Extracting Meronyms for a Biology Knowledge Base Using Distant Supervision

ABSTRACT

Knowledge of objects and their parts, meronym relations, are at the heart of many question-answering systems, but manually encoding these facts is impractical. Past researchers have tried hand-written patterns, supervised learning, and bootstrapped methods, but achieving both high precision and recall has proven elusive. This paper reports on a thorough exploration of distant supervision to learn a meronym extractor for the domain of college biology. We introduce a novel algorithm, generalizing the "at least one" assumption of multi-instance learning to handle the case where a fixed (but unknown) percentage of bag members are positive examples. Detailed experiments compare strategies for mention detection, negative example generation, leveraging out-of-domain meronyms, and evaluate the benefit of our multi-instance percentage model.

1. INTRODUCTION

We are motivated by the vision of *Digital Aristotle*, specifically an interactive knowledge-based system that can answer a wide range of questions about scientific topics like college biology [4]. Knowledge of object part-whole relations, so called meronyms, is central to many questions. For example, in order to answer the question "What part of a cell allows for selective permiability?" it's important to know that the plasma membrane is part of all cells. However it is impractical to manually encode all meronym relations, so we seek to extract them automatically from text.

The problem of meronym extraction has been studied for many years. For example, Berland and Charniak [1] applied two Hearst patterns [5] to find candidate meronyms and then used corpus statistics to order them. Girju et al. [3] added a verbal pattern and leveraged decision trees to learn semantic constraints over the part/whole's WordNet classes. While these efforts yielded promising precision, Berland and Charniak's experiments were limited to finding the parts of five specific objects. Furthermore, Girju et al.'s method required thousands of manually-labeled sentences in order to train a reliable classifier, and this effort would likely need to be repeated for different technical domains such as biology. Minimally supervised approaches have also been explored. Espresso used bootstrap learning and Web statistics to learn meronyms [14]. Ittoo and Bouma extended the method by independently bootstrapping different subtypes of the meronym relation (e.g., located-in vs member-of) [7]. Precision was high (0.67–0.82) on the most confident 500 predictions, but recall was not evaluated.

Distant supervision is a promising alternative, which has seen considerable success for relation extraction [9, 22, 12]. The basic idea is to use a database table, R, to automatically create a training set by heuristically labeling a sentence as a positive example if it contains mentions of two entities that match a tuple in R. While the resulting training set is usually quite noisy, it requires no human labeling effort. To cope with incorrectly labeled examples, Bunescu and Mooney suggest using *multiple instance learning* [2] and several authors have proposed graphical models that encode the assumption that *at least one* of the sentences matching each tuple is a true positive example [16, 6, 20]. In this paper, we generalize this model to one which assumes a certain *percentage* of matches (apriori unknown and potentially different for each relation) are true positives. We extend Hoffmann et al.'s graphical model [6] to enforce percentage constraints, train using perceptron updates, and use grid search via cross-validation to find the best percentage for a relation. In summary, this paper makes the following contributions:

- We introduce a novel generalization of multi-instance learning from the "at least one" assumption to handle the case where a fixed (but unknown) percentage of bag members are positive examples.
- We present a detailed evaluation of the efficacy of distant supervision for the problem of extracting meronyms from biological text, specifically comparing strategies for 1) mention detection, 2) negative example generation, 3) multi-instance percentage, and 4) leveraging out-of-domain meronyms.

2. OUR DISTANT SUPERVISION MODEL

As input distantly supervised extration learning requires three inputs: 1) a set of *target relations* $\mathcal{R}$ (and a null relation NA); 2) a knowledge base $\Delta = \{(a_1, R, a_2)\}$ where $R \in \mathcal{R}$ (i.e. a triple store); (3) an unlabeled textual corpus Σ. Extractors are typically learned in two stages.

