Filling Knowledge Base Gaps for Distant Supervision of Relation Extraction

Wei Xu[+] Raphael Hoffmann[^] Le Zhao[#,*] Ralph Grishman[+]
[+]New York University, New York, NY, USA
{xuwei, grishman}@cs.nyu.edu
[^]University of Washington, Seattle, WA, USA
raphaelh@cs.washington.edu
[#]Google Inc., Mountain View, CA, USA
lezhao@google.com

Abstract

Distant supervision has attracted recent interest for training information extraction systems because it does not require any human annotation but rather employs existing knowledge bases to heuristically label a training corpus. However, previous work has failed to address the problem of false negative training examples mislabeled due to the incompleteness of knowledge bases. To tackle this problem, we propose a simple yet novel framework that combines a passage retrieval model using coarse features into a state-of-the-art relation extractor using multi-instance learning with fine features. We adapt the information retrieval technique of pseudo-relevance feedback to expand knowledge bases, assuming entity pairs in top-ranked passages are more likely to express a relation. Our proposed technique significantly improves the quality of distantly supervised relation extraction, boosting recall from 47.7% to 61.2% with a consistently high level of precision of around 93% in the experiments.

1 Introduction

A recent approach for training information extraction systems is distant supervision, which exploits existing knowledge bases instead of annotated texts as the source of supervision (Craven and Kumlien, 1999; Mintz et al., 2009; Nguyen and Moschitti, 2011). To combat the noisy training data produced by heuristic labeling in distant supervision, researchers (Bunescu and Mooney, 2007; Riedel et al., 2010; Hoffmann et al., 2011; Surdeanu et al., 2012) exploited multi-instance learning models. Only a few studies have directly examined the influence of the quality of the training data and attempted to enhance it (Sun et al., 2011; Wang et al., 2011; Takamatsu et al., 2012). However, their methods are handicapped by the built-in assumption that a sentence does not express a relation unless it mentions two entities which participate in the relation in the knowledge base, leading to false negatives.

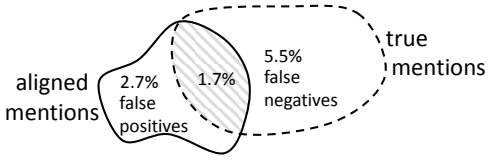

Figure 1: Noisy training data in distant supervision

In reality, knowledge bases are often incomplete, giving rise to numerous false negatives in the training data. We sampled 1834 sentences that contain two entities in the New York Times 2006 corpus and manually evaluated whether they express any of a set of 50 common Freebase[1] relations. As shown in Figure 1, of the 133 (7.3%) sentences that truly express one of these relations, only 32 (1.7%) are covered by Freebase, leaving 101 (5.5%) false negatives. Even for one of the most complete relations in Freebase, *Employee-of* (with more than 100,000 entity pairs), 6 out of 27 sentences with the pattern 'PERSON executive of ORGANIZATION' contain a fact that is not included in Freebase and are thus mislabeled as negative. These mislabelings dilute the discriminative capability of useful features and confuse the models. In this paper, we will show how reducing this source of noise can significantly improve the performance of distant supervision. In fact, our system corrects the relation labels of the above 6 sentences before training the relation extractor.

[*]This work was done while Le Zhao was at Carnegie Mellon University.

[1]http://www.freebase.com

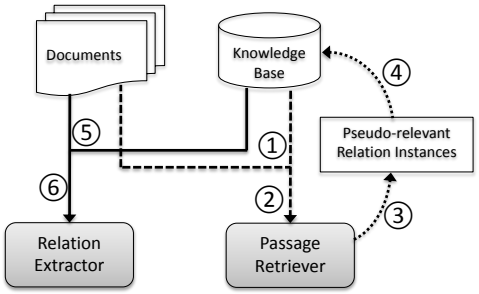

Figure 2: Overall system architecture: The system (1) matches relation instances to sentences and (2) learns a passage retrieval model to (3) provide relevance feedback on sentences; Relevant sentences (4) yield new relation instances which are added to the knowledge base; Finally, instances are again (5) matched to sentences to (6) create training data for relation extraction.

Encouraged by the recent success of simple methods for coreference resolution (Raghunathan et al., 2010) and inspired by pseudo-relevance feedback (Xu and Croft, 1996; Lavrenko and Croft, 2001; Matveeva et al., 2006; Cao et al., 2008) in the field of information retrieval, which expands or reformulates query terms based on the highest ranked documents of an initial query, we propose to increase the quality and quantity of training data generated by distant supervision for information extraction task using pseudo feedback. As shown in Figure 2, we expand an original knowledge base with possibly missing relation instances with information from the highest ranked sentences returned by a passage retrieval model (Xu et al., 2011) trained on the same data. We use coarse features for our passage retrieval model to aggressively expand the knowledge base for maximum recall; at the same time, we exploit a multi-instance learning model with fine features for relation extraction to handle the newly introduced false positives and maintain high precision.

