Learning Named Entity Hyponyms for Question Answering

Paul McNamee
JHU Applied Physics Laboratory
11100 Johns Hopkins Road
Laurel, MD 20723-6099, USA
paul.mcnamee@jhuapl.edu

Rion Snow
Stanford AI Laboratory
Stanford University
Stanford, CA 94305, USA
rion@cs.stanford.edu

Patrick Schone
Department of Defense
Fort George G. Meade, MD 20755-6000
pjschon@tycho.ncsc.mil

James Mayfield
JHU Applied Physics Laboratory
11100 Johns Hopkins Road
Laurel, MD 20723-6099, USA
james.mayfield@jhuapl.edu

Abstract

Lexical mismatch is a problem that confounds automatic question answering systems. While existing lexical ontologies such as WordNet have been successfully used to match verbal synonyms (e.g., *beat* and *defeat*) and common nouns (*tennis* is-a *sport*), their coverage of proper nouns is less extensive. Question answering depends substantially on processing named entities, and thus it would be of significant benefit if lexical ontologies could be enhanced with additional hypernymic (i.e., is-a) relations that include proper nouns, such as *Edward Teach* is-a *pirate*. We demonstrate how a recently developed statistical approach to mining such relations can be tailored to identify named entity hyponyms, and how as a result, superior question answering performance can be obtained. We ranked candidate hyponyms on 75 categories of named entities and attained 53% mean average precision. On TREC QA data our method produces a 9% improvement in performance.

1 Introduction

To correctly extract answers, modern question answering systems depend on matching words between questions and retrieved passages containing answers. We are interested in learning hypernymic (i.e., is-a) relations involving named entities because we believe these can be exploited to improve a significant class of questions.

For example, consider the following questions:

- What island produces Blue Mountain coffee?
- In which game show do participants compete based on their knowledge of consumer prices?
- What villain is the nemesis of Dudley Do-Right?

Knowledge that *Jamaica* is an island, that *The Price is Right* is a game show, and that *Snidely Whiplash* is a villain, is crucial to answering these questions.

Sometimes these relations are evident in the same context as answers to questions, for example, in *"The island of Jamaica is the only producer of Blue Mountain coffee"*; however, *"Jamaica is the only producer of Blue Mountain coffee"* should be sufficient, despite the fact that Jamaica is an island is not observable from the sentence.

The dynamic nature of named entities (NEs) makes it difficult to enumerate all of their evolving properties; thus manual creation and curation of this information in a lexical resource such as WordNet (Fellbaum, 1998) is problematic. Pasca and Harabagiu discuss how insufficient coverage of named entities impairs QA (2001). They write:

> "Because WordNet was not designed as an encyclopedia, the hyponyms of concepts such as *composer* or *poet* are illustrations rather than an exhaustive list of instances. For example, only twelve composer names specialize the concept *composer* ... Consequently, the enhancement of WordNet with NE information could help QA."

The chief contribution of this study is demonstrating that an automatically mined knowledge base, which naturally contains errors as well as correctly distilled knowledge, can be used to improve QA performance. In Section 2 we discuss prior work in identifying hypernymic relations. We then explain our methods for improved NE hyponym learning and its evaluation (Section 3) and apply the relations that are discovered to enhance question answering (Section 4). Finally we discuss our results (Section 5) and present our conclusions (Section 6).

2 Hyponym Induction

We review several approaches to learning is-a relations.

2.1 Hearst Patterns

The seminal work in the field of hypernym learning was done by Hearst (1992). Her approach was to identify discriminating lexico-syntactic patterns that suggest hypernymic relations. For example, "X, such as Y", as in *"elements, such as chlorine and fluorine"*.

2.2 KnowItAll

Etzioni et al. developed a system, *KnowItAll*, that does not require training examples and is broadly applicable to a variety of classes (2005). Starting with seed examples generated from high precision generic patterns, the system identifies class-specific lexical and part-of-speech patterns and builds a Bayesian classifier for each category. *KnowItAll* was used to learn hundreds of thousands of class instances and clearly has potential for improving QA; however, it would be difficult to reproduce the approach because of information required for each class (i.e., specifying synonyms such as *town* and *village* for *city*) and because it relies on submitting a large number of queries to a web search engine.

2.3 Query Logs

Pasca and Van Durme looked at learning entity class membership for five high frequency classes (*company*, *country*, *city*, *drug*, and *painter*), using search engine query logs (2007). They reported precision at 50 instances between 0.50 and 0.82.

