Using semantics of the arguments for predicate sense induction

Anna Rumshisky
Dept. of Computer Science
Brandeis University
Waltham, MA USA
arum@cs.brandeis.edu

Victor A. Grinberg
LightLab Imaging
Westford, MA USA
victor@ccis.neu.edu

Abstract

We present an unsupervised learning system developed to recognize predicate sense distinctions that depend on the semantics of the arguments. We develop a sense-tagged data set for 15 polysemous verbs selected for having such sense distinctions. We use this data set to evaluate the performance of our system in a word sense induction setting. Relative to the baselines, our system outperforms the best system in the SEMEVAL-2007 Task 2 on two out of three measures we use to evaluate the clustering solution.

1 Introduction

Disambiguation of polysemous words in language is usually accomplished by taking into account two aspects of the context: (1) the syntactic frame into which the word is embedded and (2) the semantics of the words with which it forms syntactic dependencies. In word sense disambiguation or induction (WSD/WSI) systems, both aspects are routinely modeled through distributional information, following the idea that semantic similarity between words must be reflected in the similarity of their habitual contexts of occurrence (Harris, 1985; Miller and Charles, 1991).

However, different ambiguities require different kinds of contextual information to be resolved (as anybody who has done any amount of semantic annotation can attest). WSD/WSI systems are not typically designed to treat different kinds of sense distinctions separately. The sense-tagged data used for training and testing of such systems is very labor-intensive to create in general. So it is not surprising that such annotation does not usually attempt to specify the relationship between the annotated senses and the context elements that distinguish between them. Many commonly used sense-tagged corpora, such as SemCor (Landes et al., 1998), PropBank (Palmer et al., 2005), OntoNotes (Hovy et al., 2006), etc., are similar in this respect. This fact makes it very difficult to address the question of what types of distinctions are detected more successfully by a given system.

The senses that are linked to specific syntactic patterns are typically easier for people to distinguish. When main differentiating factor is the semantics of the arguments, detecting such sense distinctions is often less straightforward. Our goal in the present work was to separate out and examine the contribution of the latter factor towards sense differentiation in polysemous verbs. With this goal in mind, we designed a fully unsupervised sense induction system that analyzes how different arguments contribute to disambiguation, looking at each argument position separately.

One of the main problems for the systems that seek to detect or derive senses using distributional similarity is separating occurrence counts for different senses of the word. The work in thesaurus construction or sense induction has often either used the notion of monosemous near-synonyms (Hindle, 1990; Grefenstette, 1994; Pantel and Lin, 2002) or tried to represent contexts, rather than words per se (Schütze, 1998; Mitchell and Lapata, 2008; Erk and Pado, 2008). Our approach uses the idea that words that are similar in a particular context do not need to be distributionally similar overall. Our system identifies the words selectionally similar to a given sense of the target and induces clusters of arguments activating that sense. The former is accomplished by contextualizing the vector representation of such words to the target's context.

To get an idea of how well our system performs, we decided to evaluate it as a standard sense induction system. Such an evaluation is difficult to perform without having a manually constructed resource against which to compare the induced groupings. We developed a sense-tagged data set for 15 polysemous verbs that were selected for having sense distinctions strongly dependent on the seman-

tics of a single argument.

The choice to create a data set for the testing of our system was motivated by two considerations. Firstly, the standard sense-tagged data sets (such as the ones mentioned above) conflate different kinds of contextual information, so they can not be used directly. The data sets that do specify the factors contributing to the disambiguation, such as the data from the FrameNet project (Ruppenhofer et al., 2006), are often incomplete, since only the senses linked to the completed frames are represented in the corpus. Even in case of full coverage, only a very small number of instances actually end up exemplifying the relevant sense distinctions.

Secondly, the sense annotation has often been done on corpora which are not well-balanced, such as the Wall Street Journal data. As a result, the distribution of annotated instances across senses does not reflect the actual frequency distribution between senses, as evidenced by the data sets produced for the last several Senseval competitions.

Initially, we considered using the data set used in the last SEMEVAL competition for the WSD and WSI tasks (Tasks 17 and 2, respectively) (Agirre et al., 2007), which would have allowed us to compare the performance of our system directly to the performance of the sense induction systems that participated in Task 2. However, the average per-verb entropy of the SENSEVAL data set is 0.92, suggesting that the most frequent sense dominated the data set for many of the chosen verbs. In fact, out of the 65 verbs used in the WSI task, 11 verbs had only one sense in the combined test and training data set.