Similar to iterative bootstrapping techniques (Yangarber, 2001), this mechanism uses the outputs of the first trained model to expand training data for the second model, but unlike bootstrapping it does not require iteration and avoids the problem of semantic drift. We further note that iterative bootstrapping over a single distant supervision system is difficult, because state-of-the-art systems (Surdeanu et al., 2012; Hoffmann et al., 2011; Riedel et al., 2010; Mintz et al., 2009), detect only few false negatives in the training data due to their high-precision low-recall features, which were originally proposed by Mintz et al. (2009). We present a reliable and novel way to address these issues and achieve significant improvement over the MULTIR system (Hoffmann et al., 2011), increasing recall from 47.7% to 61.2% at comparable precision. The key to this success is the combination of two different views as in co-training (Blum and Mitchell, 1998): an information extraction technique with fine features for high precision and an information retrieval technique with coarse features for high recall. Our work is developed in parallel with Min et al. (2013), who take a very different approach by adding additional latent variables to a multi-instance multi-label model (Surdeanu et al., 2012) to solve this same problem.

2 System Details

In this section, we first introduce some formal notations then describe in detail each component of the proposed system in Figure 2.

2.1 Definitions

A *relation instance* is an expression $r(e1, e2)$ where r is a binary *relation*, and $e1$ and $e2$ are two entities having such a relation, for example *CEO-of(Tim Cook, Apple)*. The knowledge-based distant supervised learning problem takes as input (1) Σ, a training corpus, (2) E, a set of entities mentioned in that corpus, (3) R, a set of relation names, and (4) Δ, a set of ground facts of relations in R. To generate our training data, we further assume (5) T, a set of entity types, as well as type signature $r(E1, E2)$ for relations.

We define the positive data set $POS(r)$ to be the set of sentences in which any related pair of entities of relation r (according to the knowledge base) is mentioned. The negative data set $RAW(r)$ is the rest of the training data, which contain two entities of the required types in the knowledge base, e.g. one person and one organization for the *CEO-of* relation in Freebase. Another negative data set with more conservative sense $NEG(r)$ is defined as the set of sentences which contain the primary entity $e1$ (e.g. person in any *CEO-of* relation in the knowledge base) and any secondary entity $e2$ of required type (e.g. organization for the *CEO-of* relation) but the relation does not hold for this pair of entities in the knowledge base.

2.2 Distantly Supervised Passage Retrieval

We extend the learning-to-rank techniques (Liu, 2011) to distant supervision setting (Xu et al., 2011) to create a robust passage retrieval system. While relation extraction systems exploit rich and complex features that are necessary to extract the exact relation (Mintz et al., 2009; Riedel et al., 2010; Hoffmann et al., 2011), passage retrieval components use coarse features in order to provide different and complementary feedback to information extraction models.

We exploit two types of lexical features: Bag-Of-Words and Word-Position. The two types of simple binary features are shown in the following example:

Sentence: *Apple founder Steve Jobs died.*
Target (Primary) entity: *Steve Jobs*
Bag-Of-Word features: *'apple' 'founder' 'died' '.'*
Word-Position features: *'apple:-2' 'founder:-1' 'died:+1' '.:+2'*

For each relation r, we assume each sentence has a binary relevance label to form distantly supervised training data: sentences in $POS(r)$ are *relevant* and sentences in $NEG(r)$ are *irrelevant*. As a pointwise learning-to-rank approach (Nallapati, 2004), the probabilities of relevance estimated by SVMs (Platt and others, 1999) are used for ranking all the sentences in the original training corpus for each relation respectively. We use LibSVM [2] (Chang and Lin, 2011) in our implementation.

2.3 Psuedo-relevance Relation Feedback

In the field of information retrieval, pseudo-relevance feedback assumes that the top-ranked documents from an initial retrieval are likely relevant, and extracts relevant terms to expand the original query (Xu and Croft, 1996; Lavrenko and Croft, 2001; Cao et al., 2008). Analogously, our assumption is that entity pairs that appear in more relevant and more sentences are more likely to express the relation, and can be used to expand knowledge base and reduce false negative noise in the training data for information extraction. We identify the most likely relevant entity pairs as follows:

initialize $\Delta' \longleftarrow \Delta$
for each relation type $r \in R$ **do**
  learn a passage (sentence) retrieval model $L(r)$ using coarse features and $POS(r) \cup NEG(r)$ as training data
  score the sentences in the $RAW(r)$ by $L(r)$
  score the entity pairs according to the scores of sentences they are involved in
  select the top ranked pairs of entities, then add the relation r to their label in Δ'
end for

We select the entity pairs whose average score of the sentences they are involved in is greater than p, where p is a parameter tuned on development data.[3] The relation extraction model is then trained using (Σ, E, R, Δ') with a more complete database than the original knowledge base Δ.

