USYD: WSD and Lexical Substitution using the Web1T Corpus

Tobias Hawker
School of Information Technologies
University of Sydney
NSW 2006, Australia
toby@it.usyd.edu.au

Abstract

This paper describes the University of Sydney's WSD and Lexical Substitution systems for SemEval-2007. These systems are principally based on evaluating the substitutability of potential synonyms in the context of the target word. Substitutability is measured using Pointwise Mutual Information as obtained from the Web1T corpus.

The WSD systems are supervised, while the Lexical Substitution system is unsupervised. The lexical sample sub-task also used syntactic category information given from a CCG-based parse to assist in verb disambiguation, while both WSD tasks also make use of more traditional features.

These related systems participated in the Coarse-Grained English All-Words WSD task (task 7), the Lexical Substitution Task (task 10) and the English Lexical Sample WSD sub-task (task 17).

1 Introduction

This paper describes closely related systems that were applied to three tasks of the SemEval-2007 workshop. The unifying characteristic of these systems is that they use the same measure of 'substitutability' for a given word and a surrounding context to perform the tasks. This measure is based on frequencies involving the word and the context from n-gram counts derived from one trillion words of Web text.

These systems participated in the English Coarse-Grained All Words and English Lexical Sample Word Sense Disambiguation (WSD) tasks, and in the Lexical Substitution task.

The Lexical Substitution system relies entirely on the substitutability measure to rank potential synonyms, and only uses manual sense inventories to preferentially select words which have been identified by lexicographers as being synonyms for the original word in some contexts. It does not make use of any machine learning, and is thus unsupervised.

The WSD systems are supervised, using a Support Vector Machine (SVM) to learn from sense-tagged examples of ambiguous words and predict the class of the test instances. Classifiers for both systems use a small number of additional feature types beyond those derived from the n-gram counts, including Bag of Words (BOW) and local context features. A single separate model was trained for each ambiguous lemma.

For verbs in the lexical sample, the classifier also uses the syntactic category assigned to the target verb by a parser as additional information for disambiguation.

2 Background

Algorithms making use of unannotated data for WSD and similar tasks are not particularly new. One strategy which resembles the substitutability technique employed by our systems is relatives-in-context (Martinez et al., 2006), an unsupervised approach which uses a web search engine to find the 'best' match for the current context, according to heuristic criteria. Monosemous relatives (Leacock

et al., 1998) increase the amount of training data for supervised learners by recruiting the contexts of synonyms in unannotated data, with the caveat that those synonyms are not themselves ambiguous. As substantial gold-standard data sets for lexical substitution have not previously been available, the SemEval data presents a promising opportunity to examine the behaviour of our method.

Gomez (2001) argues that the syntactic roles of ambiguous verbs in particular are linked to their semantic class, and thus knowledge about the syntactic function of a verb can provide information to help identify its sense. Syntactic relationships have been used to resolve ambiguity (Lin, 1997) and a reduction of ambiguity has been shown to assist in the acquisition of verb subcategorization frames (Korhonen and Preiss, 2003).

3 The Substitutability Measure

As an example to demonstrate the basic mechanism underlying the measure of substitutability, consider the sentence fragments around the verb *ruled* in:

the court ruled it was clear that

and

a republic ruled by the people

Two possible synonyms, pertaining to different senses for the verb *ruled*, are *found* and *governed*. It is clear that in a sufficiently large quantity of text, the fragments:

the court found it was clear that

and

a republic governed by the people

would be substantially more common than the sequences:

the court governed it was clear that

or

a republic found by the people

and thus *found* should be considered more substitutable in the context of the first fragment, and *governed* in the second.

Church et al. (1994) show that Pointwise Mutual Information (PMI) is a suitable measure to capture the degree to which a given word may substitute for another; we have adopted PMI as the quantified measure of substitutability in the systems used for these tasks.