Data Preparation: Given the knowledge base, Δ, and the unlabeled corpus, first identify all mentions of entities

in each sentence[1]. Next one enumerates all pairs of entities (i.e. (e_1, e_2)) in each sentence $s \in \Sigma$ and then groups these examples so each entity pair is associated with a set of sentences that mention both entities. Finally, for each entity pair and target relation $R \in \mathcal{R}$, check to see if the ground fact (e_1, R, e_2) is in Δ; if so call the quadruple (e_1, R, e_2, s) a *fact mention*. Clearly there is no guarantee that every fact mention for (e_1, R, e_2) is actually expressing R (in fact, *none* of the sentences may state R). Bunescu and Mooney [2] suggested modeling this in terms of *multi-instance learning*(MIL), where the training set consists of positive and negative *bags* of instances. One may assume that *at least one* of the instances in each positive bag is a true positive instance and that *every* instance in a negative bag is a true negative instance. For distantly supervised extractor learning, we form a positive bag from the union of all mentions of each mentioned *has-part* fact: $\{\ldots, (e_1, R, e_2, s_i), \ldots\}$. Negative bags are formed by taking (some of) the (e_1, e_2, s) matches where Δ contained no corresponding fact and are denoted $\{\ldots, (e_1, NA, e_2, s_i), \ldots\}$. All instances are represented as a vector of features, such as those used by Mintz et al [12].

Training the Extractor: From the set of positive and negative bags, one must learn a classifier that will take a new example (e_1, e_2, s) and predict whether NA or a relation $R \in \mathcal{R}$ holds. For example, MULTIR [6] uses structural perceptron-style additive updates to train a log-linear model using a simple graphical model that aggregates sentence-level predictions with a deterministic OR of the predicted relation types. For meronym extraction, there is only one target relation: $\mathcal{R} = \{\textit{has-part}\}$.

HANDLING PERCENTAGE CONSTRAINTS

So far we have made the common assumption that *at least one* instance is positive in each positive bag. One consequence of this assumption is that during MULTIR's perceptron learning, only a few instances cause updates even if the true positive instances are abundant in some bags. We propose an extension to MULTIR that allows more updates by tuning a parameter p[2].

For each bag, let y be its bag label (*has-part* or *NA*), $z = \{z_i\}$ be the instance labels and $x = \{x_i\}$ be the instances. In MULTIR, $p(z, y|x) \triangleq p(z|x)p(y|z)$. The first term is defined as $p(z|x) = \prod_i p(z_i|x_i)$ where $p(z_i|x_i)$ is a log-linear model. Predicting y given z(the second term) is deterministic. In other words, $y =$*has-part* is predicted if and only if $\exists i, z_i =$*has-part*; $y =$*NA* otherwise.

In order to handle the percentage constraint that *strictly more than p%* of a bag's instances be positive, we make the following modification. Instead of predicting $y =$*has-part* using deterministic OR (i.e. seeing only one $z_i =$*has-part*), we predict $y =$*has-part* when strictly more than p% of all the instances have a predicted label *has-part*.

During learning, when a bag's label prediction, $\hat{y}$, disagrees with the true bag label, the conditional probability $p(z|\hat{y}, x)$ is computed to determine the most probable values of z_is. The exact solution is obtained via maximum weighted bipartite matching. Pursuing efficiency, MULTIR uses an effective approximation: when the bag label is *has-part*, assigning that label to the instance with the highest score under the current linear model. However, it is likely that the other instances with high scores are also truly positive. Thus, we propose to instead assign the label *has-part* to all of the instances with the highest scores in a positive bag, until strictly more than p% denote *has-part*.

Note that MULTIR is a special case of our algorithm when $p = 0$. In both situations, at least one $z_i =$*has-part* is equivalent to more than 0% of the instances.

The discussion so far has assumed that one knows the actual percentage of positive instances in a postive bag, but p varies from relation to relation and also depends on the textual corpus in question. To find the best value of p, therefore, our algorithm performs a grid search considering different values and chooses the best using cross validation.

3. EXPERIMENTAL FRAMEWORK

Knowledge Base Δ: We start with the ontology created in Project Halo [4], which includes a seed database of meronyms from the biology domain. This includes 640 context-free, universally quantified *has-part* facts, labeled by human annotators and another 3179 *has-part* facts having some sort of context. For instance, one KB expression might state that a *plant-cell has-part chloroplast* with some qualifications such as the plant cells being photosynthetic, etc.

