Ensemble-based Semantic Lexicon Induction for Semantic Tagging

Ashequl Qadir
University of Utah
School of Computing
Salt Lake City, UT 84112, USA
asheq@cs.utah.edu

Ellen Riloff
University of Utah
School of Computing
Salt Lake City, UT 84112, USA
riloff@cs.utah.edu

Abstract

We present an ensemble-based framework for semantic lexicon induction that incorporates three diverse approaches for semantic class identification. Our architecture brings together previous bootstrapping methods for pattern-based semantic lexicon induction and contextual semantic tagging, and incorporates a novel approach for inducing semantic classes from coreference chains. The three methods are embedded in a bootstrapping architecture where they produce independent hypotheses, consensus words are added to the lexicon, and the process repeats. Our results show that the ensemble outperforms individual methods in terms of both lexicon quality and instance-based semantic tagging.

1 Introduction

One of the most fundamental aspects of meaning is the association between words and semantic categories, which allows us to understand that a "cow" is an *animal* and a "house" is a *structure*. We will use the term *semantic lexicon* to refer to a dictionary that associates words with semantic classes. Semantic dictionaries are useful for many NLP tasks, as evidenced by the widespread use of WordNet (Miller, 1990). However, off-the-shelf resources are not always sufficient for specialized domains, such as medicine, chemistry, or microelectronics. Furthermore, in virtually every domain, texts contain lexical variations that are often missing from dictionaries, such as acronyms, abbreviations, spelling variants, informal shorthand terms (e.g., "abx" for "antibiotics"), and composite terms (e.g., "may-december" or "virus/worm"). To address this problem, techniques have been developed to automate the construction of semantic lexicons from text corpora using bootstrapping methods (Riloff and Shepherd, 1997; Roark and Charniak, 1998; Phillips and Riloff, 2002; Thelen and Riloff, 2002; Ng, 2007; McIntosh and Curran, 2009; McIntosh, 2010), but accuracy is still far from perfect.

Our research explores the use of *ensemble* methods to improve the accuracy of semantic lexicon induction. Our observation is that semantic class associations can be learned using several fundamentally different types of corpus analysis. Bootstrapping methods for semantic lexicon induction (e.g., (Riloff and Jones, 1999; Thelen and Riloff, 2002; McIntosh and Curran, 2009)) collect corpus-wide statistics for individual words based on shared contextual patterns. In contrast, classifiers for semantic tagging (e.g., (Collins and Singer, 1999; Niu et al., 2003; Huang and Riloff, 2010)) label *word instances* and focus on the local context surrounding each instance. The difference between these approaches is that semantic taggers make decisions based on a single context and can assign different labels to different instances, whereas lexicon induction algorithms compile corpus statistics from multiple instances of a word and typically assign each word to a single semantic category.[1] We also hypothesize that coreference resolution can be exploited to infer semantic

[1]This approach would be untenable for broad-coverage semantic knowledge acquisition, but within a specialized domain most words have a dominant word sense. Our experimental results support this assumption.

*First Joint Conference on Lexical and Computational Semantics (*SEM)*, pages 199–208,
Montréal, Canada, June 7-8, 2012. ©2012 Association for Computational Linguistics

class labels. Intuitively, if we know that two noun phrases are coreferent, then they probably belong to the same high-level semantic category (e.g., "dog" and "terrier" are both *animals*).

In this paper, we present an ensemble-based framework for semantic lexicon induction. We incorporate a pattern-based bootstrapping method for lexicon induction, a contextual semantic tagger, and a new coreference-based method for lexicon induction. Our results show that coalescing the decisions produced by diverse methods produces a better dictionary than any individual method alone.

A second contribution of this paper is an analysis of the effectiveness of dictionaries for semantic tagging. In principle, an NLP system should be able to assign different semantic labels to different senses of a word. But within a specialized domain, most words have a dominant sense and we argue that using domain-specific dictionaries for tagging may be equally, if not more, effective. We analyze the tradeoffs between using an instance-based semantic tagger versus dictionary lookup on a collection of disease outbreak articles. Our results show that the induced dictionaries yield better performance than an instance-based semantic tagger, achieving higher accuracy with comparable levels of recall.