While previous WSD systems have made use of counts obtained from Internet search engines, for example Martinez et al. (2006), to our knowledge WSD using *corpus* data at the scale of the Web1T resource has not previously been published. Our WSD systems combine our novel PMI-Web1T features and CCG category features with additional features described in the literature. While the Web1T corpus consists only of counts, and thus is somewhat similar to the direct use of counts from Internet search engines, it is also of a known size and thus it is straightforward to determine useful quantities such as PMI, and to exhaustively catalog potential matches as for the lexical substitution task.

3.1 Web1T Corpus

The Web1T corpus (Brants and Franz, 2006) is a dataset consisting of the counts for n-grams obtained from 1 trillion (10^{12}) words of English Web text, subject to a minimum occurrence threshold (200 instances for unigrams, 40 for others). The Web1T corpus contains counts for 1, 2, 3, 4 and 5-grams, and is large enough to present serious processing difficulties: it is 25GB in compressed form.

The systems presented here thus use custom high-performance software to extract only the n-gram counts of interest from the Web1T data, including simple wildcard pattern-matching. The scale of the data rules out attempting to perform arbitrary queries — even though the counts are lexicographically ordered, disk access times and decompression overheads are severe, and case-insensitive queries are not possible. This software will be released for community use. A limitation in the implementation is that the number of tokens that can be matched in a wildcard expression is fixed at one. This limitation precluded the testing of substitutability of multi-word-expressions (MWEs) in the systems applied to the SemEval tasks.

4 Task 7: Coarse Grained-English All-Words WSD

The system for Coarse-Grained All-Words WSD was supervised, but only attempted classification for a subset of words. These words were chosen according to the amount of sense-tagged training data available, drawn from SemCor (Miller et al., 1993) and the SenseEval-3 lexical sample (Mihalcea et al.,

2004) task. Features were extracted and a classifier trained for each ambiguous content word that was either present in the SenseEval-3 lexical sample, or occurred at least 100 times in SemCor. These criteria yielded classifiers for 183 words.

For ambiguous words without sufficient available training data, the first sense baseline (determined from WordNet version 2.1 (Fellbaum, 1998)) was assigned to every instance. No manual augmentation of the information from WordNet was performed. For those words where models were being trained, the sense clusterings provided by the task organisers were used to completely unify all senses belonging to a cluster, thus attempting disambiguation at the level of the coarse senses. As the system does not attempt to disambiguate words not selected for modeling, the exclusion of the most frequent sense (MFS) baseline would be likely to have a severe adverse impact on this type of supervised approach. Extension of the substitutability measure to directly select a sense related to good substitutes, similar to the approach outlined in Lin (1997) would be one possible approach to resolve this consistently.

The classifier used for the system was an SVM (libsvm) (Chang and Lin, 2001). Linear kernels were used, as previous experiments using similar features with other data sets for WSD had shown that these kernels outperformed radial basis function and polynomial kernels; this disparity became particularly pronounced with larger number of features compared to training instances, and with the combination of different feature types. The number of unique features for each lemma was, on average, more than an order of magnitude higher than the number of training instances: 4475 compared to 289.

The features used to train the selected lemmas included the substitutability measurement, all content words within 3 sentences of the target, and immediate local context features. These are detailed below. There is no in-principle reason why CCG category features used for the Lexical Sample task (see Section 6.2) could not also be used for verbs in the all-words task. Sentences containing target verbs could have been selectively parsed and redundancy among disambiguated running text in SemCor exploited. However, the system architecture was not amenable to small modifications along these lines, and time constraints prevented implementation before the close of the evaluation period. The impact of this additional useful feature would be an interesting subject for future study.

4.1 Features

4.1.1 Substitutability: Pointwise Mutual Information

To transform the notion of substitutability into a set of features suitable for WSD, a set of potential substitute words was chosen for each modeled lemma. These words were taken from WordNet 2.1 (Fellbaum, 1998). For nouns, all synonyms, immediate hypernyms and immediate hyponyms for all senses were included. For verbs, synonyms for all senses were used. The selection of potential substitutes was stricter for verbs as the number of synonyms tended to be greater than for nouns, and these criteria kept the number of substitutes manageable.

A sliding window was used to maximise the information extracted from the Web1T corpus. All windows at all sizes covered by the Web1T corpus that included the target word were used to determine the overall substitutability.

