The Johns Hopkins SENSEVAL2 System Descriptions

David Yarowsky, Silviu Cucerzan, Radu Florian,
Charles Schafer and Richard Wicentowski
{yarowsky,silviu,rflorian,cschafer,richardw}@cs.jhu.edu

Department of Computer Science
Johns Hopkins University
Baltimore, Maryland, 21218, USA

Abstract

This article describes the Johns Hopkins University (JHU) sense disambiguation systems that participated in seven SENSEVAL2 tasks: four supervised lexical choice systems (Basque, English, Spanish, Swedish), one unsupervised lexical choice system (Italian) and two supervised all-words systems (Czech, Estonian). The common core supervised system utilizes voting-based classifier combination over several diverse systems, including decision lists (Yarowsky, 2000), a cosine-based vector model and two Bayesian classifiers. The classifiers employed a rich set of features, including words, lemmas and part-of-speech informatino modeled in several syntactic relationships (e.g. verb-object), bag-of-words context and local collocational n-grams. The all-words systems relied heavily on morphological analysis in the two highly inflected languages. The unsupervised Italian system was a hierarchical class model using the Italian WordNet.

1 The Feature Space

The JHU SENSEVAL2 systems utilized a rich feature space based on raw words, lemmas and part-of-speech (POS) tags in a variety of positional relationships to the target word. These positions include traditional bag-of-word context, local bigram and trigram collocations and several syntactic relationships based on predicate-argument structure (described in Section 1.2). Their use is illustrated on a sample English sentence for *train* in Figure 1.

1.1 Part-of-Speech Tagging and Lemmatization

Part-of-speech tagger availability varied across the languages included in this sense-disambiguation system evaluation. Transformation-based taggers (Ngai and Florian, 2001) were trained on standard data for English (Penn Treebank), Swedish (SUC-1 corpus) and Estonian (MultextEast corpus). For Czech, an available POS tagger (Hajič and Hladká, 1998), which includes lemmatization, was used. The remaining languages – Spanish, Italian and Basque – were tagged using an unsupervised tagger (Cucerzan and Yarowsky, 2000). Lemmatization was performed using a combination of supervised and unsupervised methods (Yarowsky and Wicentowski, 2000), and using existing trie-based supervised models for English.

"Many mothers do not even try to toilet **train** their children until the age of 2 years or later ..."

Feature type	Word	POS	Lemma
Context	...	...	...
Context	try	VB	try/V
Context	to	TO	to/T
Context	toilet	NN	toilet/N
Context	train	VBP	train/V
Context	their	DT	their/D
Context	...	...	...
Syntactic (predicate-argument) features			
Object	children	NNS	child/N
Prep	until	IN	until/I
ObjPrep	age	NN	age/N
Ngram collocational features			
-1 bigram	toilet	NN	toilet/N
+1 bigram	their	DT	their/D
-2/-1 trigram	to toilet *	TO-NN	to/T toilet/N *
-1/+1 trigram	to * their	TO-DT	to/T * their/D
+1/+2 trigram	their children	DT-NN	their/D child/N

Figure 1: Example sentence and extracted features

1.2 Syntactic Features

Extracted syntactic relationships in the feature space depended on the keyword's part of speech:

- for verb keywords – the head noun of the verb's object, particle/preposition and object-of-preposition were extracted when available.
- for noun keywords – the headword of any verb-object, subject-verb or noun-noun relationships identified for the keyword.
- for adjective keywords – the head noun modified by the adjective (if identifiable).

These syntactic features were extracted using simple heuristic patterns and regular expressions over the parts-of-speech surrounding the keyword.

2 Supervised Lexical Choice Systems

The supervised JHU systems utilize classifier combination merging the results of five diverse learning models.