Entity Identifier: In order to identify mentions of biological entities in sentences, we use a simple procedure that automatically marks substrings by taking the longest possible match from a compiled dictionary of biological terms. In the dictionary, each concept is associated with a few different lexical phrases (e.g. "cell of plant" is recognized as *plant-cell*). The entity types are then determined by a curated ontology (e.g., *plant-cell* has the type *living-entity*; in total there are nine types).

Unlabeled Corpus Σ: We use a digitized version of a popular high school biology textbook [15] comprised of 41,892 sentences in 56 chapters. For syntactic analysis, we ran the Stanford CoreNLP pipeline [8], substituting the parser by Charniak-Johnson using a biomedical model [10]. In the whole data set used in the following experiments, around 20 thousand sentences are used as fact mentions for roughly 4 thousand facts.

Features: We use the commonly adopted "Mintz features" [13], which include features such as the word sequence between two entity arguments as well as the dependency path connecting them. Since overfitting was a concern, we experimented with filtering features that occurred fewer than k times in the training set, but to our surprise this reduced F1 scores for all attempted values of k.

Evaluation: A held-out test set was set up and validated by two human annotators at the start of the project, consisting of 172 *has-part* facts and 206 *NA* facts. In most experiments, we report precision and recall for k-fold cross-validation ($k = 5$) over the training set as well as performance on the test set.

In the following sections, we evaluate various versions of the distant supervision process. In each case, when evaluating one aspect, we run the experiment using the best combination of other aspects. For example, when considering the use of coreference in mention detection, we use the best setting for creating negative examples.

[1]A common approach on news articles uses named-entity recognition (NER) to find the entity mentions and simple string match to disambiguate the entities; a more sophisticated approach might use named-entity linking (NEL).

[2]For simplicity, we describe the extension for one single relation.

4. EXPANDING MENTIONS WITH COREF

As mentioned in the previous section, in the baseline condition entities are identified using simple string match against a precompiled dictionary. When two mentions of the same entities appear in the same sentence, only one mention is used depending on the token distance from the mention of the other argument (the shorter one is chosen). However, sometimes the closest mention of an entity with respect to the other is a pronoun or nominal rather than its full name. For instance, consider the task of extracting *(X-Chromosome,* has-part*, Gene)* from the sentence "One of the sex-determining chromosomes is X chromosome and it has around 1,100 genes.", the pronoun "it" that refers to "X chromosome" has a more direct syntactic connection with "genes". The best mentions of a concept are chosen based on the length of their dependency path. Besides pronouns, we also found that partitive nouns are often used in the corpus. Only "nucleitide bases" will be identified as a concept in the phrase "a sequence of nucleitide bases". We manually selected 29 partitive nouns (e.g. collection, pair, etc) and expand mentions like the previous example to their whole phrases including the partitive nouns.

	Recall	Precision	F1
CV	0.664	0.820	0.733
CV+COREF	**0.674**	**0.821**	**0.740**
TEST	0.663	**0.857**	0.748
TEST+COREF	**0.744**	0.795	**0.769**

Table 1: An accurate mention detection lifts the performance.

The top two rows of Table 1 show the 5-fold cross-validation (CV) results on the training set. The following two rows display the results on the held-out test set (TEST). Using coreference to improve mention selection yields a small, but definite imrpovement.

5. GENERATING NEGATIVE EXAMPLES

While the distant supervision framework is clear about how to generate positive examples, negative examples are not so straightforward. If one knew that the underlying knowledge base were complete, then one could add every sentence whose an entity pair failed to match a fact as a negative example. Of course, if one had a complete knowledge base, one would not need to build an extractor in the first place. There is also the issue of skew — increasing the number of negative examples increases precision and reduces recall. Depending on one's objective, varying points along this tradeoff may be desired.

In our dataset, 887 out of 3819 *has-part* facts have at least one corresponding sentence that mentions both arguments. These form the positive bags. Another 105 argument pairs are manually labeled as definitively not having a *has-part* relation; of these 105 pairs, 77 have matching sentences, which are taken as negative examples. We call this setup "BASE".