The counts of interest for determining the PMI for a single substitute in a single window position include the unigram frequency of the substitute itself the overall frequency of the context, irrespective of the word in the target position; and crucially, the frequency of the substitute in that context. For a given substitute and context, an overall PMI is determined as a single quantity, obtained by simply adding the PMI together from each window position of each size covered in the data:

$$\begin{aligned} PMI &= \sum_{n=2}^{5}\sum_{i=1}^{n} \log_2 \frac{\text{observation}_{n,i}}{\text{expectation}_{n,i}} \\ &= \sum_{n=2}^{5}\sum_{i=1}^{n} \log_2 \frac{\#(\text{sub} + \text{context}_{n,i})}{p(\text{sub}) \cdot p(\text{context}_{n,i}) \cdot N_n} \end{aligned}$$

Here n represents the window size (varying from 2 to 5), i is the position within the window, and N_n indicates the total number of n-grams present in the corpus for a given value of n. Following Church et al. (1994) the Maximum Likelihood Estimate (MLE) is used for both probabilities in the

denominator. $p(\text{substitute})$ is estimated from the unigram frequency of the substitute word, while $p(\text{context})$ is derived from the counts of the context ignoring the token in the target location.

Features were also created that harnessed the idea that it is not only the level of substitutability for each candidate word that is useful, but also that it may be informative to recognise that some words are better substitutes than others. This information was captured by adding additional features consisting of the pairwise differences between PMI values for all candidate substitute words. To further draw the differing levels of substitutability into relief, features representing the rank of each pair's PMI difference were also included.

Finally, each of the above feature types yields real-valued features. Before being used in classification, these features were converted to binary features using supervised Entropy-Based Discretisation (Fayyad and Irani, 1993). This process characterises the partition selection as a message coding problem: the class labels in the training data are a message to be encoded given that the value of the feature is known for each instance, and the process aims to minimise the length of that message. This is achieved by recursively bifurcating each feature's values at the partition point that would result in the shortest message. Useful boundaries are those where knowing which side of the partition the feature value falls on can be used to reduce the message length beyond any increase required to specify the partition. The algorithm terminates when the existing partitions cannot be divided further and still satisfy this condition. If this occurs when attempting to find the first partition, the feature is dropped altogether.

4.1.2 Bag of Words in broad context

Bag of words (BOW) features were introduced to represent the presence or absence of almost all words within a window of three sentences of the target word. A small stop list (approximately 50 words) was used to remove common closed-class words such as prepositions and conjunctions. The words were lemmatised before being transformed into features, and were not weighted for their distance from the target word. No attribute subset selection was performed on the BOW features.

Document	Attempted	Precision	Recall	F1
d001	0.986	0.625	0.617	0.621
d002	0.958	0.598	0.573	0.585
d003	0.948	0.610	0.578	0.593
d004	0.929	0.606	0.563	0.583
d005	0.965	0.471	0.455	0.463
Total	0.953	0.588	0.560	0.574

Table 1: Coarse-Grained WSD results

4.1.3 Local Context Features

The sentence containing the target word was tagged for Part of Speech (POS) using the POS tagger in the C&C parser tools. For four tokens either side of the target lemma, features were formed from the displacement of the token concatenated with:

- The POS tag
- The lemmatised word
- The POS and lemma together

Also included were features combining the above information for pairs of tokens before, after, and either side of the target word. Finally, a feature representing the POS tag of the target word was added, providing such information as number and tense.

The portion of the context used to form these features is identical with that used to determine substitutability of potential synonyms using the Web1T-based features. Combining the abstract substitutability features with features that use the particular tokens in the local context helps to maximise the utility of information present near the target word by approaching it from multiple perspectives.

4.2 Results and Discussion

The results of the system are shown in Table 1

The first-sense baseline achieves scores of 0.788 for precision, recall and F1, and thus outperforms our system for all documents.

