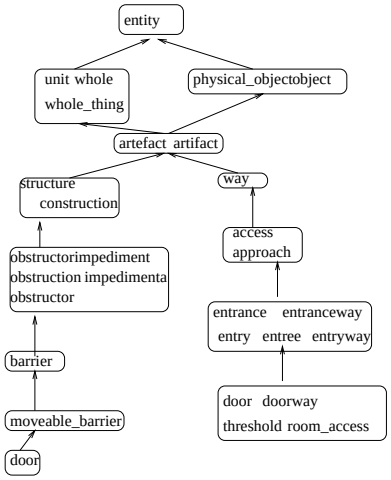

Figure 1: Hypernyms of the first two senses of door in WordNet

$$score_i = \sum_{d \in disjuncts_i} \frac{valid\ responses_{i,d}}{responses_{i,d}} \times p(d) \quad (1)$$

We need to decide whether to accept multi-words such as *room_access*. Removing them would reduce the replacement possibilities from a resource such as WordNet, but make evaluation easier. Systems devising their own entries do not typically include multi-words, and so would perhaps be at an unfair disadvantage if these were included.

4 A Proposal

From the discussion in the previous section we suggest the following for a lexical substitution exercise.

- Participation
  1. we ask for semantically similar replacements, allowing for slightly more general words as replacements, since near synonyms are not always available
  2. we allow participants to replace target words with any words, but not multi-words, from any inventory
  3. we ask participants to group alternative replacements in brackets which are to be seen as conjuncts (all apply)
  4. we ask participants to supply disjuncts in separate brackets with attached probability distribution if required (system cannot discriminate these, but assumes a human would)
- Annotation
  1. in advance we collect possible replacements from any available inventories, and allow annotators to supply their own
  2. for words in the target data annotators supply as large a set of possible replacements, to be read conjunctively (they are all valid replacements), and as few sets as possible to be read disjunctively (it is not clear which of the alternative replacement sets is correct in this case)
  3. we check a sample of (or all) system responses not in the annotators' collective repository to ensure that these were not appropriate.
- Scoring
  1. Each test item is given a score of 1
  2. scoring is performed as in equation 1 above

5 Conclusions

In this paper we have argued that an application oriented WSD task is required in addition to the sense-tagging methodology. If we were able to provide participants with an application requiring a WSD module, this would enable us to prove the utility of the WSD, but without an available plug and play application a lot of effort would be spent on other subtasks related to the application, and SENSEVAL would become quite a different enterprise. We suggest a lexical substitution task since it is relevant to a number of applications such as term expansion in information retrieval and text simplification. A chief advantage of an application-oriented task such as this is that not only would it permit systems to demonstrate performance in the context of an appropriate inventory but could allow users to select their own inventory.

6 Further Work Needed

Further work is required to determine if a lexical substitution task, such as the one we propose, is feasible. We need to investigate if human annotators could be asked to provide in advance a gold-standard of possible substitutes, given access to appropriate inventories. We need to ascertain how time consuming and costly this process would be. Is it more or less time-consuming than sense tagging? Is selecting from a pool of suggested replacements a valid possibility? Can human annotators readily ignore syntactic and collocational constraints in favour of semantic resemblance or would it be appropriate to require participants to adhere to such constraints? How likely is it that a good replacement would not be thought of in advance, thereby requiring a thorough check on non-valid responses afterwards and could we supply training data for a task such as this?

We acknowledge that our examples have been from English. Work is required to see whether and how this approach might work in other languages, and whether man-made resources exist which might be appropriate to the task.

One large by-product of the past SENSEVAL exercises has been the production of sense labelled data sets for further evaluation. If a good deal of time is to be set up into creating data resources supplied with valid lexical replacements it would be good to know if and how such a resource might be used by the WSD community, other researchers in NLP and as a resource for lexicographers.

7 Acknowledgments

This work was supported by the EPSRC-funded RASP project (grant GR/N36493), and the EU 5th Framework project MEANING – Developing Multilingual Web-scale Language Technologies (IST-2001-34460).

References

Eneko Agirre, German Rigau, and J. Atserias. 2000. Combining supervised and unsupervised lexical knowledge methods for word sense disambiguation. *Computers and the Humanities. Senseval Special Issue*, 34(1–2):103–108.

Sue Atkins. 1993. Tools for computer-aided lexicography: the Hector project. In *Papers in Computational Lexicography: COMPLEX 93*, Budapest.

Michele Banko, Vibhu Mittal, and Michael Witbrock. 2000. Headline generation based on statistical translation. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*, pages 318–325.

P. Brown, S.A.D. Pietra, V.J.D. Pietra, and R.L. Mercer. 1991. Word-sense disambiguation using statistical methods. In *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics*, pages 264–270.

John Carroll and Diana McCarthy. 2000. Word sense disambiguation using automatically acquired verbal preferences. *Computers and the Humanities. Senseval Special Issue*, 34(1–2):109–114.

Siobhan Devlin and John Tait. 1998. The use of a psycholinguistic database in the simplification of text for aphasic readers. In John Nerbonne, editor, *Linguistic Databases*, volume CSLI Lecture Notes Number 77, pages 161–173. CSLI Publications, Stanford CA.

Adam Kilgarriff and Joseph Rosenzweig. 2000. Framework and results for english SENSEVAL. *Computers and the Humanities. Senseval Special Issue*, 34(1–2):15–48.

Adam Kilgarriff. 1997. What is word sense disambiguation good for? In *Proceedings of Natural Language Processing in the Pacific Rim*, pages 209–214.

Adam Kilgarriff. 1998. Gold standard datasets for evaluating word sense disambiguation programs. *Computer Speech and Language*, 12(3):453–472.

Philip Resnik and David Yarowsky. 2000. Distinguishing systems and distinguishing senses: New evaluation methods for word sense disambiguation. *Natural Language Engineering*, 5(3):113–133.

Hinrich Schütze. 1998. Automatic word sense discrimination. *Computational Linguistics*, 24(1):97–123.