2.4 Distantly Supervised Relation Extraction

We use a state-of-the-art open-source system, MULTIR (Hoffmann et al., 2011), as the relation extraction component. MULTIR is based on multi-instance learning, which assumes that at least one sentence of those matching a given entity-pair contains the relation of interest (Riedel et al., 2010) in the given knowledge base to tolerate false positive noise in the training data and superior than previous models (Riedel et al., 2010; Mintz et al., 2009) by allowing overlapping relations. MULTIR uses features which are based on Mintz et al. (2009) and consist of conjunctions of named entity tags, syntactic dependency paths between arguments, and lexical information.

3 Experiments

For evaluating extraction accuracy, we follow the experimental setup of Hoffmann et al. (2011), and use their implementation of MULTIR[4] with 50 training iterations as our baseline. Our complete system, which we call IRMIE, combines our passage retrieval component with MULTIR. We use the same datasets as in Hoffmann et al. (2011) and Riedel et al. (2010), which include 3-years of New York Times articles aligned with Freebase. The **sentential extraction** evaluation is performed on a small amount of manually annotated sentences, sampled from the union of matched sentences and

[2] http://www.csie.ntu.edu.tw/~cjlin/libsvm

[3] We found $p = 0.5$ to work well in practice.

[4] http://homes.cs.washington.edu/~raphaelh/mr/

Test Data Set	Original Test Set				Corrected Test Set			
	$\tilde{P}$	$\tilde{R}$	$\tilde{F}$	$\Delta\tilde{F}$	$\tilde{P}$	$\tilde{R}$	$\tilde{F}$	$\Delta\tilde{F}$
MULTIR	80.0	44.6	62.3		92.7	47.7	70.2	
IRMIE	84.6	56.1	70.3	+8.0	92.6	61.2	76.9	+6.7
MULTIRLEX	91.8	43.0	67.4		79.6	57.0	68.3	
IRMIELEX	89.2	52.5	70.9	+3.5	78.0	69.2	73.6	+5.3

Table 1: Overall sentential extraction performance evaluated on the original test set of Hoffmann et al. (2011) and our corrected test set: Our proposed relevance feedback technique yields a substantial increase in recall.

system predictions. We define S^e as the sentences where some system extracted a relation and S^F as the sentences that match the arguments of a fact in Δ. The sentential precision and recall is computed on a randomly sampled set of sentences from $S^e \cup S^F$, in which each sentence is manually labeled whether it expresses any relation in R.

Figure 3 shows the precision/recall curves for MULTIR with and without pseudo-relevance feedback computed on the test dataset of 1000 sentence used by Hoffmann et al. (2011). With the pseudo-relevance feedback from passage retrieval, IRMIE achieves significantly higher recall at a consistently high level of precision. At the highest recall point, IRMIE reaches 78.5% precision and 59.2% recall, for an F1 score of 68.9%.

Because the two types of lexical features used in our passage retrieval models are not used in MULTIR, we created another baseline MULTIRLEX by adding these features into MULTIR in order to rule out the improvement from additional information. Note that the sentences are sampled from the union of Freebase matches and sentences from which some systems in Hoffmann et al. (2011) extracted a relation. It underestimates the improvements of the newly developed systems in this paper. We therefore also created a new test set of 1000 sentences by sampling from the union of Freebase matches and sentences where MULTIRLEX or IRMIELEX extracted a relation. Table 1 shows the overall precision and recall computed against these two test datasets, with and without adding lexical features into multi-instance learning models. The performance improvement by using pseudo-feedback is significant ($p < 0.05$) in McNemar's test for both datasets.

4 Conclusion and Perspectives

This paper proposes a novel approach to address an overlooked problem in distant supervision: the knowledge base is often incomplete causing numerous false negatives in the training data. It greatly improves a state-of-the-art multi-instance learning model by correcting the most likely false negatives in the training data based on the ranking of a passage retrieval model.

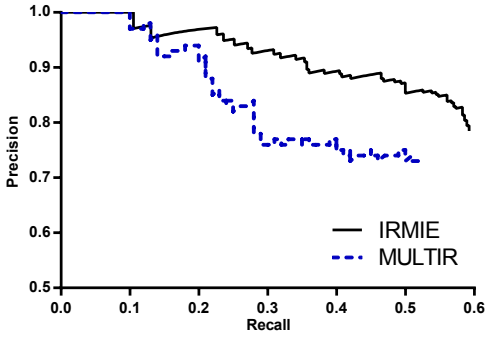

Figure 3: Sentential extraction: precision/recall curves using exact same training and test data, features and system settings as in Hoffmann et al. (2011).

