SENSEVAL-3: Third International Workshop on the Evaluation of Systems
for the Semantic Analysis of Text, Barcelona, Spain, July 2004
Association for Computational Linguistics

# The University of Maryland SENSEVAL-3 System Descriptions

**Clara Cabezas, Indrajit Bhattacharya, and Philip Resnik**
University of Maryland, College Park, MD 20742 USA
clarac@umiacs.umd.edu, indrajit@cs.umd.edu, resnik@umd.edu

## Abstract

For SENSEVAL-3, the University of Maryland (UMD) team focused on two primary issues: the portability of sense disambigation across languages, and the exploitation of real-world bilingual text as a resource for unsupervised sense tagging. We validated the portability of our supervised disambiguation approach by applying it in seven tasks (English, Basque, Catalan, Chinese, Romanian, Spanish, and "multilingual" lexical samples), and we experimented with a new unsupervised algorithm for sense modeling using parallel corpora.

## 1 Supervised Sense Tagging for Lexical Samples

### 1.1 Tagging Framework

For the English, Basque, Catalan, Chinese, Romanian, Spanish, and "multilingual" lexical samples, we employed the UMD-SST system developed for SENSEVAL-2 (Cabezas et al., 2001); we refer the reader to that paper for a detailed system description. Briefly, UMD-SST takes a supervised learning approach, treating each word in a task's vocabulary as an independent problem of classification into that word's sense inventory. Each training and test item is represented as a weighted feature vector, with dimensions corresponding to properties of the context. As in SENSEVAL-2, our system supported the following kinds of features:

- Local context. For each $i$ = 1, 2, and 3, and for each word $w$ in the vocabulary, there is a feature $L_i : w$ representing the presence of word $w$ at a distance of $i$ words to the left of the word being disambigated; there is a corresponding set of features $R_i : w$ for the local context to the right of the word.
- Wide context. Each word $v$ in the training set vocabulary has a corresponding feature indicating its presence. For SENSEVAL-3, wide context features were taken from the entire training or test instance. In other settings, one might make further distinctions, e.g. between words in the same paragraph and words in the document.

We also experimented with the following additional kinds of features for English:

- Grammatical context. We use a syntactic dependency parser (Lin, 1998) to produce, for each word to be disambiguated, features identifying relevant syntactic relationships in the sentence where it occurs. For example, in the sentence *The U.S. government announced a new visa waiver policy*, the word *government* would have syntactic features like DET:THE, MOD:U.S., and SUBJ-OF:ANNOUNCED.

- Expanded context. In information retrieval, we and other researchers have found that it can be useful to expand the representation of a document to include informative words from similar documents (Levow et al., 2001). In a similar spirit, we create a set of expanded-context features $ext : e$ by (a) treating the WSD context as a bag of words, (b) issuing it as a query to a standard information retrieval system that has indexed a large collection of documents, and (c) including the non-stopword vocabulary of the top $n$ documents returned. So, for example, in a context containing the sentence *The U.S. government announced a new visa waiver policy*, the query might retrieve news articles like "US to Extend Fingerprinting to Europeans, Japanese" (Bloomberg.com, April 2, 2004), leading to the addition of features like EXT:EUROPEAN, EXT:JAPANESE, EXT:FINGERPRINTING EXT:VISITORS, EXT:TOURISM, and so forth.

| Lexical Sample | Coarse (prec/rec) | Fine (prec/rec) |
|---|---|---|
| UMD-SST | 0.643/0.643 | 0.568/0.568 |
| UMD-SST-gram | 0.600/0.600 | 0.576/0.576 |
| UMD-SST-docexp | 0.541/0.542 | 0.516/0.491 |

Table 1: UMD-SST variations on the SENSEVAL-2 English lexical sample task

As described by Cabezas et al. (2001), we have adopted the framework of support vector machines (SVMs) in order to perform supervised classification. Because we used a version of SVM learning designed for binary classification tasks, rather than the multi-way classification needed for disambiguating among $N$ senses, we constructed a family of SVM classifiers for each word $w$ — one for each of the word's $N_w$ senses. All positive training examples for a sense $s_i$ of $w$ were treated as negative training examples for all the other senses $s_j$, $j \neq i$.