2.1 Core Algorithm Design

The lexical choice task can be cast as a classification task: training data is given in the form of a set of word-document pairs $\mathcal{T} = [(w_i, D_{ij}), S_{ij}]_{i,j}$ (S_{ij} being the sense associated with the document D_{ij} of keyword w_i), labeled with the corresponding gold standard class. The goal is to establish the classification of a set of unlabeled word-document pairs $\mathcal{T}' = \{(w_i, D'_{ij})\}_{i,j}$, not previously seen in the training data. The training data $\mathcal{T}$ is used to estimate class probabilities and then the sense classification is made by choosing the class with the maximum a posteriori class probability:

$$S = \arg\max_{S'} P\left(S'|D\right) = \arg\max_{S'} P(S') \cdot P(D|S')$$

The disambiguation models used in our experiments are feature-based models. A feature is a boolean function defined as $f_w : F \times \mathcal{D} \to \{0,1\}$, where F is the entire set of features and $\mathcal{D}$ is the document space. An overview of the exploited feature space was given in Section 1.

2.2 Vector-based Algorithms

Our Bayesian and cosine-based models use a common vector representation, capturing both traditional bag-of-words features and the extended Ngram and predicate-argument features in a single data structure.

In these models, a vector is created for each document in the collection:

$$D_i = (D_{ij})_{j=1,|F|}$$

where F is the entire utilized feature space

$$D_{ij} = \frac{c_{ij}}{N_i} W_j$$

where c_{ij} is the the number of times the feature f_j appears in document D_i, N_i is the number of words in the document D_i and W_j is the weight associated with the feature f_j.

To avoid confusion between the same word in multiple feature roles, feature values are marked with their positional type (e.g. *children_object*, *toilet_L*, and *their_R* as distinct from *children*, *toilet* and *their* in unmarked bag-of-words context).

The basic sense disambiguation algorithm proceeds as follows:

1. Vectors in the training data are assigned to classes based on their classification;
2. For each vector in the test data, the a posteriori class distribution is computed as
$$P(S|D) = \frac{\text{Sim}(D, C_S)}{\sum_{S'} \text{Sim}(D, C_{S'})}$$
where C_S is the centroid corresponding to the sense S and Sim is the similarity measure used by the algorithm (cosine or Bayes).
3. The sample D is labeled with sense S if $S = \arg\max_{S'} P(S'|D)$.

2.2.1 The Cosine-based Model

In this model, traditional cosine similarity is used to compute similarity between a document D and a centroid C. The weight associated with a feature (F_j) is its inverse document frequency $W_j = \log \frac{N}{N_j}$, where N is the total number of documents and N_j is the number of documents containing feature f_j. Function words and POS tags were excluced from the cosine vectors.

2.2.2 The Bayesian Models

In the Bayes model, the Bayes similarity is computed as:

$$\text{Sim}(D_i, S_j) = P(D_i, S_j) = P(S_j)\, P(D_i|S_j)$$

and the following assumption of independence is made:

$$P(D_i|C_S) = \prod_{f_j \in D_i} P(f_j|C_S)$$

The probability distribution $P(f_j|C_S)$ is obtained by smoothing the word relative frequencies in the cluster C_S. Given the lack of independence between the word-based and lemma-based feature spaces, these are utilized in two separate Bayesian models with output combined in Section 2.5.

2.3 Decision Lists

The decision list model we used in our system is a non-hierarchical variant of the method of interpolated decision lists described in Yarowsky (2000). For each feature f_i a smoothed log of likelihood ratio ($\log \frac{P(f_i|S_j)}{P(f_i|\neg S_j)}$) is computed for each sense S_j, with smoothing based on an empirically estimated function of feature type and relative frequency. Candidate features are ordered by this smoothed ratio (putting the best evidence first), and the remaining probabilities are computed via the interpolation of the global and history-conditional probabilities. By utilizing the single strongest-matching evidence in context, non-independent feature spaces combine readily without inflated confidence, and can be mapped to accurate and robust probability estimates as shown in Figure 2.