2.4 Dependency Patterns

Snow et al. have described an approach with several desirable properties: (1) it is weakly-supervised and only requires examples of hypernym/hyponym relations and unannotated text; (2) the method is suitable for both common and rare categories; and, (3) it achieves good performance without post filtering using the Web (2005; 2006). Their method relies on dependency parsing, a form of shallow parsing where each word modifies a single parent word.

Hypernym/hyponym word pairs where the words[1] belong to a single WordNet synset were identified and served to generate training data in the following way: making the assumption that when the two words co-occur, evidence for the is-a relation is present, sentences containing both terms were extracted from unlabeled text. The sentences were parsed and paths between the nouns in the dependency trees were calculated and used as features in a supervised classifier for hypernymy.

3 Learning Named Entity Hyponyms

The present work follows the technique described by Snow et al.; however, we tailor the approach in several ways. First, we replace the logistic regression model with a support vector machine (SVM-Light). Second, we significantly increase the size of training corpora to increase coverage. This beneficially increases the density of training and test vectors. Third, we include additional features not based on dependency parses (e.g., morphology and capitalization). Fourth, because we are specifically interested in hypernymic relations involving named entities, we use a bootstrapping phase where training data consisting primarily of common nouns are used to make predictions and we then manually extract named entity hyponyms to augment the training data. A second learner is then trained using the entity-enriched data.

3.1 Data

We rely on large amounts of text; in all our experiments we worked with a corpus from the sources given in Table 1. Sentences that presented difficulties in parsing were removed and those remaining

[1] Throughout the paper, use of the term *word* is intended to include named entities and other multiword expressions.

Table 1: Sources used for training and learning.

	Size	Sentences	Genre
TREC Disks 4,5	81 MB	0.70 M	Newswire
AQUAINT	1464 MB	12.17 M	Newswire
Wikipedia (4/04)	357 MB	3.27 M	Encyclopedia

Table 2: Characteristics of training sets.

	Pos. Pairs	Neg. Pairs	Total Features
Baseline	7975	63093	162528
+NE	9331	63093	164298
+Feat	7975	63093	162804

were parsed with MINIPAR (Lin, 1998). We extracted 17.3 million noun pairs that co-occurred in at least one sentence. All pairs were viewed as potential hyper/hyponyms.

Our three experimental conditions are summarized in Table 2. The baseline model used 71068 pairs as training data; it is comparable to the weakly-supervised hypernym classifier of Snow et al. (2005), which used only dependency parse features, although here the corpus is larger. The entity-enriched data extended the baseline training set by adding positive examples. The +Feat model uses additional features besides dependency paths.

3.2 Bootstrapping

Our synthetic data relies on hyper/hyponym pairs drawn from WordNet, which is generally rich in common nouns and lacking in proper nouns. But certain lexical and syntactic features are more likely to be predictive for NE hyponyms. For example, it is uncommon to precede a named entity with an indefinite article, and certain superlative adjectives are more likely to be used to modify classes of entities (e.g., "the *youngest* coach", "the *highest* peak"). Accordingly we wanted to enrich our training data with NE exemplars.

By manually reviewing highly ranked predictions of the baseline system, we identified 1356 additional pairs to augment the training data. This annotation took about a person-day. We then rescanned the corpus to build training vectors for these co-occurring nouns to produce the +NE model vectors.

Table 3: Features considered for +Feat model.

Feature	Comment
Hypernym contained in hyponym	*Sands Hotel* is-a *hotel*
Length in chars / words	Chars: 1-4, 5-8, 9-16, 17+ Words: 1, 2, 3, 4, 5, 6, 7+
Has preposition	*Treaty of Paris*; *Statue of Liberty*
Common suffixes	*-ation*, *-ment*, *-ology*, etc...
Figurative term	Such as *goal*, *basis*, or *problem*
Abstract category	Like *person*, *location*, *amount*
Contains digits	Usually not a good hyponym
Day of week; month of year	Indiscriminately co-occurs with many nouns.
Presence and depth in WordNet graph	Shallow hypernyms are unlikely to have entity hyponyms. Presence in WN suggests word is not an entity.
Lexname of 1st synset in WordNet	Root classes like *person*, *location*, *quantity*, and *process*.
Capitalization	Helps identify entities.
Binned document frequency	Partitioned by base 10 logs

3.3 Additional Features

The +Feat model incorporated an additional 276 binary features which are listed in Table 3. We considered other features such as the frequency of patterns on the Web, but with over 17 million noun pairs this was computationally infeasible.