2 Related Work

Several techniques have been developed for *semantic class induction* (also called *set expansion*) using bootstrapping methods that consider co-occurrence statistics based on nouns (Riloff and Shepherd, 1997), syntactic structures (Roark and Charniak, 1998; Phillips and Riloff, 2002), and contextual patterns (Riloff and Jones, 1999; Thelen and Riloff, 2002; McIntosh and Curran, 2008; McIntosh and Curran, 2009). To improve the accuracy of induced lexicons, some research has incorporated negative information from human judgements (Vyas and Pantel, 2009), automatically discovered negative classes (McIntosh, 2010), and distributional similarity metrics to recognize concept drift (McIntosh and Curran, 2009). Phillips and Riloff (2002) used co-training (Blum and Mitchell, 1998) to exploit three simple classifiers that each recognized a different type of syntactic structure. The research most closely related to ours is an ensemble-based method for automatic thesaurus construction (Curran, 2002). However, that goal was to acquire fine-grained semantic information that is more akin to synonymy (e.g., words similar to "house"), whereas we associate words with high-level semantic classes (e.g., a "house" is a *transient structure*).

Semantic class tagging is closely related to *named entity recognition* (NER) (e.g., (Bikel et al., 1997; Collins and Singer, 1999; Cucerzan and Yarowsky, 1999; Fleischman and Hovy, 2002)). Some bootstrapping methods have been used for NER (e.g., (Collins and Singer, 1999; Niu et al., 2003) to learn from unannotated texts. However, most NER systems will not label nominal noun phrases (e.g., they will not identify "the dentist" as a *person*) or recognize semantic classes that are not associated with proper named entities (e.g., symptoms).[2] ACE mention detection systems (e.g., (ACE, 2007; ACE, 2008)) can label noun phrases that are associated with 5-7 semantic classes and are typically trained with supervised learning. Recently, (Huang and Riloff, 2010) developed a bootstrapping technique that induces a semantic tagger from unannotated texts. We use their system in our ensemble.

There has also been work on extracting semantic class members from the Web (e.g., (Paşca, 2004; Etzioni et al., 2005; Kozareva et al., 2008; Carlson et al., 2009)). This line of research is fundamentally different from ours because these techniques benefit from the vast repository of information available on the Web and are therefore designed to harvest a wide swath of general-purpose semantic information. Our research is aimed at acquiring domain-specific semantic dictionaries using a collection of documents representing a specialized domain.

3 Ensemble-based Semantic Lexicon Induction

3.1 Motivation

Our research combines three fundamentally different techniques into an ensemble-based bootstrapping framework for semantic lexicon induction: pattern-based dictionary induction, contextual semantic tagging, and coreference resolution. Our motivation for using an ensemble of different tech-

[2] Some NER systems will handle special constructions such as dates and monetary amounts.

niques is driven by the observation that these methods exploit different types of information to infer semantic class knowledge. The coreference resolver uses features associated with coreference, such as syntactic constructions (e.g., appositives, predicate nominals), word overlap, semantic similarity, proximity, etc. The pattern-based lexicon induction algorithm uses corpus-wide statistics gathered from the contexts of all instances of a word and compares them with the contexts of known category members. The contextual semantic tagger uses local context windows around words and classifies each word instance independently from the others.

Since each technique draws its conclusions from different types of information, they represent independent sources of evidence to confirm whether a word belongs to a semantic class. Our hypothesis is that, combining these different sources of evidence in an ensemble-based learning framework should produce better accuracy than using any one method alone. Based on this intuition, we create an ensemble-based bootstrapping framework that iteratively collects the hypotheses produced by each individual learner and selects the words that were hypothesized by at least 2 of the 3 learners. This approach produces a bootstrapping process with improved precision, both at the critical beginning stages of the bootstrapping process and during subsequent bootstrapping iterations.

3.2 Component Systems in the Ensemble

In the following sections, we describe each of the component systems used in our ensemble.

3.2.1 Pattern-based Lexicon Induction

The first component of our ensemble is Basilisk (Thelen and Riloff, 2002), which identifies nouns belonging to a semantic class based on collective information over lexico-syntactic pattern contexts. The patterns are automatically generated using AutoSlog-TS (Riloff, 1996). Basilisk begins with a small set of seed words for each semantic class and a collection of unannotated documents for the domain. In an iterative bootstrapping process, Basilisk identifies candidate nouns, ranks them based on its scoring criteria, selects the 5 most confident words for inclusion in the lexicon, and this process repeats using the new words as additional seeds in subsequent iterations.

3.2.2 Lexicon Induction with a Contextual Semantic Tagger

The second component in our ensemble is a contextual semantic tagger (Huang and Riloff, 2010). Like Basilisk, the semantic tagger also begins with seed nouns, trains itself on a large collection of unannotated documents using bootstrapping, and iteratively labels new instances. This tagger labels noun instances and does not produce a dictionary.