2.4 Additional Details

The English task differs slightly from the other lexical-choice tasks in that phrasal verbs are explicitly marked in the training and test data. To make reasonable use of this information, when a phrasal verb is marked, only corresponding phrasal senses are considered; conversely when a phrasal

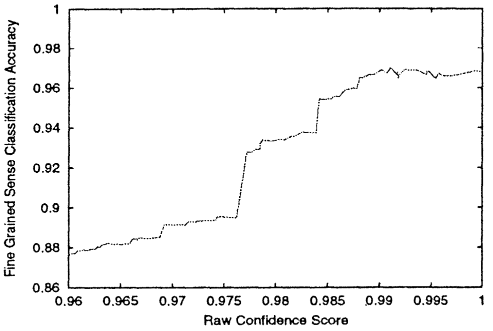

Figure 2: Mapping between raw confidence scores and classification accuracy for English decision lists

verb is not marked, no phrasal senses are considered. Likewise, when a training or test sentence matches a compound noun in the observed sense inventory (e.g. *art_gallery%1:06:00::*) only the matching phrasal sense(s) are considered unless there is at least one non-phrasal sense tagged in the training data for that compound (indicating the potential for both compositional and non-compositional interpretations).

2.5 Classifier Combination

Several classifier combination approaches were investigated in the system development phase. They are outlined below, along with their cross-validated performance on the English lexical-sample training data (in Table 1). In each case four individual classifiers were combined: the cosine model, two Bayes models (one based on words and one based on lemmas[1]), and the decision-list model.

The first two model combination approches simply averages the output of the participating classifiers over each candidate sense tag, in terms of $P(S_j|D_i)$ and $rank(S_j|D_i)$ respectively, with each classifier given an equal vote[2].

The remaining methods assign potentially variable weights to the votes of different classifiers. Interestingly, Equal Weighting of all four classifiers slightly outperforms classifier weighting proportional to each model's aggregate accuracy (Performance-Weighted voting), similar to the technique used for classifier combination in part-of-speech tagging in van Halteren et al. (1998). Finally, it was observed that on sentences where decision lists have high model confidence their accuracy exceeds other classifiers. Thus the most effective approach, based on training-data cross validation, was found to be a very basic Thresholded Model Voting:

- If the decision_list_confidence≥ 0.985 (an empirically selected threshold) then return the output of the decision list;
- Otherwise, each system votes for the sense that is most likely under it and, another vote is obtained from the most probable class yielded by linear interpolation of the 4 classifiers.

This simple top-performing approach was utilized in the evaluation system, and is reasonably close to the performance of an Oracle upper bound for classifier combination (using the output of the single best classifier on each test instance – unknowable in practice).

	Accuracy	
Classifier Combination Method	Fine	Coarse
Model Averaging (excluding decision lists):		
Probability interpolation voting	.657	.728
Rank-averaged voting	.652	.709
Weighted Model Voting (includes decision lists):		
Equal-weighted Model Voting	.667	.736
Performance-Weighted Voting	.655	.724
Thresholded Model Voting	**.676**	**.746**
Oracle Voting (Upper Bound)	.734	.761

Table 1: Comparison of classifier combination methods on English (using 5-fold cross-validation)

3 Supervised All-Words Systems

3.1 Estonian All-words Task

Because of the importance of morphological analysis in a highly inflected language such as Estonian, a lemmatizer based on Yarowsky and Wicentowski (2000) was first applied to all words in the training data (and, at evaluation time, the test data). For each lemma, the $P\,(\text{sense}|\text{lemma})$ distribution was measured on the training data. For all lemmas exhibiting only one sense in the training data, this sense was returned. Likewise, if there was insufficient data for word-specific training (the sum of the minority sense examples for the word in training data was below a threshold) the majority sense in training was returned for all instances of that lemma. In the remaining cases where a lemma had more than one sense in training, with sufficient minority examples to adequately be modeled, the generic JHU lexical sample sense classifier was trained and applied.

3.2 Czech All-words Task

Czech is another example of a highly inflected language. A part-of-speech tagger and lemmatizer kindly provided by Jan Hajič of Charles University (Hajič and Hladká, 1998) was first applied to the data. Consistent with the spirit of evaluating sense disambiguation rather than morphology, the JHU system focused on those words where more than one sense was possible for a root word (e.g.

[1] On training-set cross-validation it was observed that the two systems were uncorrelated enough to make it useful to keep both of them.