To increase the number of negative examples for training, we exploit the irrelecive nature of *has-part* and create a negative instance by swapping the arguments of each positive *has-part* fact: if $(e_1, \textit{has-part}, e_2)$ holds, so will (e_2, NA, e_1). This heuristic(REV) yields an additional 887 negative instances. Furthermore, a transitive closure(TRANS) of existing negative relations lifts the number of negative facts to 2566.[3]

	Recall	Precision	F1	#+	#-
CV(BASE)	**0.856**	0.389	0.533	887	77
CV(REV)	0.706	0.748	0.725	887	963
CV(TRANS)	0.674	**0.821**	**0.740**	887	2566
TEST(BASE)	**0.884**	0.596	0.712	887	77
TEST(REV)	0.709	0.718	0.713	887	963
TEST(TRANS)	0.744	**0.795**	**0.769**	887	2566

Table 2: Performance on different negative data. "#+" indicates the number of *has-part* facts in the training set; "#-" indicates the number of *NA* facts.

Table 2 shows the results. Since additional negative examples improve precision at the expense of recall, it is unsurprising that BASE has the highest recall in both evaluations, while TRANS has the highest precision. However, it is notable that TRANS also has the highest F1 scores by a sizable amount.

6. OPTIMIZING % IN POSITIVE BAGS

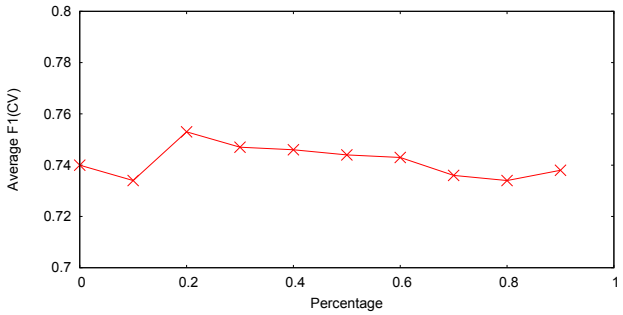

Figure 1: Grid search of an optimal p

The actual percentage of instances in a positive bag that truly denote the corresponding fact is a function of both the relation and the corpus. Without knowing the value of p a priori, our learner picks the best value based on the results of automated cross validation experiments. Figure 1 shows the average F1 values discovered by a grid search from 0.0 to 0.9 with a 0.1 increment on p. Because the grid search finds a peak in F1 at $p = 0.2$, that value is used when evaluating on the held-out test set (Table 3). Our approach for finding and exploiting percentage constraints results in a small, but distinct improvement in F1.

7. LEVERAGING OUT-OF-DOMAIN MERONYMS

The manner in which meronym relations are expressed varies from domain to domain. In the case of the biology domain, we were able to use distant supervision to exploit the nascent, manually-created knowledge-base. Given that WordNet [11] encodes a huge number of part-whole relations

[3]Surdeanu *et al.* [19] and Sun *et al.* [18] propose only taking *(a, R, c)* as a negative example if *(a, R, b)* is a positive fact for $b \neq c$, but this heuristic depends on the relation being functional, and an entity can have many parts.

	Recall	Precision	F1
CV($p = 0$)	0.674	**0.821**	0.740
CV($p = 0.2$)	**0.730**	0.776	**0.753**
TEST($p = 0$)	0.744	**0.795**	0.769
TEST($p = 0.2$)	**0.791**	0.786	**0.788**

Table 3: Performance comparing MultiR and MultiR with percentage.

over synset pairs, it is natural to wonder if these general facts could be exploited to increase the performance in the biology domain.