To adapt it for our purposes, we ran the bootstrapping process over the training texts to induce a semantic classifier. We then applied the classifier to the same set of training documents and compiled a lexicon by collecting the set of nouns that were assigned to each semantic class. We ignored words that were assigned different labels in different contexts to avoid conflicts in the lexicons. We used the identical configuration described by (Huang and Riloff, 2010) that applies a 1.0 confidence threshold for semantic class assignment.

3.2.3 Coreference-Based Lexicon Construction

The third component of our ensemble is a new method for semantic lexicon induction that exploits coreference resolution. Members of a coreference chain represent the same entity, so all references to the entity should belong to the same semantic class. For example, suppose *"Paris"* and *"the city"* are in the same coreference chain. If we know that *city* is a *Fixed Location*, then we can infer that *Paris* is also a *Fixed Location*.

We induced lexicons from coreference chains using a similar bootstrapping framework that begins with seed nouns and unannotated texts. Let S denote a set of semantic classes and W denote a set of unknown words. For any $s \in S$ and $w \in W$, let $N_{s,w}$ denote the number of instances of s in the current lexicon[3] that are coreferent with w in the text corpus. Then we estimate the probability that word w belongs to semantic class s as:

$$P(s|w) = \frac{N_{s,w}}{\sum_{s' \in S} N_{s',w}}$$

We hypothesize the semantic class of w, $SemClass(w)$ by:

$$SemClass(w) = \arg\max_s P(s|w)$$

[3] In the first iteration, the lexicon is initialized with the seeds.

To ensure high precision for the induced lexicons, we use a threshold of 0.5. All words with a probability above this thresold are added to the lexicon, and the bootstrapping process repeats. Although the coreference chains remain the same throughout the process, the lexicon grows so more words in the chains have semantic class labels as bootstrapping progresses. Bootstrapping ends when fewer than 5 words are learned for each of the semantic classes.

Many noun phrases are singletons (i.e., they are not coreferent with any other NPs), which limits the set of words that can be learned using coreference chains. Furthermore, coreference resolvers make mistakes, so the accuracy of the induced lexicons depends on the quality of the chains. For our experiments, we used Reconcile (Stoyanov et al., 2010), a freely available supervised coreference resolver.

3.3 Ensemble-based Bootstrapping Framework

Figure 1 shows the architecture of our ensemble-based bootstrapping framework. Initially, each lexicon only contains the seed nouns. Each component hypothesizes a set of candidate words for each semantic class, based on its own criteria. The word lists produced by the three systems are then compared, and we retain only the words that were hypothesized with the same class label by at least two of the three systems. The remaining words are discarded. The consenus words are added to the lexicon, and the bootstrapping process repeats. As soon as fewer than 5 words are learned for each of the semantic classes, bootstrapping stops.

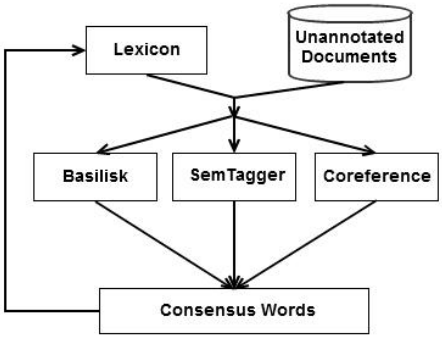

Figure 1: Ensemble-based bootstrapping framework

We ran each individual system with the same seed words. Since bootstrapping typically yields the best precision during the earliest stages, we used the semantic tagger's trained model immediately after its first bootstrapping iteration. Basilisk generates 5 words per cycle, so we report results for lexicons generated after 20 bootstrapping cycles (100 words) and after 80 bootstrapping cycles (400 words).

3.4 Co-Training Framework

The three components in our ensemble use different types of features (views) to identify semantic class members, so we also experimented with co-training. Our co-training model uses an identical framework, but the hypotheses produced by the different methods are all added to the lexicon, so each method can benefit from the hypotheses produced by the others. To be conservative, each time we added only the 10 most confident words hypothesized by each method.

In contrast, the ensemble approach only adds words to the lexicon if they are hypothesized by two different methods. As we will see in Section 4.4, the ensemble performs much better than co-training. The reason is that the individual methods do not consistently achieve high precision on their own. Consequently, many mistakes are added to the lexicon, which is used as training data for subsequent bootstrapping. The benefit of the ensemble is that consensus is required across two methods, which serves as a form of cross-checking to boost precision and maintain a high-quality lexicon.

4 Evaluation

4.1 Semantic Class Definitions

We evaluated our approach on nine semantic categories associated with disease outbreaks. The semantic classes are defined below.