Due to such distribution across the senses, the resulting data set did not seem particularly well suited for testing the discriminatory powers of a sense-induction system. This was the case especially since the evaluation schemes used in Task 2 relied to a large extent on the most frequent sense to assess the systems' performance.[1]

For the above reasons, we do not use the SEMEVAL Task 2 data set for evaluation directly. Instead, we perform an indirect comparison using the data set we developed. We use several measures to evaluate the clustering solution quality, and compare our system's performance to the performance the SEMEVAL Task 2 systems. Relative to the baselines, our system outperforms the best system in the SEMEVAL Task 2 on two out of three measures.

The rest of the paper is organized as follows. Section 2 gives a detailed description of the algorithm. Section 3 describes the data set we developed for testing. Section 4 describes the evaluation we performed. Finally, in the last section, we discuss the results and some possible applications of the system.

[1] For further discussion of this, see Section 4.

2 System Architecture

Our system produces clusters of *selectional equivalents* (Rumshisky, 2008) for each sense of the target word.

A lexical item w_1 is a *selectional equivalent* of lexical item w_2 with respect to grammatical relation R, if in the argument position defined by R, one of the senses of w_1 selects for the same aspect of meaning as one of the senses of w_2. Such selectional equivalence can also apply to two lexical items w_1 an w_2 with distinct relations R_1 and R_2. For example, the verbs *threaten* and *struggle* are selectional equivalents with respect to relations *dobj* (direct object) and *iobj_against* (relation between the governing verb and the head of a prepositional phrase introduced by *against*).

In the discussion below, we will use the term *selector* to refer to the words with which the target word forms syntactic dependencies, regardless of whether the target word is the head or the dependent element in a syntactic relation.

The data set we used for evaluation in this paper consisted of verbs, and the direct object relation was used both for the target and for its selectional equivalents. All the computations were performed over the 100M word British National Corpus (BNC, 2000). We used Robust Accurate Statistical Parsing (RASP) (Briscoe and Carroll, 2002) to extract grammatical relations.

For each target word t and relation R, we execute the steps described in detail below.

2.1 Establishing the set of words to be clustered

Given a corpus, we first identify the set of selectors with which the target word occurs in relation R, and then take the inverse image of that set under the relation R^{-1}. For example, for t = *acquire*, $R = \mathsf{obj}$, the first operation gives the set of nouns that occur in direct object position with *acquire*. The second operation gives us the set of potential selectional equivalents for different senses of *acquire*. Words that occur with a given selector only once, as well as those that occur with less than two of the target's selectors, are discarded.

We sort the resulting words according to how many selectors of the target they co-occur with.[2]

2.2 Identifying reliable selectors

Since every word w in the resulting set occurs with some of the same selectors as the target t, it could potentially be selectionally equivalent to one of the target's senses. We need to identify selectors that for both t and w behave in the following manner: (1) activate the appropriate sense (2) are good disambiguators, i.e. activate only one sense and are not likely to occur with the other senses. Such selectors can be polysemous themselves, yet always occur in the same sense when combining with t or w. If a selector occurs frequently with both t and w, several explanations are possible:

(i) A selector activates the appropriate sense for both t and w, and that sense is fairly frequent for both words:

a. *take on*/*acquire* a new *importance*

(ii) (*Parallel Sense Distinctions*.) If the verbs have more than one selectionally equivalent sense, a selector could activate the wrong pair of senses:

a. *acquire*/*possess* a new *significance* (QUALITY)
b. *acquire*/*possess* a powerful *weapon* (POSSESSION)

(iii) (*Selector Polysemy*.) Different senses of that selector may activate unrelated interpretations for the two verbs:

a. *take on* a greater *share* of the load
b. *acquire* the *shares* of the company

In our model, we make an assumption that the first case is the dominant one, while the other two cases are much more rare. Under such conditions, selectors that are strongly associated with both t and w must be the ones that pick the corresponding sense for each of them.[3]

For every word in the set of candidates for selectional equivalence, we obtain a set of reliable selectors as follows:

1. For each selector s that occurs both with t and w, compute association score $\textit{assoc}_R(s, w)$ and $\textit{assoc}_R(s, t)$.