After some experimentations, we chose only *substance* and *part* meronyms (not *member* meronyms). Furthermore, we excludes meronym pairs involving a location or a named entity. Initial efforts of matching these facts against a broad corpus led to dramatic performance drops, so we restrict the matching corpus to Wikipedia articles. Unfortunately, WordNet synsets use a set of semantic types that are incompatible with those in the Halo biology ontology. Therefore, we make the features type-free, e.g.

$$\textit{Living-Entity} - [nsubj] \rightarrow \textit{have} \leftarrow [dobj] - \textit{Organ}$$

is reduced to

$$\textit{ARG1} - [nsubj] \rightarrow \textit{have} \leftarrow [dobj] - \textit{ARG2}$$

We compare the results on the test set when training the model with(TEST+WN) or without(TEST) additional WordNet meronyms (around 700 ground facts). We also show the results when different percentage values are set ($p = 0$ or $p = 0.2$). While the experiment (Table 4) shows small improvement due to the additional training instances, the gains are small. Inspection of the data makes it clear that the absence of type features exerts a significant hit on performance. However, we believe that the results suggest that with further effort on ontological alignment there would be bigger benefits from the additional examples.

	p	Recall	Prec	F1	#+	#-
TEST	0.0	0.384	**0.892**	0.537	887	2566
TEST+WN	0.0	**0.401**	0.841	**0.543**	1578	2566
TEST	0.2	0.523	**0.811**	0.636	887	2566
TEST+WN	0.2	**0.558**	0.793	**0.655**	1578	2566

Table 4: Adding additional out-of domain meronym facts during training yields small improvements despite the lack of type features.

8. COMPARED SYSTEMS

We also evaluate the results in the context of other available systems, which are listed below:

- Pattern-based (PAT): Our implementation of 2 Hearst patterns (*whole* `'s` *part* and *part* `of` *whole*) and 2 verbal patterns (e.g. *part* `form` *whole*, *whole* `consist of` *part*).
- Always Predict *has-part* (AP): For any entity pair, this baseline always predicts *has-part*.
- Entity type based (TYPE): The relation type for each entity pair is determined by the most frequent relation given the entity type pair in the training data.
- MINTZ [13]: Treat each instance in a positive bag as a positive instance and each instance in a negative bag as a negative instance. A linear SVM model is trained on the artificially labeled instances. The bag label is predicted by deterministically OR-ing individual instance predictions.

TEST	Recall	Precision	F1
PAT	0.465	**0.847**	0.601
AP	**1.000**	0.455	0.645
TYPE	0.709	0.777	0.742
MINTZ	0.849	0.582	0.690
MULTIR	0.744	0.795	0.769
MULTIR(p)	0.791	0.786	**0.788**

Table 5: MULTIR(p) is the system when the percentage value is 0.2**.**

Table 5 shows that our distantly supervised relation extractor is able to achieve reasonably high numbers in both precision and recall among all other systems. Moreover, our proposed extension beyond the "at least one" assumption further lifts the performance.

9. RELATED WORK AND CONCLUSION

In addition to the work described in the introduction, we note that Wang et al.'s p-posterior mixture-model kernels [21] are similar to our density model for positive bags. Our methods differ in two respects, however. First, they seek to predict a label for a complete bag not for instances within the bag; this is akin to the distinction between aggregate and sentential extraction. Secondly, our approaches use dramatically different mechanisms — SVMs vs perceptron updates. Our work is also related to that of Snow et al., who present a probabilistic joint model for taxonomy induction [17].

Given the importance of meronyms in question answering, extraction of these relations from text deserves more attention. This paper has presented a thorough exploration of distant supervision to this task, comparing various strategies for mention detection, negative example generation, and example transfer. In addition, we presented a novel generalization of multi-instance learning that replaces the "at least one" assumption to handle the case where an unknown but fixed percentage of bag members are positive examples. Our results show that the approach for finding and exploiting percentage constraints results in a small, but distinct improvement in F1.

10. REFERENCES

[1] M. Berland and E. Charniak. Finding parts in very large corpora. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, pages 57–64. Association for Computational Linguistics, 1999.

[2] R. Bunescu and R. Mooney. Learning to extract relations from the web using minimal supervision. In *In Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL-07*, 2007.

[3] R. Girju, A. Badulescu, and D. Moldovan. Learning semantic constraints for the automatic discovery of part-whole relations. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, pages 1–8. Association for Computational Linguistics, 2003.