Animal: Mammals, birds, fish, insects and other animal groups. (e.g., *cow, crow, mosquito, herd*)

[4] http://www.nlm.nih.gov/research/umls/
[5] http://www.maxmind.com/app/worldcities
[6] http://www.listofcountriesoftheworld.com/
[7] http://names.mongabay.com/most_common_surnames.htm
[8] http://www.sec.gov/rules/other/4-460list.htm
[9] http://www.utexas.edu/world/univ/state/
[10] http://www.uta.fi/FAST/GC/usabacro.html/

Semantic Class	External Word List Sources
Animal	**WordNet**: [animal], [mammal family], [animal group]
Body Part	**WordNet**: [body part], [body substance], [body covering], [body waste]
DisSym	**WordNet**: [symptom], [physical condition], [infectious agent]; **Wikipedia**: common and infectious diseases, symptoms, disease acronyms; **UMLS Thesaurus**[4]: diseases, abnormalities, microorganisms (Archaea, Bacteria, Fungus, Virus)
Fixed Loc.	**WordNet**: [geographic area], [land], [district, territory], [region]; **Wiki**:US-states; **Other**:cities[5], countries[6]
Human	**WordNet**: [person], [people], [personnel]; **Wikipedia**: people names, office holder titles, nationalities, occupations, medical personnels & acronyms, players; **Other**: common people names & surnames[7]
Org	**WordNet**: [organization], [assembly]; **Wikipedia**: acronyms in healthcare, medical organization acronyms, news agencies, pharmaceutical companies; **Other**: companies[8], US-universities[9], organizations[10]
Plant & Food	**WordNet**: [food], [plant, flora], [plant part]
Temp. Ref.	**WordNet**: [time], [time interval], [time unit],[time period] **TimeBank**: TimeBank1.2 (Pustejovsky et al., 2003) TIMEX3 expressions
Trans. Struct.	**WordNet**: [structure, construction], [road, route], [facility, installation], [work place]

Table 1: External Word List Sources

Body Part: A part of a human or animal body, including organs, bodily fluids, and microscopic parts. (e.g., *hand, heart, blood, DNA*)

Diseases and Symptoms (DisSym): Diseases and symptoms. We also include fungi and disease carriers because, in this domain, they almost always refer to the disease that they carry. (e.g. *FMD, Anthrax, fever, virus*)

Fixed Location (Fixed Loc.): Named locations, including countries, cities, states, etc. We also include directions and well-defined geographic areas or geo-political entities. (e.g., *Brazil, north, valley*)

Human: All references to people, including names, titles, professions, and groups. (e.g., *John, farmer, traders*)

Organization (Org.): An entity that represents a group of people acting as a single recognized body, including named organizations, departments, governments, and their acronyms. (e.g., *department, WHO, commission, council*)

Temporal Reference (Temp. Ref.): Any reference to a time or duration, including months, days, seasons, etc. (e.g., *night, May, summer, week*)

Plants & Food[11]**:** plants, plant parts, or any type of food. (e.g., *seed, mango, beef, milk*)

Transient Structures (Trans. Struct.): Transient physical structures. (e.g., *hospital, building, home*)

Additionally, we defined a **Miscellaneous** class for words that do not belong to any of the other categories. (e.g., *output, information, media, point*).

[11]We merged plants and food into a single category as it is difficult to separate them because many food items are plants.

4.2 Data Set

We ran our experiments on ProMED-mail[12] articles. ProMED-mail is an internet based reporting system for infectious disease outbreaks, which can involve people, animals, and plants grown for food. Our ProMED corpus contains 5004 documents. We used 4959 documents as (unannotated) training data for bootstrapping. For the remaining 45 documents, we used 22 documents to train the coreference resolver (Reconcile) and 23 documents as our test set. The coreference training set contains MUC-7 style (Hirschman, 1997) coreference annotations[13]. Once trained, Reconcile was applied to the 4959 unannotated documents to produce coreference chains.

4.3 Gold Standard Semantic Class Annotations

To obtain gold standard annotations for the test set, two annotators assigned one of the 9 semantic class labels, or Miscellaneous, to each head noun based on its surrounding context. A noun with multiple senses could get assigned different semantic class labels in different contexts. The annotators first annotated 13 of the 23 documents, and discussed the cases where they disagreed. Then they independelty annotated

[12]`http://www.promedmail.org/`

[13]We omit the details of the coreference annotations since it is not the focus of this research. However, the annotators measured their agreement on 10 documents and achieved MUC scores of Precision = .82, Recall = .86, F-measure = .84.

the remaining 10 documents and measured inter-annotator agreement with Cohen's Kappa (κ) (Carletta, 1996). The κ score for these 10 documents was 0.91, indicating a high level of agreement. The annotators then adjudicated their disagreements on all 23 documents to create the gold standard.