Unfortunately we are currently unable to explain this relatively poor performance. It is possible that an error of a similar nature to the one which affected the initial results for the lexical sample system (see Section 6.3) was also present in this system, although we have not been unable to identify such a

problem. As of the time of writing, the gold standard labels for the test data was not yet available, and thus pinpointing the reason for the disappointing performance is difficult.

5 Task 10: English Lexical Substitution

5.1 Methodology

As for the WSD systems, the Lexical Substitution system concentrated on words whose occurrence in local contexts similar to that of the target was more frequent than expected in the Web1T corpus.

Aside from preferring sets of potential synonyms obtained from lexical resources, the system is entirely unsupervised. Consequently, no sense-annotated corpus resources were used.

The lexical resources used were WordNet version 2.1 (Fellbaum, 1998) and the Macquarie Thesaurus (Bernard, 1985), a pre-defined, manually constructed Thesaurus. The only information used from these resources was a list of potential synonyms for all listed senses that matched the target word's part-of-speech. These synonyms were used to preferentially choose potential substitutes obtained from the corpus data, as described below. The union of potential synonyms from both resources was used, although MWEs were not included due to limitations with the corpus. Although these lexical resources were not augmented, the system was capable of producing substitutes not present in these resources by using high-scoring words found in the corpus. The ordering of synonyms in these resources was not used directly, nor was their association with particular senses.

The PMI for potential substitutes that occurred in the target position of each local context window was determined using the Web1T corpus, as for coarse WSD above. The strategy differed slightly from the supervised process employed for WSD however, in that rather than testing a fixed set of potential substitutes, every word that occurred in the correct location in a matching context was considered as a substitute. This introduced an additional computational burden which restricted the set of n-grams used to 4 and 5 grams. In particular, this is because the set of words occurring in the target position grew prohibitively large for 2 and 3 grams.

As for WSD, the PMI for each potential substitute was combined by summing the individual PMIs over all locations and size of n-gram where it occurred. This sum was used to rank the substitutes. After the production of the ranked list, the set of synonyms obtained from the lexical resources was used for preferential selection. Substitutes in the ranked list that also occurred in the synonym pool were chosen first. The exact manner of the preferential selection differed for the two evaluation measures the system participated in.

	P	R	Mode P	Mode R
all	11.23	10.88	18.22	17.64
Further Analysis				
NMWT	11.68	11.34	18.46	17.90
NMWS	12.48	12.10	19.25	18.63
RAND	11.47	11.01	19.14	18.35
MAN	10.95	10.73	17.20	16.84

Table 2: BEST results

For the BEST measure, the highest PMI-ranked substitute that occurred in the synonym pool was given as the only substitute. If no substitutes from the synonym pool were present in the ranked list, the top three substitutes from the list were given.

For the out-of-ten (OOT) measure, the ten highest-ranked substitutes that were in the synonym pool were given. If fewer than 10 substitutes were present in the list, the remaining best ranked substitutes not in the synonym pool were used to make up the ten answers.

As with the Coarse-Grained All Word WSD, limitations in the current implementation of the Web1T processing software meant that it was not possible to examine MWEs, and there was thus no provision to detect or handle MWEs in the system. For this reason, the MW measure was not produced by the system.

5.2 Results and Discussion

The results for the BEST and OOT measures are given in tables 2 and 3 respectively. While the results for the other tasks are reported as a decimal fraction of 1, the results here are percentage scores, in line with the results given by the task organisers.

Notably, recall is always lower than precision. If no substitutes were found to have finite PMI at any

	P	R	Mode P	Mode R
all	36.07	34.96	43.66	42.28
Further Analysis				
NMWT	37.62	36.17	44.71	43.35
NMWS	40.13	38.89	46.25	44.77
RAND	35.67	34.26	42.90	41.13
MAN	36.52	35.78	44.50	43.58

Table 3: OOT results

position, no substitute was rendered by the system. This meant a small number of examples in the submitted system had no answer provided. The system's design meant that no attempt was made to provide any answer when counts were zero for all Web1T queries. No answer was given for around 3% of the evaluation set. As the query retrieval software was limited to single word substitutions, this should be expected to occur for MWEs more frequently than for single word substitutions. The results for both BEST and OOT confirms this, showing that the system's performance is uniformly better when MWEs are excluded.