2. Combine the two association scores using a combiner function $\psi(\textit{assoc}_R(s, w), \textit{assoc}_R(s, t))$ and choose the top-k selectors that maximize it.

Each w is then represented as a k-dimensional vector $\overline{w} = \langle f(s) \rangle$, where $f(s)$ is selector scoring function that determines the value for each selector based on its association scores.

For example, consider the verbs *acquire* and *lack* which are selectionally equivalent with respect to one of the senses of *acquire* (*take on a certain characteristic*). We would like for $\textit{assoc}_{obj}(\textit{importance-n,acquire-v})$ and $\textit{assoc}_{obj}(\textit{importance-n,lack-v})$ to produce a combined value that is high enough to allow *importance* to be identified as a reliable selector.

In this paper, we report results obtained by several configurations of our system, which vary with respect to the association score used, the method used to pick the top-k selectors, and the selector scoring function. We used two types of association scores: pointwise mutual information (mi) and mi normalized by a log factor of the tuple count $\textit{freq}(s, R, w)$.[4] Selector scoring function used either just the association score $\textit{assoc}_R(s, w)$ or the product of selector's association scores with w and t. The resulting configurations are summarized in Table 1.

	$\textit{assoc}_R(s, w)$	$f(s)$
1	$mi(s, Rw) = \log \frac{P(s,R,w)}{P(s)P(R,w)}$	$\textit{assoc}_R(s, w)$
2	$mi(s, Rw) \cdot \log(\textit{freq}(s, R, w))$	$\textit{assoc}_R(s, w)$
3	$mi(s, Rw)$	$\textit{assoc}_R(s, w) \cdot \textit{assoc}_R(s, t)$
4	$mi(s, Rw) \cdot \log(\textit{freq}(s, R, w))$	$\textit{assoc}_R(s, w) \cdot \textit{assoc}_R(s, t)$

Table 1: System configurations. 1 - MI, 2 - MI-FACTOR, 3 - MI-PRODUCT, 4 - MI-FACTOR-PRODUCT.

We use the geometric mean of the two association scores as the combiner function to sort selectors for each w, which induces a sorting order with the sequence of equivalence classes located along the hyperbolic curves. Note that even if the relevant sense is infrequent for the target, but predominant for w, the combined score would still be fairly high.

Clearly, automatically identifying all good disambiguators is not feasible. Our goal is to choose enough selectors correctly so that the selectional equivalents for each sense can be grouped together.

2.3 Producing clusters of selectional equivalents

We use group-average agglomerative clustering to produce clusters of selectional equivalents $C_i = \{w\}$ for each sense of the target word, with each w represented as a k-dimensional vector. For the

[2] For efficiency, we restrict the number of elements to be clustered to 4000, selecting the words that co-occur with a higher number of the target's selectors.

[3] A selector that is strongly associated with both t and w must occur "frequently enough" with each of them. Ideally, the frequency of distribution on the senses for w and t must be taken into account, since the relevant sense may be much more prominent for one word than for the other.

[4] Log factor de-emphasizes the elements with low occurrence counts.

present experiments, we use $k = 15$. Similarity for two elements w_1 and w_2 is computed as the numeric equivalent of set intersect (i.e. sum of minimums) for the top-15 selectors chosen for each of the elements. We do not apply normalization used in the standard numerical extensions of Jaccard and Dice measures.

In our clustering algorithm, we keep a list of selectors for each node in the dendrogram. When two clusters are merged, a union of their selector lists is computed. Each selector is assigned a score that is a weighted average of its scores in the merged clusters (weighted by the number of elements in the cluster).

2.4 Cluster rank

We sort all the nodes in the dendrogram by computing the following score for each node C_i:

$$\textit{rank}(C_i) = \textit{IntraAPS}(C_i) \cdot \log(|C_i|) \cdot \log(\sum_{s \in C_i} f_i(s))$$

where $f_i(s)$ is the score assigned to the selector within cluster C_i, $|C_i|$ is the number of elements in C_i, and *IntraAPS*(C_i) is the average pairwise similarity between the elements of the cluster.

In the present experiments, we used the top 20 clusters that maximized this score.

2.5 Selector-cluster association

Using the obtained clusters, we can estimate which sense of the target a selector is likely to occur with. We compute an association score for each of the chosen clusters C_i and selector s:

$$\textit{assoc}(s, C_i) = \frac{\sum_{w \in C_i} mi(s, Rw)}{|C_i|}$$

where $mi(s, Rw) = \log \frac{P(s,R,w)}{P(s)P(R,w)}$.