4.4 Dictionary Evaluation

To assess the quality of the lexicons, we estimated their accuracy by compiling external word lists from freely available sources such as Wikipedia[14] and WordNet (Miller, 1990). Table 1 shows the sources that we used, where the bracketed items refer to WordNet hypernym categories. We searched each WordNet hypernym tree (also, instance-relationship) for all senses of the word. Additionally, we collected the manually labeled words in our test set and included them in our gold standard lists.

Since the induced lexicons contain individual nouns, we extracted only the head nouns of multi-word phrases in the external resources. This can produce incorrect entries for non-compositional phrases, but we found this issue to be relatively rare and we manually removed obviously wrong entries. We adopted a conservative strategy and assumed that any lexicon entries not present in our gold standard lists are incorrect. But we observed many correct entries that were missing from the external resources, so our results should be interpreted as a lower bound on the true accuracy of the induced lexicons.

We generated lexicons for each method separately, and also for the ensemble and co-training models. We ran Basilisk for 100 iterations (500 words). We refer to a Basilisk lexicon of size N using the notation $B[N]$. For example, $B400$ refers to a lexicon containing 400 words, which was generated from 80 bootstrapping cycles. We refer to the lexicon obtained from the semantic tagger as *ST Lex*.

Figure 2 shows the dictionary evaluation results. We plotted *Basilisk's* accuracy after every 5 bootstrapping cycles (25 words). For *ST Lex*, we sorted the words by their confidence scores and plotted the accuracy of the top-ranked words in increments of 50. The plots for *Coref*, *Co-Training*, and *Ensemble B[N]* are based on the lexicons produced after each bootstrapping cycle.

[14] www.wikipedia.org/

The ensemble-based framework yields consistently better accuracy than the individual methods for *Animal*, *Body Part*, *Human* and *Temporal Reference*, and similar if not better for *Disease & Symptom*, *Fixed Location*, *Organization*, *Plant & Food*. However, relying on consensus from multiple models produce smaller dictionaries. Big dictionaries are not always better than small dictionaries in practice, though. We believe, it matters more whether a dictionary contains the most frequent words for a domain, because they account for a disproportionate number of instances. Basilisk, for example, often learns infrequent words, so its dictionaries may have high accuracy but often fail to recognize common words. We investigate this issue in the next section.

4.5 Instance-based Tagging Evaluation

We also evaluated the effectiveness of the induced lexicons with respect to instance-based semantic tagging. Our goal was to determine how useful the dictionaries are in two respects: (1) do the lexicons contain words that appear frequently in the domain, and (2) is dictionary look-up sufficient for instance-based labeling? Our bootstrapping processes enforce a constraint that a word can only belong to one semantic class, so if polysemy is common, then dictionary look-up will be problematic.[15]

The instance-based evaluation assigns a semantic label to each instance of a head noun. When using a lexicon, all instances of the same noun are assigned the same semantic class via dictionary look-up. The semantic tagger (SemTag), however, is applied directly since it was designed to label instances.

Table 2 presents the results. As a baseline, the *W.Net* row shows the performance of WordNet for instance tagging. For words with multiple senses, we only used the first sense listed in WordNet. The *Seeds* row shows the results when performing dictionary look-up using only the seed words. The remaining rows show the results for Basilisk (B100 and B400), coreference-based lexicon induction (Coref), lexicon induction using the semantic tagger (ST Lex), and the original instance-based tagger (SemTag). The following rows show the results for co-training (after 4 iterations and 20 iterations)

[15] Only coarse polysemy across semantic classes is an issue (e.g., "plant" as a living thing vs. a factory).