Table 1 shows the performance of our approach on the English lexical sample task from the previous SENSEVAL exercise (SENSEVAL-2), including the basic system (UMD-SST), the basic system with grammatical features added (UMD-SST-gram), and the basic system with document expansion features added (UMD-SST-docexp). (We have not done a run with both sets of features added.) The results showed a possible potential benefit for using grammatical features, in the fine-grained scoring. However, we deemed the benefit too small to rely upon, and submitted our official SENSEVAL-3 runs using UMD-SST without the grammatical or document-expansion features.

### 1.2 SENSEVAL-3 Lexical Sample Tasks

For SENSEVAL-3, the modularity of our system made it very easy to participate in the many lexical sample tasks, including the multilingual lexical sample, where the "sense inventory" consisted of vocabulary items from a second language.[1] Indeed, we participated in several tasks without having anyone on the team who could read the language. (Whether or not this was a good idea remains to be seen.) For Basque, Catalan, and Spanish, we used the lemmatized word forms provided by the task organizers; for the other languages, including English, we used only simple tokenization.

Table 2 shows the UMD-SST system's official

[1]Data format problems prevented us from participating in the Italian lexical sample task.

| Lexical Sample | Precision (%) | Recall (%) |
|---|---|---|
| Basque | 65.6 | 58.7 |
| Catalan | 81.5 | 80.3 |
| Chinese | 51.3 | 51.2 |
| English | 66.0 | 66.0 |
| Romanian | 70.7 | 70.7 |
| Spanish | 82.5 | 82.5 |
| Multilingual | 58.8 | 58.8 |

Table 2: UMD-SST results (fine-grained) on SENSEVAL-3 lexical sample tasks

| Lexical Sample | Coarse (prec/rec) | Fine (prec/rec) |
|---|---|---|
| UMD-SST | 0.709/0.709 | 0.660/0.660 |
| UMD-SST-gram | 0.703/0.703 | 0.655/0.655 |
| UMD-SST-docexp | 0.691/0.680 | 0.637/0.627 |

Table 3: UMD-SST variations on the SENSEVAL-3 English lexical sample task

SENSEVAL-3 performance on the lexical sample runs in which we participated, using fine-grained scores.

In unofficial runs, we also experimented with the grammatical and document-expansion features. Table 3 shows the results, which indicate that on this task the additional features did not help and may have hurt performance slightly. Although we have not yet reached any firm conclusions, we conjecture that value potentially added by these features may have been offset by the expansion in the size of the feature space; in future work we plan to explore feature selection and alternative learning frameworks.

## 2 Unsupervised Sense Tagging using Bilingual Text

### 2.1 Probabilistic Sense Model

For the past several years, the University of Maryland group has been exploring unsupervised approaches to word sense disambiguation that take advantage of parallel corpora (Diab and Resnik, 2002; Diab, 2003). Recently, Bhattacharya et al. (2004) (in a UMD/Montreal collaboration) have developed a variation on this bilingual approach that is inspired by the central insight of Diab's work, but recasts it in a probabilistic framework. A generative model, it is a variant of the graphical model of Bengio and Kermorvant (2003), which groups semantically related words from the two languages into "senses"; translations are generated by probabilis-

tically choosing a sense and then words from the sense.

Briefly, the model of Bhattacharya et al. uses probabilistic analysis and independence assumptions: it assumes that senses and words have certain occurrence probabilities and that the choice of the word can be made independently once the sense has been decided. Here interaction between different words arising from the same sense comes into play, even if the words are not related through translations, and this interdependence of the senses through common words plays a role in sense disambiguation.