3.4 Evaluation

To compare our different models we created a test set of 75 categories. The classes are diverse and include personal, corporate, geographic, political, artistic, abstract, and consumer product entities. From the top 100 responses of the different learners, a pool of candidate hyponyms was created, randomly reordered, and judged by one of the authors. To assess the quality of purported hyponyms we used average precision, a measure in ranked information retrieval evaluation, which combines precision and recall.

Table 4 gives average precision values for the three models on 15 classes of mixed difficulty[2]. Performance varies considerably based on the hypernym category, and for a given category, by classifier. *N* is the number of known correct instances found in the pool that belong to a given category.

Aggregate performance, as mean average precision, was computed over all 75 categories and is

[2]These are not the highest performing classes

Table 4: Average precision on 15 categories.

	N	Baseline	+NE	+Feat
chemical element	78	0.9096	0.9781	0.8057
african country	48	0.8581	0.8521	0.4294
prep school	26	0.6990	0.7098	0.7924
oil company	132	0.6406	0.6342	0.7808
boxer	109	0.6249	0.6487	0.6773
sculptor	95	0.6108	0.6375	0.8634
cartoonist	58	0.5988	0.6109	0.7097
volcano	119	0.5687	0.5516	0.7722
horse race	23	0.4837	0.4962	0.7322
musical	80	0.4827	0.4270	0.3690
astronaut	114	0.4723	0.5912	0.5738
word processor	26	0.4437	0.4426	0.6207
chief justice	115	0.4029	0.4630	0.5955
perfume	43	0.2482	0.2400	0.5231
pirate	10	0.1885	0.3070	0.2282

Table 5: Mean average precision over 75 categories.

	Baseline	+NE	+Feat
MAP	0.4801	0.5001 (+4.2%)	0.5320 (+10.8%)

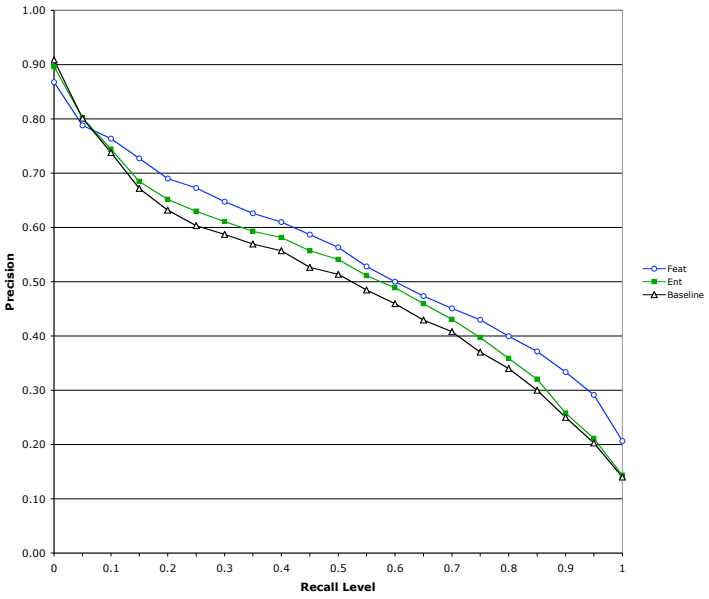

Figure 1: Precision-recall graph for three classifiers.

given in Table 5. Both the +NE and +Feat models yielded improvements that were statistically significant at a 99% confidence level. The +Feat model gained 11% over the baseline condition. The maximum F-score for +Feat is 0.55 at 70% recall.

Mean average precision emphasizes precision at low ranks, so to capture the error characteristics at multiple operating points we present a precision-recall graph in Figure 1. The +NE and +Feat models both attain superior performance at all but the lowest recall levels. For question answering this is important because it is not known which entities will be the focus of a question, so the ability to deeply mine various entity classes is important.

Table 6 lists top responses for four categories.

3.5 Discussion

53% mean average precision seems good, but is it good enough? For automated taxonomy construction precision of extracted hyponyms is critically important; however, because we want to improve question answering we prefer high recall and can tolerate some mistakes. This is because only a small set of passages that are likely to contain an answer are examined in detail, and only from this subset of passages do we need to reason about potential hyponyms. In the next section we describe an experiment which confirms that our learned entity hyponyms are beneficial.