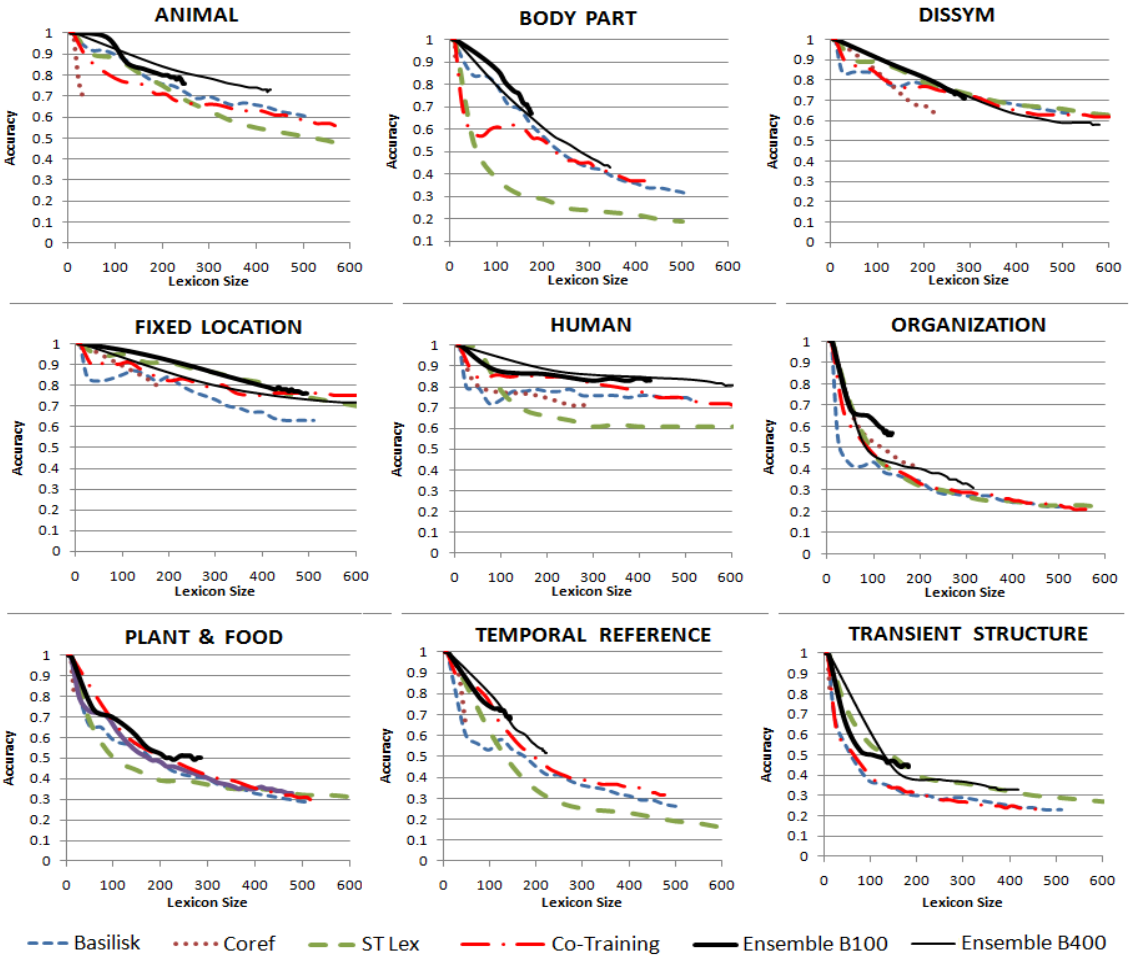

Figure 2: Dictionary Evaluation Results

and for the ensemble (using Basilisk size 100 and size 400). Table 3 shows the micro & macro average results across all semantic categories.

Table 3 shows that the dictionaries produced by the *Ensemble w/B100* achieved better results than the individual methods and co-training with an F score of 80%. Table 2 shows that the ensemble achieved better performance than the other methods for 4 of the 9 classes, and was usually competitive on the remaining 5 classes. WordNet *(W.Net)* consistently produced high precision, but with comparatively lower recall, indicating that WordNet does not have sufficient coverage for this domain.

4.6 Analysis

Table 4 shows the performance of our ensemble when using only 2 of the 3 component methods. Removing any one method decreases the average F-measure by at least 3-5%. Component pairs that include induced lexicons from coreference (*ST Lex+Coref* and *B100+Coref*) yield high precision but low recall. The component pair *ST Lex+B100* produces higher recall but with slightly lower accuracy. The ensemble framework boosted recall even more, while maintaining the same precision.