[4] D. Gunning, V. Chaudhri, P. Clark, Y. Chaw, M. Greaves, B. Grosof, A. Leung, D. McDonald, S. Mishra, B. Pacheco, A. Spaulding, D. Tecuci, and J. tien. Project halo update — progress toward digital aristotle. *AI Magazine*, 31(3):33–58, 2010.

[5] M. Hearst. Automatic Acquisition of Hyponyms from Large Text Corpora. In *Procs. of the 14th International Conference on Computational Linguistics*, pages 539–545, Nantes, France, 1992.

[6] R. Hoffmann, C. Zhang, X. Ling, L. Zettlemoyer, and D. S. Weld. Knowledge-based weak supervision for information extraction of overlapping relations. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, volume 1, pages 541–550, 2011.

[7] A. Ittoo and G. Bouma. On learning subtypes of the part-whole relation: do not mix your seeds. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1328–1336. Association for Computational Linguistics, 2010.

[8] H. Lee, A. Chang, Y. Peirsman, N. Chambers, M. Surdeanu, and D. Jurafsky. Deterministic coreference resolution based on entity-centric, precision-ranked rules. *Computational Linguistics*, (Just Accepted):1–54, 2013.

[9] T. Lengauer, R. Schneider, P. Bork, D. L. Brutlag, J. I. Glasgow, H.-W. Mewes, and R. Zimmer, editors. *Constructing Biological Knowledge Bases by Extracting Information from Text Sources*. AAAI, 1999.

[10] D. McClosky and E. Charniak. Self-training for biomedical parsing. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Short Papers*, pages 101–104. Association for Computational Linguistics, 2008.

[11] G. Miller. WordNet: an on-line lexical database. *International Journal of Lexicography*, 3(4):235–312, 1991.

[12] M. Mintz, S. Bills, R. Snow, and D. Jurafsky. Distant supervision for relation extraction without labeled data. In *ACL-IJCNLP*, 2009.

[13] M. Mintz, S. Bills, R. Snow, and D. Jurafsky. Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2-Volume 2*, pages 1003–1011. Association for Computational Linguistics, 2009.

[14] P. Pantel and M. Pennacchiotti. Espresso: Leveraging generic patterns for automatically harvesting semantic relations. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 113–120. Association for Computational Linguistics, 2006.

[15] J. Reece, L. Urry, M. Cain, S. Wasserman, P. Minorsky, P. Jackson, and N. Campbell. In *Campbell Biology*. Benjamin Cummings, 2010.

[16] S. Riedel, L. Yao, and A. McCallum. Modeling relations and their mentions without labeled text. In *Proceedings of the Sixteenth European Conference on Machine Learning (ECML-2010)*, pages 148–163, 2010.

[17] R. Snow, D. Jurafsky, and A. Y. Ng. Semantic taxonomy induction from heterogenous evidence. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 801–808. Association for Computational Linguistics, 2006.

[18] A. Sun, R. Grishman, W. Xu, and B. Min. New york university 2011 system for kbp slot filling. In *Proceedings of the Text Analytics Conference*, 2011.

[19] M. Surdeanu, S. Gupta, J. Bauer, D. McClosky, A. X. Chang, V. I. Spitkovsky, and C. D. Manning. StanfordâĂŹs distantly-supervised slot-filling system. In *Proceedings of the Text Analytics Conference*, 2011.

[20] M. Surdeanu, J. Tibshirani, R. Nallapati, and C. D. Manning. Multi-instance multi-label learning for relation extraction. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, 2012.

[21] H.-Y. Wang, Q. Yang, and H. Zha. Adaptive p-posterior mixture-model kernels for multiple instance learning. In *Proceedings of the 25th international conference on Machine learning*, ICML '08, pages 1136–1143, New York, NY, USA, 2008. ACM.

[22] F. Wu and D. Weld. Autonomously semantifying Wikipedia. In *Proceedings of the ACM Sixteenth Conference on Information and Knowledge Management (CIKM-07)*, Lisbon, Porgugal, 2007.