The resulting score indicates how likely selector s is to pick the sense of the target associated with C_i. The difference between the scores obtained for different senses with a given selector indicates how strongly that selector tends to prefer one of the senses. If the difference is small, the selector must either equally likely select for either of the senses, or select for both senses at once.

2.6 Using clusters in WSI task

The obtained dendrogram was adapted for use with the standard word sense induction task as follows. Given a set of sentences containing the target word, we extracted the selector for the appropriate grammatical relation. For each selector, we then computed the selector-cluster association score with each of the high-ranking clusters. The sentences containing selector s were tagged with the cluster that had maximum *assoc*(s, C_i). The sentences that were tagged with intersecting clusters (i.e. clusters containing at least some of the same selectional equivalents of the target) were then grouped together.

This method has an obvious handicap relative to the full WSI systems, namely, that we do disambiguation based on only one selector. Consequently, we would expect it to do poorly in situations where a larger context is required for disambiguation.

Here are the clusters obtained for the verbs *conclude* and *grasp* using this method:

verb: conclude
gloss #1: finish
cluster: *begin-v continue-v resume-v prolong-v start-v commence-v open-v initiate-v reopen-v re-open-v*
selectors: negotiation-n, discussion-n, investigation-n, proceedings-n, conversation-n, inquiry-n, talk-n, debate-n, friendship-n, deliberation-n, exploration-n, round-n, argument-n, conquest-n, tour-n, ...
gloss #2: reach an agreement
cluster: *sign-v renegotiate-v agree-v negotiate-v*
selectors: deal-n, pact-n, contract-n, treaty-n, agreement-n, covenant-n, settlement-n, ceasefire-n, arrangement-n, armistice-n, truce-n, ...

verb: grasp
gloss #1: understand, comprehend
cluster: *appreciate-v recognise-v recognize-v realise-v realize-v assess-v demonstrate-v reflect-v illustrate-v explain-v understand-v acknowledge-v underline-v emphasize-v stress-v emphasise-v*
selectors: importance-n, nature-n, significance-n, potential-n, value-n, difference-n, extent-n, fact-n, point-n, complexity-n, implication-n, relationship-n, principle-n, effect-n, meaning-n, situation-n, truth-n, reality-n, concept-n, role-n, aspect-n, necessity-n, idea-n, ...
gloss #2: grab hold of something
cluster: *put-v hold-v thrust-v touch-v raise-v rest-v lift-v rub-v*
selectors: hand-n, arm-n, chin-n, elbow-n, finger-n, shoulder-n head-n, leg-n, receiver-n, knife-n, wrist-n, hair-n, back-n, sword-n, ...

The standard sense-annotated data sets, such as, for example, the ones that have been developed within the framework of the SENSEVAL competitions in the recent years (Preiss and Yarowsky, 2001; Mihalcea and Edmonds, 2004; Agirre et al., 2007), are not suitable for testing in this case.

3 Data set

We tested our system on an independently developed data set that was created using the British National Corpus (BNC), which is more balanced than the more commonly used annotated Wall Street

Journal data. We selected 15 polysemous verbs with sense distinctions that were judged to depend for disambiguation on semantics of the argument in direct object position.

A set of senses was created for each verb using a modification of the Corpus Pattern Analysis (CPA) technique (Pustejovsky et al., 2004). A set of complements was examined in the Sketch Engine, a lexicographic tool that lists significant collocates that co-occur with a given target word in the specified grammatical relation (Kilgarriff et al., 2004). If a clear division was observed between semantically different groups of the collocates in direct object position, the verb was selected. For each group, a separate sense was added to the sense inventory for the target. For example, for the verb *acquire*, a separate sense was added for each of the following sets of direct objects:

(1) *Take on certain characteristics*
shape, meaning, color, form, dimension, reality, significance, identity, appearance, characteristic, flavor
(2) *Purchase or become the owner of property*
land, stock, business, property, wealth, subsidiary, estate, stake

The resulting sense inventory was used to annotate 200 sentences for each verb. The annotators (two undergraduate linguistics majors) were instructed to mark each sentence with the most fitting sense.