We observe that some of the smallest lexicons produced the best results for instance-based semantic tagging (e.g., *Organization*). Our hypothesis is that consensus decisions across different methods helps to promote the acquisition of high frequency domain words, which are crucial to have in the dictionary. The fact that dictionary look-up performed better than an instance-based semantic tagger also suggests that coarse polysemy (different senses that

Method	Animal	Body Part	DisSym	Fixed Loc.	Human	Org.	Plant & Food	Temp. Ref.	Trans. Struct.
	P R F	P R F	P R F	P R F	P R F	P R F	P R F	P R F	P R F
Individual Methods									
W.Net	92 88 90	93 59 72	99 77 87	86 58 69	83 55 66	86 44 59	65 79 71	93 85 **89**	85 64 73
Seeds	100 54 70	92 55 69	100 59 74	95 10 18	100 22 36	100 41 58	100 61 **76**	100 52 69	100 09 17
B100	99 77 86	94 73 **82**	100 66 80	96 23 37	96 31 47	91 58 71	82 64 72	68 83 75	67 22 33
B400	94 90 92	51 86 64	100 69 81	97 35 51	91 51 65	79 77 78	46 82 59	49 94 64	83 78 **80**
Coref	90 67 77	92 55 69	66 83 73	65 46 54	57 50 53	54 68 60	81 61 69	60 74 67	45 09 15
ST Lex	94 89 91	68 77 72	80 91 85	91 74 82	79 43 55	84 62 71	51 68 58	73 91 81	82 49 61
SemTag	91 90 90	52 68 59	77 90 83	91 78 **84**	81 48 60	80 63 70	43 82 56	77 93 84	83 53 64
Co-Training									
pass4	64 76 70	67 73 70	91 79 85	91 39 54	98 44 61	83 69 76	43 68 53	73 94 82	49 36 42
pass20	60 89 71	56 91 69	88 91 90	83 64 72	92 54 68	72 77 74	28 71 40	65 98 78	46 40 43
Ensembles									
w/B100	93 94 **94**	74 77 76	93 81 86	92 73 81	94 55 **70**	90 78 **84**	56 89 68	55 94 70	79 75 77
w/B400	94 93 93	65 91 75	96 87 **91**	89 75 81	92 56 **70**	79 79 79	47 86 61	53 94 68	63 55 58

Table 2: Instance-based Semantic Tagging Results (P = Precision, R = Recall, F = F-measure)

Method	Micro Average	Macro Average
	P R F	P R F
Individual Systems		
W.Net	88 66 75	87 68 76
Seeds	99 35 52	99 40 57
B100	89 50 64	88 55 68
B400	77 66 71	77 74 75
Coref	65 59 62	68 57 62
ST Lex	82 72 77	78 72 75
SemTag	80 74 77	75 74 74
Co-Training		
pass4	77 61 68	73 64 68
pass20	69 74 71	65 75 70
Ensembles		
w/B100	83 77 **80**	81 80 **80**
w/B400	79 78 78	75 79 77

Table 3: Micro & Macro Average for Semantic Tagging

Method	Micro Average	Macro Average
	P R F	P R F
Ensemble with component pairs		
ST Lex+Coref	92 59 72	92 57 70
B100+Coref	92 40 56	94 44 60
ST Lex+B100	82 69 75	81 75 77
Ensemble with all components		
ST Lex+B100+Coref	83 77 **80**	81 80 **80**

Table 4: Ablation Study of the Ensemble Framework for Semantic Tagging

cut across semantic classes) is a relatively minor issue within a specialized domain.

5 Conclusions

Our research combined three diverse methods for semantic lexicon induction in a bootstrapped ensemble-based framework, including a novel approach for lexicon induction based on coreference chains. Our ensemble-based approach performed better than the individual methods, in terms of both dictionary accuracy and instance-based semantic tagging. In future work, we believe this approach could be enhanced further by adding new types of techniques to the ensemble and by investigating better methods for estimating the confidence scores from the individual components.

Acknowledgments

We are grateful to Lalindra de Silva for manually annotating data, Nathan Gilbert for help with Reconcile, and Ruihong Huang for help with the semantic tagger. We gratefully acknowledge the support of the National Science Foundation under grant IIS-1018314 and the Defense Advanced Research Projects Agency (DARPA) Machine Reading Program under Air Force Research Laboratory (AFRL) prime contract no. FA8750-09-C-0172. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the view of the DARPA, AFRL, or the U.S. government.

References

ACE. 2007. NIST ACE evaluation website. In *http://www.nist.gov/speech/tests/ace/2007*.

ACE. 2008. NIST ACE evaluation website. In *http://www.nist.gov/speech/tests/ace/2008*.

Daniel M. Bikel, Scott Miller, Richard Schwartz, and Ralph Weischedel. 1997. Nymble: a high-performance learning name-finder. In *Proceedings of ANLP-97*, pages 194–201.

A. Blum and T. Mitchell. 1998. Combining Labeled and Unlabeled Data with Co-Training. In *Proceedings of the 11th Annual Conference on Computational Learning Theory (COLT-98)*.