4 QA Experiments

4.1 QACTIS

To evaluate the usefulness of our learned NE hyponyms for question answering, we used the QACTIS system (Schone et al., 2005). QACTIS was fielded at the 2004-2006 TREC QA evaluations and placed fifth at the 2005 workshop. We worked with a version of the software from July 2005.

QACTIS uses WordNet to improve matching of question and document words, and a resource, the Semantic Forest Dictionary (SFD), which contains many hypernym/hyponym pairs. The SFD was populated through both automatic and manual means (Schone et al., 2005), and was updated based on questions asked in TREC evaluations through 2004.

4.2 Experimental Setup

We used factoid questions from the TREC 2005-2006 QA evaluations (Voorhees and Dang, 2005) and measured performance with mean reciprocal rank (MRR) and percent correct at rank 1.

All runs made use of WordNet 2.0, and we examined several other sources of hypernym knowl-

Table 6: Top responses for four categories using the +Feat model. Starred entries were judged incorrect.

	Sculptor	Horse Race	Astronaut	Perfume
1	Evelyn Beatrice Longman	Tevis Cup	Mark L Polansky	* Avishag
2	Nancy Schon	Kenilworth Park Gold Cup	Richard O Covey	Ptisenbon
3	Phidias	Cox Plate	George D Nelson	Poeme
4	Stanley Brandon Kearl	Grosser Bugatti Preis	Guion Bluford Jr	Parfums International
5	Andy Galsworthy	Melbourne Cup	Stephen S Oswald	Topper Schroeder
6	Alexander Collin	* Great Budda Hall	Eileen Collins	* Baccarin
7	Rachel Feinstein	Travers Stakes	Leopold Eyharts	Pink Lady
8	Zurab K Tsereteli	English Derby	Daniel M Tani	Blue Waltz
9	Bertel Thorvaldsen	* Contrade	Ronald Grabe	WCW Nitro
10	Cildo Meireles	Palio	* Frank Poole	Jicky

Table 7: Additional knowledge sources by size.

	Classes	Class Instances
Baseline	76	11,066
SFD	1,140	75,647
SWN	7,327	458,370
+Feat	44,703	1,868,393

edge. The baseline condition added a small subset of the Semantic Forest Dictionary consisting of 76 classes seen in earlier TREC test sets (e.g., nationalities, occupations, presidents). We also tested: (1) the full SFD; (2) a database from the Stanford Wordnet (SWN) project (Snow et al., 2006); and, (3) the +Feat model discussed in Section 3. The number of classes and entries of each is given in Table 7.

4.3 Results

We observed that each source of knowledge benefitted questions that were incorrectly answered in the baseline condition. Examples include learning a meteorite (Q84.1), a university (Q93.3), a chief operating officer (Q108.3), a political party (Q183.3), a pyramid (Q186.4), and a movie (Q211.5).

In Table 8 we compare performance on questions from the 2005 and 2006 test sets. We assessed performance primarily on test questions that were deemed likely to benefit from hyponym knowledge – questions that had a readily discernible category (e.g., *"What film ..."*, *"In what country ..."*) – but we also give results on the entire test set.

The WordNet-only run suffers a large decrease compared to the baseline. This is expected because WordNet lacks coverage of entities and the baseline condition specifically populates common categories of entities that have been observed in prior TREC evaluations. Nonetheless, WordNet is useful to the system because it addresses lexical mismatch that does not involve entities.

The full SFD, the SWN, and the +Feat model achieved 17%, 2%, and 9% improvements in answer correctness, respectively. While no model had exposure to the 2005-2006 TREC questions, the SFD database was manually updated based on training on the TREC-8 through TREC-2004 data sets. It approximates an upper bound on gains attributable to addition of hyponym knowledge: it has an unfair advantage over the other models because recent question sets use similar categories to those in earlier TRECs. Our +Feat model, which has no bias towards TREC questions, realizes larger gains than the SWN. This is probably at least in part because it produced a more diverse set of classes and a significantly larger number of class instances. Compared to the baseline condition the +Feat model sees a 7% improvement in mean reciprocal rank and a 9% improvement in correct first answers; both results represent a doubling of performance compared to the use of WordNet alone. We believe that these results illustrate clear improvement attributable to automatically learned hyponyms.

The rightmost columns in Table 8 reveal that the magnitude of improvements, when measured over all questions, is less. But the drop off is consistent with the fact that only one third of questions have clear need for entity knowledge.