[2] Decision lists are not included because they only assign a probability to their selected classifier output but not to lower-ranked candidates.

the -1 and -2 suffixes in the Czech inventory). In these cases, the fine-grained output of the Czech lemmatizer was ignored (in both training and test) and a generic lexical-sample sense classifier was applied to the sense-distinction tags extracted from the lemmatized training data (see Section 2), using the classification models employed in Estonian. Whenever insufficient numbers of minority tagged examples were available for training a word-specific classifier, the majority sense for the POS-level lemma was returned. Likewise, if only one possible sense tag was observed for any POS-level lemma analysis, then this unambiguous sense tag was returned.

4 Unsupervised Italian System

The Italian task stands out from the group of lexical choice tasks because no labelled training was data provided for Italian; instead a subset of the Italian Wordnet was provided. To obtain a sense classifier for Italian, we employed an unsupervised method that used hierarchical class models of the Wordnet relationships among words (synonymy, hypernomy, etc) and a large unannotated corpus of Italian newspaper data to obtain sense centroids.

First, every relationship type in the Italian Wordnet received an initial weight, based on a roughly estimated measure of the relative dissimilarity of two words in that relationship. For instance, the *synonymy* relationship received a small weight (words are semantically "close"), while other relationships (*has_near_synonym*, *causes*, *has_hypernym*) received proportionately larger weights (words are more semantically distant). Starting from the senses S of a target k, the wordnet relationships graph was explored, up to a given distance (two links away), creating "clouds" of similar words, M_S, together with a similarity[3] to the original sense, S.

For each of the words w in M_S, we extracted sentences from the unannotated corpus that contained the word w, and then considered them as being examples of context for the sense S of target k, and assigned them to the centroid C_S (the centroid of the sense S) with a weight corresponding to the similarity between the word w and the sense S (computed using the wordnet graph). After all the documents were distributed, the test documents were also assigned to the most probable cluster, similar to the other lexical choice tasks.

The centroids were then allowed to adjust in a manner similar to k-means clustering. At each step, the centroids were recomputed, after which each document migrated to the closest cluster (i.e. $\arg\max_S P\left(C_S|D\right)$), and the process was repeated. After the process converged, each test document was assigned the label corresponding to the sense centroid it converged into. This process is completely unsupervised, and the only structured resource that was used is the provided Italian Wordnet subset.

Task	Accuracy on Test Data	
	Fine-Grained	Coarse-Grained
Basque	.757	.971
English	.642	.713
Spanish	.712	–
Swedish	.701	1.00
Italian	.353	.423
Czech	.935	–
Estonian	.666	–

Table 2: Official JHU system performance

5 Results

Table 2 lists the official performance of the JHU systems on unseen test data in the final SENSEVAL2 evaluation. Coarse-grained performance scores are based on a hierarchical sense clustering given by the task organizers in 4 of the languages. In the lexical sample tasks, these scores were obtained after correction of a simple bug in the merger of final system output as provided for in the SENSEVAL evaluation protocols.

As illustrated in the comparative performance tables elsewhere in this volume, the JHU systems are consistently very successful across all 7 languages and 3 major system types described here.

References

S. Cucerzan and D. Yarowsky. 2000. Language independent minimally supervised induction of lexical probabilities. In *Proceedings of ACL-2000*, pages 270–277, Hong Kong.

J. Hajič and Hladká. 1998. Tagging inflective languages: Prediction of morphological categories for a rich, structured tagset. In *Proceedings of COLING/ACL-98*, pages 483–490, Montréal.

G. Ngai and R. Florian. 2001. Transformation-based learning in the fast lane. In *Proceedings of NAACL-2001*, pages 40–47, Pittsburgh.

H. van Halteren, J. Zavrel and W. Daelemans. 1998. Improving Data Driven Wordclass Tagging by System Combination In *Proceedings of COLING/ACL-1998*, pages 491–497, Montreal.

D. Yarowsky and R. Wicentowski. 2000. Minimally supervised morphological analysis by multimodal alignment. In *Proceedings of ACL-2000*, pages 207–216, Hong Kong.

D. Yarowsky. 2000. Hierarchical decision lists for word sense disambiguation. *Computers and the Humanities*, 34(2):179–186.

[3] The weight on a path was computed as the sum of the weights on the path, and the similarity was computed as $\mathrm{Sim}\left(w,S\right) = e^{-c(w,S)}$ – large weights result in 0 similarity.