Jean Carletta. 1996. Assessing agreement on classification tasks: the kappa statistic. *Comput. Linguist.*, 22:249–254, June.

Andrew Carlson, Justin Betteridge, Estevam R. Hruschka Jr., and Tom M. Mitchell. 2009. Coupling semi-supervised learning of categories and relations. In *HLT-NAACL 2009 Workshop on Semi-Supervised Learning for NLP*.

M. Collins and Y. Singer. 1999. Unsupervised Models for Named Entity Classification. In *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP/VLC-99)*.

S. Cucerzan and D. Yarowsky. 1999. Language Independent Named Entity Recognition Combining Morphologi cal and Contextual Evidence. In *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP/VLC-99)*.

J. Curran. 2002. Ensemble Methods for Automatic Thesaurus Extraction. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing*.

O. Etzioni, M. Cafarella, D. Downey, A. Popescu, T. Shaked, S. Soderland, D. Weld, and A. Yates. 2005. Unsupervised named-entity extraction from the web: an experimental study. *Artificial Intelligence*, 165(1):91–134, June.

M.B. Fleischman and E.H. Hovy. 2002. Fine grained classification of named entities. In *Proceedings of the COLING conference*, August.

L. Hirschman. 1997. MUC-7 Coreference Task Definition.

Ruihong Huang and Ellen Riloff. 2010. Inducing domain-specific semantic class taggers from (almost) nothing. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*.

Z. Kozareva, E. Riloff, and E. Hovy. 2008. Semantic Class Learning from the Web with Hyponym Pattern Linkage Graphs. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-08)*.

T. McIntosh and J. Curran. 2008. Weighted mutual exclusion bootstrapping for domain independent lexicon and template acquisition. In *Proceedings of the Australasian Language Technology Association Workshop*.

T. McIntosh and J. Curran. 2009. Reducing Semantic Drift with Bagging and Distributional Similarity. In *Proceedings of the 47th Annual Meeting of the Association for Computational Linguistics*.

T. McIntosh. 2010. Unsupervised Discovery of Negative Categories in Lexicon Bootstrapping. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*.

G. Miller. 1990. Wordnet: An On-line Lexical Database. *International Journal of Lexicography*, 3(4).

V. Ng. 2007. Semantic Class Induction and Coreference Resolution. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*.

Cheng Niu, Wei Li, Jihong Ding, and Rohini K. Srihari. 2003. A bootstrapping approach to named entity classification using successive learners. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics (ACL-03)*, pages 335–342.

M. Paşca. 2004. Acquisition of categorized named entities for web search. In *Proc. of the Thirteenth ACM International Conference on Information and Knowledge Management*, pages 137–145.

W. Phillips and E. Riloff. 2002. Exploiting Strong Syntactic Heuristics and Co-Training to Learn Semantic Lexicons. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing*, pages 125–132.

J. Pustejovsky, P. Hanks, R. Saurí, A. See, R. Gaizauskas, A. Setzer, D. Radev, B. Sundheim, D. Day, L. Ferro, and M. Lazo. 2003. The TIMEBANK Corpus. In *Proceedings of Corpus Linguistics 2003*, pages 647–656.

E. Riloff and R. Jones. 1999. Learning Dictionaries for Information Extraction by Multi-Level Bootstrapping. In *Proceedings of the Sixteenth National Conference on Artificial Intelligence*.

E. Riloff and J. Shepherd. 1997. A Corpus-Based Approach for Building Semantic Lexicons. In *Proceedings of the Second Conference on Empirical Methods in Natural Language Processing*, pages 117–124.

E. Riloff. 1996. Automatically Generating Extraction Patterns from Untagged Text. In *Proceedings of the Thirteenth National Conference on Artificial Intelligence*, pages 1044–1049. The AAAI Press/MIT Press.

B. Roark and E. Charniak. 1998. Noun-phrase Co-occurrence Statistics for Semi-automatic Semantic

Lexicon Construction. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics*, pages 1110–1116.

Veselin Stoyanov, Claire Cardie, Nathan Gilbert, Ellen Riloff, David Buttler, and David Hysom. 2010. Coreference resolution with Reconcile. In *Proceedings of the ACL 2010 Conference Short Papers*, pages 156–161.

M. Thelen and E. Riloff. 2002. A Bootstrapping Method for Learning Semantic Lexicons Using Extraction Pa ttern Contexts. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing*, pages 214–221.

V. Vyas and P. Pantel. 2009. Semi-automatic entity set refinement. In *Proceedings of North American Association for Computational Linguistics / Human Language Technology (NAACL/HLT-09)*.