In sense annotation, the annotators are frequently forced to choose a sense when no disambiguation can really be performed (Palmer et al., 2007), especially since sense inventories often contain overlapping senses that can be activated simultaneously (Pustejovsky and Boguraev, 1993). Our goal was to create, for each target word, a set of instances where humans had no trouble disambiguating between different senses. Our annotators were therefore instructed to mark a sentence as "N/A" if (1) The sense inventory was missing the relevant sense (2) More than one sense seemed to fit (3) The sense was impossible to determine from the context.[5]

The average inter-annotator agreement (ITA) for our data set was 95%, with disagreements resolved in adjudication. Table 4 shows the following characteristics for each verb: 1) ITA (percentage of instances where the annotators selected the same sense for the verb), 2) MFS (percentage of instances that belong to the most frequent sense), 3) the number of senses and number of instances, and 4) entropy of the distribution of instances across senses. The last row of each column gives the average for the column, weighted by the number of instances for each verb.

To get an idea of how well the verbs in our data set could be disambiguated by a supervised system relying solely on nouns in direct object position, we also ran on our data a Maximum Entropy classifier with 10-fold cross-validation.[6] The obtained accuracy values are shown in Table 4.

4 Evaluation

Our system uses semantics of a single argument to do the disambiguation, but since the verbs in our data have been selected for effectiveness of single-argument semantics disambiguation, it is reasonable to compare the performance of our system to that of the general sense induction systems. One handicap that such evaluation imposes on our system is that since no other context is available, one selector can only be associated with one sense of the target. We found that our system performed well even despite this handicap.

In Task-2 of SEMEVAL-2007 (Agirre and Soroa, 2007), Van Rijsbergen's F-measure was used to rank the participating sense induction systems. Under this metric, the *1cluster1word* baseline (all occurrences of the target word grouped together) outperforms all the clustering systems that competed in the task. This is due to the known problems with this measure (Meila, 2003).

A number of other metrics have been proposed in the literature to evaluate the quality of a particular clustering solution against a gold standard (Amigó et al., 2008; Meila, 2003; Zhao and Karypis, 2004). The metric must support certain reasonable constraints, such as giving a lower score to the solution that merges two clusters that correspond to different senses, or unnecessarily splits a single sense.[7] We also wanted to see the comparison produced by metrics that (1) do not require set matching to evaluate a particular clustering solution, and/or (2) consider the quality of mapping in both directions.

We used the following metrics to evaluate the performance of our system: (1) F-measure (2) BCubed P&R (Amigó et al., 2008) (3) mutual information as used in Meila (2003). We review the latter two measures below:

[5] For further details on data set creation, see Rumshisky and Batiukova (2008).

[6] We used the Maximum Entropy classifier from the CARAFE project available at `http://sourceforge.net/projects/carafe`.

[7] See, for example, Amigó et al. (2008) for similar considerations.

"BCubed" measures: We used the harmonic mean of BCubed precision and recall, which are defined for a given clustering solution C and a sense assignment solution S on data set D as follows:

BCubed Precision = $\frac{\sum_e \frac{|C(e) \cap S(e)|}{|C(e)|}}{n}$

BCubed Recall = $\frac{\sum_e \frac{|C(e) \cap S(e)|}{|S(e)|}}{n}$

where $e \in D$ is an element of the data set, $C(e)$ is the cluster to which e belongs, and $S(e)$ is the sense category to which e belongs, and $n = |D|$.

Entropy/MI measures: We used the standard mutual information measure of two variables defined by the clustering solution and the sense assignment $I(C, S)$ in the way delineated in Meila (2003):

$$I(C,S) = \sum_{k,k'} P(k,k') \log \frac{P(k,k')}{P(k)P(k')}$$

where $c_k \in C$ is a cluster from the clustering solution C, and $s_{k'} \in S$ is a sense from the sense assignment S, and $P(k,k') = \frac{|c_k \cap s_{k'}|}{n}$. The range for $I(C,S)$ depends on the entropy values of the two variables, $H(C)$ and $H(S)$:

$$0 \leq I(C,S) \leq \min(H(C), H(S))$$

Meila (2003) proposed a related *variant of information* measure $VI = H(C) + H(S) - 2I(C,S)$ which suffers from the same problem, i.e. its maximum depends on the respective entropy values. Since we needed to perform comparisons across different data sets, we used $I(C,S)$ normalized by $\max(H(C), H(S))$:

NormalizedMI = $\frac{I(C,S)}{\max(H(S),H(C))}$

which allowed us to retain the $(0, 1)$ range and certain other desirable properties, such as:

NormalizedMI(*1c1word*,S)= 0
NormalizedMI(*1c1inst*,S)= $H(S)/\log n$
NormalizedMI(S,S)= 1

In Task-2 of SEMEVAL-2007, the participant sense induction systems were evaluated using Wall Street Journal data annotated with OntoNotes senses (Hovy et al., 2006). While we could not reuse that data set with our system, we performed a set of comparisons of our system's performance relative to the characteristics of our data set.