The model takes as its starting point the idea of a "translation pair" — a pair of words $e$ and $f$ that are aligned in two sentences (here "English" and "non-English") that are translations of each other. For example, in the English-Spanish sentence pair *Me gusta la ciudad/I like the city*, one would find the translation pairs $(I, me)$, $(like, gusta)$, $(the, la)$, and $(city, ciudad)$.[2] Those familiar with statistical machine translation (MT) models will note that a translation pair is equivalent to a link in a word-level alignment, and in fact we obtain translation pairs from sentence-aligned parallel text by training a statistical MT model (using GIZA++, (Och and Ney, 2003)) and using the word-level alignments that result.

The probabilistic sense model makes the assumption that the English word $w_e$ and the non-English word $w_f$ in a translation pair share the same precise sense, or, in other words, that the set of sense labels for the words in the two languages is the same and may be collapsed into one set of senses that is responsible for both English and non-English words. Thus the one latent variable in the model is the sense label $T$ generating both words, represented by variables $W_e$ and $W_f$. The model also makes the assumption that words in both languages are conditionally independent given the sense label. The generative parameters $\theta$ for the model are the prior probability $P(t)$ of each sense $t$ and the conditional probabilities $P(w_e|t)$ and $P(w_f|t)$ of each word $w_e$ and $w_f$ in the two languages given the sense. The generation of a translation pair by this model may be viewed as a two-step process that first selects a sense according to the priors on the senses, and then selects a word from each language using the conditional probabilities for that sense. This may be imagined as a factoring of the joint distribution: $P(W_e, W_f, T) = P(T)P(W_e|T)P(W_f|T)$.

Given WordNet as a sense inventory, the set of English senses is as defined in the WordNet database. Since the model assumes the sense labels for the two languages are the same, it must use the same WordNet labels for the non-English words as well. Rather than considering all possible words in the non-English vocabulary for each WordNet sense $t$, we permit an association between non-English word $w_f$ and WordNet sense $t$ if $w_f$ is the translation of any English word $w_e$ in $t$'s synonym set. We use the popular EM algorithm (Dempster et al., 1977) to estimate the model's parameters from a set of translation pairs derived from a parallel corpus.

## 3 Using the Model for WSD

Like Diab's system, the sense model of Bhattacharya et al. requires a parallel corpus in order to estimate its parameters, and as currently implemented it can only assign sense tags to words in parallel text. In order to perform WSD experiments on English test items, therefore, two steps are necessary: obtaining translation pairs from an English-F parallel corpus in order to create a model, and translating test items from English into F in order to obtain word-level alignments that can be used as the basis for disambiguation.

In order to accomplish these steps, Bhattacharya et al. (2004) used the pseudo-translation approach of Diab and Resnik (2002): they created the model using an English-Spanish parallel corpus constructed by using Systran to translate a large collection of English text, and they obtained parallel Spanish text for the test items in the same fashion. On the nouns (only) in the SENSEVAL-2 English all-words task, they obtained precision and recall of 0.624 and 0.616, respectively, improving on Diab's precision and recall of 0.618 and 0.572.[3]

Our goal for SENSEVAL-3 was to investigate the use of human-translated text, rather than pseudo-translated text, in creating the model. To that end, we used three sources of sentence-aligned parallel text:

- Spanish-English: A set of 107,222 sentence pairs sampled from modern Bible translations,

[2] Speakers of both Spanish and English will observe that translation pairs may well include words that are not exactly translations of each other.

[3] Their paper includes a refinement of the probabilistic model that improves performance further, to precision and recall of 0.672 and 0.651.

| Language pair | Precision (%) | Recall (%) |
|---|---|---|
| English-Chinese | 0.445 | 0.445 |
| English-Spanish | 0.444 | 0.444 |
| English-French | 0.445 | 0.445 |

Table 4: Unsupervised probabilistic model results (fine-grained) on the SENSEVAL-2 English all-words task

United Nations Proceedings, and newswire translations from FBIS (the Foreign Broadcast Information Service).