5 Discussion

Although there is a significant body of work in automated ontology construction, few researchers have examined the relationship between their methods

Table 8: QA Performance on TREC 2005 & 2006 Data

	Hyponym-Relevant Subset (242)		All Questions (734)	
	MRR	**% Correct**	**MRR**	**% Correct**
WN-alone	0.189 (-45.6%)	12.8 (-51.6%)	0.243 (-29.0%)	18.26 (-30.9%)
Baseline	0.348	26.4	0.342	26.4
SFD	0.405 (+16.5%)	31.0 (+17.2%)	0.362 (+5.6%)	27.9 (+5.7%)
SWN	0.351 (+1.0%)	26.9 (+1.6%)	0.343 (+0.3%)	26.6 (+0.5%)
Feat	0.373 (+7.4%)	28.9 (+9.4%)	0.351 (+2.5%)	27.3 (+3.1%)

for knowledge discovery and improved question-answering performance. One notable study was conducted by Mann (2002). Our work differs in two ways: (1) his method for identifying hyponyms was based on a single syntactic pattern, and (2) he looked at a comparatively simple task – given a question and one answer sentence containing the answer, extract the correct named entity answer.

Other attempts to deal with lexical mismatch in automated QA include rescoring based on syntactic variation (Cui et al., 2005) and identification of verbal paraphrases (Lin and Pantel, 2001).

The main contribution of this paper is showing that large-scale, weakly-supervised hyponym learning is capable of producing improvements in an end-to-end QA system. In contrast, previous studies have generally presented algorithmic advances and showcased sample results, but failed to demonstrate gains in a realistic application. While the hypothesis that discovering is-a relations for entities would improve factoid QA is intuitive, we believe these experiments are important because they show that automatically distilled knowledge, even when containing errors that would not be introduced by human ontologists, is effective in question answering systems.

6 Conclusion

We have shown that highly accurate statistical learning of named entity hyponyms is feasible and that bootstrapping and feature augmentation can significantly improve classifier accuracy. Mean average precision of 53% was attained on a set of 75 categories that included many fine-grained entity classes. We also demonstrated that mining knowledge about entities can be directly applied to question answering, and we measured the benefit on TREC QA data. On a subset of questions for which NE hyponyms are likely to help we found that learned hyponyms generated a 9% improvement in performance compared to a strong baseline.

References

Hang Cui, Renxu Sun, Keya Li, Min-Yen Kan, and Tat-Seng Chua. 2005. Question answering passage retrieval using dependency relations. In *SIGIR 2005*, pages 400–407.

Oren Etzioni, Michael Cafarella, Doug Downey, Ana M. Popescu, Tal Shaked, Stephen Soderland, Daniel S. Weld, and Alexander Yates. 2005. Unsupervised Named-Entity Extraction from the Web: An Experimental Study. *Artificial Intelligence*, 165(1):191–134.

Christine Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. MIT Press.

Marti A. Hearst. 1992. Automatic acquisition of hyponyms from large text corpora. In *ACL 1992*, pages 539–545.

Dekang Lin and Patrick Pantel. 2001. Discovery of inference rules for question-answering. *Natural Language Engineering*, 7(4):343–360.

Dekang Lin. 1998. Dependency-based evaluation of minipar. In *Workshop on the Evaluation of Parsing Systems*.

Gideon S. Mann. 2002. Fine-grained proper noun ontologies for question answering. In *COLING-02 on SEMANET*, pages 1–7.

Marius Pasca and Benjamin Van Durme. 2007. What you seek is what you get: Extraction of class attributes from query logs. In *IJCAI-07*, pages 2832–2837.

Marius Pasca and Sanda M. Harabagiu. 2001. The informative role of wordnet in open-domain question answering. In *Proceedings of the NAACL 2001 Workshop on WordNet and Other Lexical Resources*.

Patrick Schone, Gary Ciany, Paul McNamee, James Mayfield, and Thomas Smith. 2005. QACTIS-based Question Answering at TREC 2005. In *Proceedings of the 14th Text REtrieval Conference*.

Rion Snow, Daniel Jurafsky, and Andrew Y. Ng. 2005. Learning syntactic patterns for automatic hypernym discovery. In *NIPS 17*.

Rion Snow, Daniel Jurafsky, and Andrew Y. Ng. 2006. Semantic taxonomy induction from heterogenous evidence. In *ACL 2006*, pages 801–808.