SEMEVAL Task-2 used two kinds of evaluation: supervised and unsupervised.[8] We used the unsupervised method in our comparison. This method used a set-matching evaluation technique optimizing F-measure. The set matching stage found the optimum cluster for each sense, and averaged the F-measure of the best-matching cluster across all senses. Two relevant baselines were computed for the data set: (1) all instances for the given target word clustered together (*1cluster1word*) and (2) each instance treated as a separate cluster (*1cluster1inst*).

Since our own data set only included verbs, we recomputed the metrics for the verbs in SEMEVAL data set, based on the published clustering solutions for each participating system. In addition to the F-measure based metric, we computed the BCubed and NormalizedMI metrics both for our system and for the SEMEVAL data.

Table 2 summarizes the values obtained for these metrics by four configurations of our system. Table 3 gives the values obtained for the same metrics for each of the systems in SEMEVAL Task-2. We compute the metrics for the verbs in the SEMEVAL dataset, based on the published clustering solutions for each participating system. We give the values over the full dataset (i.e. both test and training data).

Variant	F-measure	% 1c1w	BCubed	% 1c1w	Norm. MI	% 1c1i
1c1inst	.038	6.5	.040	6.7	.188	100
1c1word	.584	100	.599	100	0	0
mi-fact	.586	100.3	.522	87.1	.138	73.4
mi-fact-prod	.572	97.9	.540	90.2	.061	32.4
mi	.504	86.3	.439	73.3	.103	54.8
mi-prod	.544	93.2	.469	78.3	.101	53.7

Table 2: Performance of our system for different clustering configurations

System	F-measure	% 1c1w	BCubed	% 1c1w	Norm. MI	% 1c1i
1c1inst	.035	4.6	.039	5.0	.118	100
1c1word	.755	100	.776	100	0	0
I2R	.528	69.9	.505	65.1	.051	43.2
UBC-AS	.750	99.3	.769	99.1	.005	4.2
UMND2	.640	84.8	.638	82.2	.006	5.1
UOY	.383	50.7	.253	32.6	.048	40.7
upv_si	.607	80.4	.520	67.0	.044	37.3

Table 3: SEMEVAL Task-2 system performance

The reported values in both tables are averages across all target words in the data set. To aid com-

[8] Supervised evaluation used a set-matching technique under which the obtained accuracy depends strongly on the majority baseline for each word in the data set.

parison across data sets, next to the actual value obtained by each system, we give the ratio of that value to the best performing baseline.

The verbs in our test data set have a significantly higher degree of polysemy compared to the SEMEVAL data. While the average number of senses per verb in our data and in SEMEVAL data is very similar (3.73 and 3.54, respectively), the distributions of senses differ. The average per-verb entropy for our data set is 1.4, as compared with the 0.9 value for the SEMEVAL data. Consequently, our data has a much lower majority baseline and therefore is potentially more difficult to classify. Note that the average number of instances per target in our data set was similar to the SEMEVAL data set, so the higher value of *1c1word* baseline for NormalizedMI reflects only the difference in the entropy of the annotated data.

Table 4 shows the F-measure values obtained for two baselines and for our best-performing configuration (*mi-factor*), for each verb in our data set. The random baseline was computed in the following way: for each verb, we randomly split the instances into clusters of the same number and size as the sense classes in the annotated data, and calculated the resulting F-measure, averaged over 10 runs.

5 Conclusion and Future Work

The resulting system for clustering selectors of polysemous words can be used in a number of ways. For example, Gamallo (2005) uses a similar approach to clustering of argument positions to improve prepositional phrase attachment in Portuguese.