- French-English: a set of 1,008,591 sentence pairs from the Europarl corpus (Koehn, 2003)
- Chinese-English: a set of 440,223 sentence pairs from FBIS.

In order to tag new test sentences, we used machine translation from English test items into each of Spanish, French, and Chinese. We used Systran for Spanish and French, and for Chinese we used an implementation of the alignment template framework for statistical MT (Kumar and Byrne, 2003). Once having obtained the translations for test sentences, we used GIZA++ to create word-level alignments within which translation pairs could be identified. We used the probabilistic model *only for WSD of nouns*, where nouns were identified using an automatic part-of-speech tagger. For other parts of speech, we used the first-listed WordNet sense.

Time limitations prevented us from completing SENSEVAL-3 runs in time for this writing. Table 4 shows the performance of the system on the SENSEVAL-2 English all-words task. This performance level places the approach in the middle group of performers on this task at the time of SENSEVAL-2 (with scores roughly in the range of 0.44-0.46), as contrasted to the top group (with scores in the range of 0.56-0.69). In the near future, we hope to repeat the experiment with the SENSEVAL-3 English all-words and lexical sample test data, and also to explore evidence combination from multiple language pairs.

## 4 Conclusions

Our precision and recall in the SENSEVAL-3 lexical sample tasks was in all cases very close to the median reported by the task organizer, thus demonstrating an ability to obtain credible performance with a simple, robust approach. Additionally, we explored the use of additional features, and we experimented with applying a new unsupervised probabilistic model using human-translated rather than pseudo-translated parallel text, with equivocal results for the various extensions beyond the basic system. In the future we plan to focus our attention more heavily on the learning paradigm and probabilistic modeling, with the particular aim of more effectively exploiting local and document-level context for both sense disambiguation and lexical selection.

## Acknowledgements

The work described in this paper was supported in part by ONR MURI Contract FCPO.810548265, National Science Foundation grant EIA0130422, and Department of Defense contract RD-02-5700. The authors gratefully acknowledge the assistance of Mona Diab, Aaron Elkiss, Adam Lopez, and Grazia Russo-Lassner in data preparation details, and the kind help of Shankar Kumar and Yonggang Deng in using their framework for statistical machine translation.

## References

Yoshua Bengio and Christopher Kermorvant. 2003. Extracting hidden sense probabilities from bitexts. Technical report, TR 1231, Departement d'informatique et recherche operationnelle, Universite de Montreal.

Indrajit Bhattacharya, Lise Getoor, and Yoshua Bengio. 2004. Unsupervised sense disambiguation using bilingual probabilistic models. In *Meeting of the Association for Computational Linguistics*.

Clara Cabezas, Philip Resnik, and Jessica Stevens. 2001. Supervised sense tagging using support vector machines. In *Proceedings of the Second International Workshop on Evaluating Word Sense Disambiguation Systems (SENSEVAL-2)*, Toulouse, France, July.

A.P. Dempster, N.M. Laird, and D.B. Rubin. 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, B 39:1–38.

Mona Diab and Philip Resnik. 2002. An unsupervised method for word sense tagging using parallel corpora. In *40th Anniversary Meeting of the Association for Computational Linguistics (ACL-02)*, Philadelphia, July.

Mona Diab. 2003. *Word Sense Disambiguation*

*within a Multilingual Framework*. Ph.D. thesis, University of Maryland.

Philipp Koehn. 2003. Europarl: A multilingual corpus for evaluation of machine translation. unpublished manuscript.

S. Kumar and W. Byrne. 2003. A weighted finite state transducer implementation of the alignment template model for statistical machine translation. In *Proceedings of HLT-NAACL*, Edmonton, Canada, May.

Gina Levow, Douglas Oard, and Philip Resnik. 2001. Rapidly retargetable interactive translingual retrieval. In *Human Language Technology Conference (HLT-2001)*, San Diego, CA, March.

Dekang Lin. 1998. Dependency-based evaluation of MINIPAR. In *Workshop on the Evaluation of Parsing Systems*, Granada, Spain, May.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.