As a consequence of the properties of the Web1T corpus, the system chooses substitutes on the basis of information that is derived from at most four words either side of the target word. It is thus encouraging that it is able to outperform the baselines on each evaluation measure.

Interestingly, for the BEST evaluation the performance on the randomly selected (RAND) examples outperforms that on the manually selected (MAN) examples. For the OOT evaluation the situation is reversed. This could indicate that, depending on the motivation for the manual selections, the system is not particularly well-suited to selecting an obvious singular substitution, but is quite capable of ranking reasonably acceptable ones near the top of the list.

6 Task 17: Coarse Grained English Lexical Sample sub-task

6.1 Approach

The Lexical Sample system used features identical to those described for the Coarse-Grained All-Words task, with the addition of the CCG supertag feature, discussed below. Labeled data used for training the classifier models in this system consisted of only the instances in the training data supplied for the task, although Web1T corpus was of course used to provide extensive information in the form of features for those instances. As for the All-Words system, an individual SVM model was trained using linear kernels for each lemma being disambiguated. The contextual BOW features were not selected from within a window as for the All-Words system; instead the entire context provided in the training and test data was used.

Unlike the other systems, the Lexical Sample system produced a prediction for every instance in the test data, as the MWE limitation of the Web1T processing software did not present an impediment.

6.2 CCG Verb Categories

The Lexical sample data was parsed using the Clark and Curran CCG parser (Clark and Curran, 2004). Existing tagging and parsing models, derived from CCGBank are included with the parser package, and were used without adjustment. Gold-standard parses available for the source data were not used.

The syntactic combination category ("supertags") assigned to target verbs by the parser were used as features. This category label encodes information about the types of the other sentential components used when building a parse. A forward slash indicates that the current token requires a component of the specified type to the right; a backwards slash requires one to the left. The C&C parser includes a supertagger, but this supertagger assigns multiple labels with varying degrees of confidence, and when the parse is performed, the supertag labels are subject to revision in determining the most likely parse. The feature used for the Lexical Sample system uses the final, parser-determined supertag.

As an example, consider the occurrence of the verb *find* in the following two fragments where it has different senses:

managers did not find out about questionable billing

and

or new revenues are found by Congress

The first fragment has a (simplified) supertag of (S\NP)/PP, while the second is playing a different grammatical role, and hence has a different supertag: S\NP. While these supertags are generally not exclusively associated with a single sense in particular, their distribution is sufficiently distinct

over different senses that features derived from them are informative for the WSD task. To form features, the system uses the supertags obtained from the parser as binary features, with a slight simplification: by removing distinctions between the argument types of the main S component, generalisation is facilitated among instances of verbs which differ slightly on a local level but combine with other parts of the sentence similarly.

6.3 Results and Discussion

Unfortunately, the component of the lexical sample system responsible for assigning identifiers for evaluation contained a systematic error, resulting in a mismatch between the predictions of the system and the correct labels as used in evaluation. The system assumed that for each lemma in the test set, the instances in the test data file would have lexicographically ascending identifiers, and matched predictions to identifiers using this assumption. This was not the case in the task data, and yielded a result for the submission that severely underestimated the performance of the system. We calculated a baseline of 0.788 for the Lexical Sample sub-task, using the Most Frequent Sense for each lemma in the training data. The result for the systems initial submission was 0.743 (precision, recall, accuracy and F1 are all identical, as the system provides an answer for every instance).

However, as the mismatch is systematic, and only occurred after the classifier had made its predictions, it was possible to correct almost all of the alignment by post-processing the erroneous answer file. By holding the order of predictions constant, but lexicographically sorting instance identifiers within each lemma, predictions were re-matched with their intended identifiers. Using the test labels provided by the task organisers, the accuracy of the system after repairing the mismatch was 0.891.