Using our system, we can also easily produce heterogeneous sets of arguments that select for the same sense. For example, consider the word *position* whose meaning is so underspecified that it almost always requires a modifier in order to be disambiguated. In the BNC, the top modifiers of *position* (ncmod relation in RASP; with collocation ranking computed with log-factor adjusted MI score) are *sitting*, *predicative*, and *dominant*. Using the dendrogram obtained for the ncmod relation for *position*, we can sort the clusters whose selector lists include a given collocate, so that the cluster in which the given collocate has the highest average MI is placed at the top. For the collocates above, this method places at the top the clusters with the following selector lists[9]:

sitting: stooped-j 11.4, kneeling-j 11.3, recumbent-j 11.0, seated-j 10.1, commanding-j 8.5, standing-j 7.5, ...
cluster: [*figure-n posture-n*]

predicative: attributive-j 14.1, predicative-j 13.7, postnominal-j 13.2, clausal-j 13.1, predicate-n 10.0, postverbal-j 9.1, syntactic-j 8.8, prenominal-j 8.3, ordinal-j 6.8, adjectival-j 6.1, ...
cluster: [*construction-n adjective-n*]

dominant: interactionist-n 11.4 marxist-j 10.3, pluralist-n 10.2, philosophical-j 8.8, popperian-j 8.7, antiracist-j 7.8, phenomenological-j 7.5, kantian-j 7.4, structuralist-j 7.4, essentialist-j 7.3, functionalist-j 7.2, dominant-j 7.1, holist-j 6.9, doctrinal-j 6.6, materialist-n 6.4, theoretical-j 5.4, ideological-j 5.1, ...
cluster: [*conception-n perspective-n critique-n*]

Notice that the phrase *dominant position* is actually ambiguous. The second cluster in the sorted list for *dominant* identifies the other sense:

dominant: monopolistic-j 8.1, leading-j 8.0, competing-j 7.6, respected-j 6.1, rival-j 6.0, monopoly-n 5.5, established-j 5.3, dominant-j 5.2, competitive-j 4.6, well-established-j 4.5, ...
cluster: [*manufacturer-n firm-n producer-n provider-n supplier-n*]

The resulting heterogeneous selector sets could be used to improve ambiguity resolution in statistical machine translation. Another application of this system would be to utilize multiple relations per instance. Selector-cluster association scores currently used to classify sense-tagged instances of the target word can be used to choose which relations are to affect disambiguation.

In summary, we have presented a system designed to assess the impact of semantics of the argument on different types of ambiguities. For the cases when ambiguity may be resolved by the semantics of the arguments, our system outperforms full-context WSI systems. This system allows for a number of interesting applications that should be investigated in the future.

References

E. Agirre and A. Soroa. 2007. Semeval-2007 task 02: Evaluating word sense induction and discrimination systems. In *Proceedings of SemEval-2007*, pages 7–12, Prague, Czech Republic, June. ACL.

E. Agirre, L. Màrquez, and R. Wicentowski, editors. 2007. *Proceedings of SemEval-2007*. ACL, Prague, Czech Republic, June.

E. Amigó, J. Gonzalo, and J. Artiles. 2008. A comparison of extrinsic clustering evaluation metrics based on formal constraints. Techical report, Departamento de Lenguajes y Sistemas Informticos (UNED), Madrid, Spain.

BNC. 2000. *The British National Corpus*. The BNC Consortium, University of Oxford, *http://www.natcorp.ox.ac.uk/*.

[9] MI values for each selector, averaged across all elements of the cluster, are given next to the POS-marked lemma

Word	No. Senses	No. Inst.	ITA %	Entropy	MFS	MaxEnt accuracy	random	1c1word	mi-fact
absorb	7	196	92.4	2.49	.30	.58	.20	.33	.36
acquire	4	186	92.1	1.86	.44	.44	.30	.45	.59
admit	2	163	98.7	1.00	.53	.71	.51	.67	.74
assume	3	191	90.8	1.55	.45	.73	.39	.52	.48
conclude	2	178	97.5	0.96	.62	.89	.55	.68	.51
cut	4	166	92.3	1.33	.58	.51	.49	.61	.78
deny	3	190	97.2	1.49	.49	.62	.38	.54	.55
dictate	2	193	98.9	0.53	.88	.97	.79	.85	.62
drive	11	174	97.6	2.64	.41	.40	.23	.34	.39
edit	2	176	98.0	0.98	.57	.82	.57	.67	.62
enjoy	2	193	86.2	0.93	.66	.70	.57	.70	.53
fire	6	162	97.3	1.87	.54	.73	.37	.49	.58
grasp	3	178	97.6	1.25	.49	.84	.45	.61	.85
know	2	172	92.6	0.98	.58	.79	.54	.67	.56
launch	3	196	89.9	1.24	.63	.74	.52	.62	.66
Average	3.73	180.9	94.5	1.41	.545	.699	.457	.584	.586