In the future, we would like to more tightly integrate a coarser featured estimator of sentential relevance and a finer featured relation extractor, such that a single joint-model can be learned.

Acknowledgments

Supported in part by NSF grant IIS-1018317, the Air Force Research Laboratory (AFRL) under prime contract number FA8750-09-C-0181 and the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior National Business Center contract number D11PC20154. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of AFRL, IARPA, DoI/NBC, or the U.S. Government.

References

Avrim Blum and Tom M. Mitchell. 1998. Combining labeled and unlabeled sata with co-training. In *Proceedings of the 11th Annual Conference on Computational Learning Theory (COLT)*, pages 92–100.

Razvan C. Bunescu and Raymond J. Mooney. 2007. Learning to extract relations from the web using minimal supervision. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Guihong Cao, Jian-Yun Nie, Jianfeng Gao, and Stephen Robertson. 2008. Selecting good expansion terms for pseudo-relevance feedback. In *Proceedigns of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, pages 243–250.

Chih-Chung Chang and Chih-Jen Lin. 2011. Libsvm: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2(3):27.

Mark Craven and Johan Kumlien. 1999. Constructing biological knowledge bases by extracting information from text sources. In *Proceedings of the Seventh International Conference on Intelligent Systems for Molecular Biology (ISMB)*, pages 77–86.

Raphael Hoffmann, Congle Zhang, Xiao Ling, Luke S. Zettlemoyer, and Daniel S. Weld. 2011. Knowledge-based weak supervision for information extraction of overlapping relations. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 541–550.

Victor Lavrenko and W. Bruce Croft. 2001. Relevance-based language models. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, pages 120–127.

Tie-Yan Liu. 2011. *Learning to Rank for Information Retrieval*. Springer-Verlag Berlin Heidelberg.

Irina Matveeva, Chris Burges, Timo Burkard, Andy Laucius, and Leon Wong. 2006. High accuracy retrieval with multiple nested ranker. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, pages 437–444.

Bonan Min, Ralph Grishman, Li Wan, Chang Wang, and David Gondek. 2013. Distant supervision for relation extraction with an incomplete knowledge base. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL 2013)*.

Mike Mintz, Steven Bills, Rion Snow, and Daniel Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In *Proceedigns of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing (ACL)*, pages 1003–1011.

Ramesh Nallapati. 2004. Discriminative models for information retrieval. In *Proceedigns of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, pages 64–71.

Truc Vien T Nguyen and Alessandro Moschitti. 2011. End-to-end relation extraction using distant supervision from external semantic repositories. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 277–282.

John Platt et al. 1999. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in Large Margin Classifiers*, 10(3):61–74.

Karthik Raghunathan, Heeyoung Lee, Sudarshan Rangarajan, Nathanael Chambers, Mihai Surdeanu, Dan Jurafsky, and Christopher Manning. 2010. A multi-pass sieve for coreference resolution. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 492–501. Association for Computational Linguistics.

Sebastian Riedel, Limin Yao, and Andrew McCallum. 2010. Modeling relations and their mentions without labeled text. In *Proceedigns of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML/PKDD)*, pages 148–163.

Ang Sun, Ralph Grishman, Wei Xu, and Bonan Min. 2011. New york university 2011 system for kbp slot filling. In *Text Analysis Conference 2011 Workshop*.

Mihai Surdeanu, Julie Tibshirani, Ramesh Nallapati, and Christopher D. Manning. 2012. Multi-instance multi-label learning for relation extraction. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 455–465.

Shingo Takamatsu, Issei Sato, and Hiroshi Nakagawa. 2012. Reducing wrong labels in distant supervision for relation extraction. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 721–729.

Chang Wang, James Fan, Aditya Kalyanpur, and David Gondek. 2011. Relation extraction with relation topics. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1426–1436.

Jinxi Xu and W. Bruce Croft. 1996. Query expansion using local and global document analysis. In Hans-Peter Frei, Donna Harman, Peter Schäuble, and Ross Wilkinson, editors, *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, pages 4–11. ACM.

Wei Xu, Ralph Grishman, and Le Zhao. 2011. Passage retrieval for information extraction using distant supervision. In *Proceedings of the International Joint Conference on Natural Language Processing (IJCNLP)*, pages 1046–1054.

Roman Yangarber. 2001. *Scenario customization for information extraction*. Ph.D. thesis, Department of Computer Science, Graduate School of Arts and Science, New York University.