As the parser does not have 100% coverage, the parse of the test sentence did not succeed in every instance. This in turn caused some supertag features to be misaligned with other feature types before the error was rectified. This meant that a small fraction of instances were given predictions in the submitted data that differed from those produced by the corrected system. When the already-trained models were used to re-predict the classes of the correctly aligned test instances, a further small improvement to a result of 0.893 was achieved.

It is interesting that the results (after correcting the misaligned identifiers) for the patched system is approaching the Inter Tagger Agreement (ITA) level reported for OntoNotes sense tags by the task organisers – 90%. This could be seen as an encouraging outcome of the movement towards coarser-grained sense inventories for the WSD tasks, it is difficult for automated systems to agree with humans more often than they agree with each other.

7 Conclusion

Substantially similar information in the form of a PMI-based substitutability measure from the Web1T corpus was used in all USYD systems. That this information yielded positive results in different semantic-ambiguity related tasks, both supervised and unsupervised, demonstrates the usefulness of the data at the scale of the Web1T corpus, and there are still many more approaches to using this resource for semantic processing that could be explored.

The systems demonstrated outstanding performance on the Lexical Sample WSD task – nearly at the level of the reported ITA. Good unsupervised performance above the baseline was also achieved on the Lexical Substitution task, and uncharacteristically poor performance was seen for the Coarse-Grained all words task. It will be interesting to compare the performance of the All-Words system to that of the Lexical Sample system when the labels for the test data are released, and potentially improve the performance of that system.

8 Acknowledgements

Many thanks to Jon Patrick, James Curran, and Matthew Honnibal for their invaluable assistance, insights and advice.

References

J. R. L. Bernard, editor. 1985. *The Macquarie Thesaurus*. The Macquarie Library, Sydney.

Thorsten Brants and Alex Franz. 2006. Web 1T 5-gram corpus version 1.1. Technical report, Google Research.

Chih-Chung Chang and Chih-Jen Lin. 2001. *LIBSVM: A Library for Support Vector Machines*.

Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

Kenneth Ward Church, Willam Gale, Patrick Hanks, Donald Hindle, and Rosamund Moon. 1994. Lexical substitutability. In B. T. S. Atkins and A. Zampolli, editors, *Computational Approaches to the Lexicon*, pages 153–177. Oxford University Press.

Stephen Clark and James R. Curran. 2004. Parsing the WSJ using CCG and log-linear models. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, pages 104–111. Barcelona, Spain.

Usama M. Fayyad and Keki. B. Irani. 1993. Multi-interval discretization of continuous-valued attributes for classification learning. In *Proceedings of the Thirteenth International Joint Conference on Artificial Intelligence*, pages 1022–1029. Chambery, France.

Christiane Fellbaum, editor. 1998. *Wordnet: An Electronic Lexical Database*. MIT Press.

Fernando Gomez. 2001. An algorithm for aspects of semantic interpretation using an enhanced wordnet. In *Proceedings of NAACL-2001*, pages 1–8.

Anna Korhonen and Judita Preiss. 2003. Improving subcategorization acquisition using word sense disambiguation. In *ACL '03: Proceedings of the 41st Annual Meeting on Association for Computational Linguistics*, pages 48–55.

Claudia Leacock, Martin Chodorow, and George A. Miller. 1998. Using corpus statistics and WordNet relations for sense identification. *Computational Linguistics*, 24:147–165.

Dekang Lin. 1997. Using syntactic dependency as local context to resolve word sense ambiguity. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics*.

David Martinez, Eneko Agirre, and Xinglong Wang. 2006. Word relatives in context for word sense disambiguation. In *Proceedings of the 2006 2006 Australasian Language Technology Workshop (ALTW 2006)*, pages 42–50.

Rada Mihalcea, Timothy Chklovski, and Adam Kilgarriff. 2004. The senseval-3 english lexical sample task. In Rada Mihalcea and Phil Edmonds, editors, *Senseval-3: Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*, pages 25–28. Association for Computational Linguistics.

George. A. Miller, Claudia. Leacock, Tengi Randee, and Ross Bunker. 1993. A semantic concordance. In *Proceedings of the 3rd DARPA Workshop on Human Language Technology*, pages 303–308.