Table 4: Per-word characteristics of the data set and system performance

T. Briscoe and J. Carroll. 2002. Robust accurate statistical annotation of general text. *Proceedings of LREC 2002, Las Palmas, Canary Islands*, pages 1499–1504.

K. Erk and S. Pado. 2008. A structured vector space model for word meaning in context. In *Proceedings of EMNLP*.

P. Gamallo, A. Agustini, and G. Lopes. 2005. Clustering syntactic positions with similar semantic requirements. *Computational Linguistics*, 31(1):107–145.

G. Grefenstette. 1994. *Explorations in Automatic Thesaurus Discovery*. Kluwer Academic Publishers, Norwell, MA, USA.

Z. Harris. 1985. Distributional structure. In J. Katz, editor, *Philosophy of Linguistics*, pages 26–47. Oxford University Press, New York.

D. Hindle. 1990. Noun classification from predicate-argument structures. In *Proceedings of ACL*, pages 268–275, Pittsburgh, PA. ACL.

E. Hovy, M. Marcus, M. Palmer, L. Ramshaw, and R. Weischedel. 2006. OntoNotes: The 90% solution. In *Proceedings of HLT-NAACL, Companion Volume: Short Papers*, pages 57–60, New York City, USA, June. ACL.

A. Kilgarriff, P. Rychly, P. Smrz, and D. Tugwell. 2004. The Sketch Engine. *Proceedings of Euralex, Lorient, France*, pages 105–116.

S. Landes, C. Leacock, and R.I. Tengi. 1998. Building semantic concordances. In C. Fellbaum, editor, *Wordnet: an electronic lexical database*. MIT Press, Cambridge (Mass.).

M. Meila. 2003. Comparing clusterings. Techical Report TR418, University of Washington, Department of Statistics.

R. Mihalcea and P. Edmonds, editors. 2004. *Senseval-3: Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*, Barcelona, Spain, July. ACL.

G. Miller and W. Charles. 1991. Contextual correlates of semantic similarity. *Language and Cognitive Processes*, 6(1):1–28.

J. Mitchell and M. Lapata. 2008. Vector-based models of semantic composition. In *Proceedings of ACL*, pages 236–244.

M. Palmer, D. Gildea, and P. Kingsbury. 2005. The Proposition Bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31(1):71–106.

M. Palmer, H. Dang, and C. Fellbaum. 2007. Making fine-grained and coarse-grained sense distinctions, both manually and automatically. *Journal of Natural Language Engineering*.

P. Pantel and D. Lin. 2002. Discovering word senses from text. In *Proceedings of ACM SIGKDD02*.

J Preiss and D. Yarowsky, editors. 2001. *Proceedings of the Second Int. Workshop on Evaluating WSD Systems (Senseval 2)*. ACL2002/EACL2001.

J. Pustejovsky and B. Boguraev. 1993. Lexical knowledge representation and natural language processing. *Artif. Intell.*, 63(1-2):193–223.

J. Pustejovsky, P. Hanks, and A. Rumshisky. 2004. Automated Induction of Sense in Context. In *COLING 2004, Geneva, Switzerland*, pages 924–931.

A. Rumshisky and O. Batiukova. 2008. Polysemy in verbs: systematic relations between senses and their effect on annotation. In *HJCL-2008*, Manchester, England.

A. Rumshisky. 2008. Resolving polysemy in verbs: Contextualized distributional approach to argument semantics. In A. Lenci, editor, *Distributional Models of the Lexicon, A Special Issue of Rivista di Linguistica*.

J. Ruppenhofer, M. Ellsworth, M. Petruck, C. Johnson, and J. Scheffczyk. 2006. *FrameNet II: Extended Theory and Practice*.

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

Y. Zhao and G. Karypis. 2004. Empirical and theoretical comparisons of selected criterion functions for document clustering. *Machine Learning*, 55:311–331.