The Workshops of the Thirtieth AAAI Conference on Artificial Intelligence
Knowledge Extraction from Text: Technical Report WS-16-10

Chinese Relation Extraction by Multiple Instance Learning

Yu-Ju Chen and **Jane Yung-jen Hsu**
National Taiwan University
Department of Computer Science & Information Engineering
Taipei, Taiwan
{r01922049, yjhsu}@csie.ntu.edu.tw

Abstract

Relation extraction, which learns semantic relations of concept pairs from text, is an approach for mining commonsense knowledge. This paper investigates an approach for relation extraction, which helps expand a commonsense knowledge base with little labor work. We proposed a framework that learns new pairs from Chinese corpora by adopting concept pairs in Chinese commonsense knowledge base as seeds. Multiple instance learning is utilized as the learning algorithm for predicting relation for unseen pairs. The performance of our system could be improved by learning multiple iterations. The results in each iteration are manually evaluated and processed to next iteration as seeds. Our experiments extracted new pairs for relations "*AtLocation*", "*CapableOf*", and "*HasProperty*". This study showed that new pairs could be extracted from text without huge humans work.

Introduction

As a source of knowledge, text plays the role of corpus for relation extraction, which learns concept pairs linked with some relations and transforms sentences to knowledge graph. For one relation, a set of seed pairs is given for training a model, which could predict whether the relation exists in each new pair. To achieve good performance, state-of-the-art supervised learning requires a large labeled training set, which is often expensive to prepare. As an alternative, distant supervision, a semi-supervised learning method, was adopted to extract relations from unlabeled corpora. A training set consisting of a large amount of sentences can be weakly labeled automatically based on a set of concept pairs for any given relation in a knowledge base.

However, labels generated with heuristics can be quite noisy. When the sources of sentences in the training set are not correlated with the knowledge base, the automatic labeling mechanism is unreliable. Instead of assuming all sentences are labeled correctly in the training set, multiple instance learning learns from bags of instances, provided that each positive bag contains at least one positive instance while negative bags contain only negative instances.

By implementing the relation extraction problem as multiple instance learning, data is transformed from sentences to features and stored in a bag. The features used for training include syntactic and lexical features. After being transformed as features, sentences are packed as bags by the occurrence of concept pairs.

We conducted experiments on relation extraction in Chinese using concept pairs in ConceptNet[1], a commonsense knowledge base, as the seeds for labeling a set of predefined relations. The training bags were generated from the Sinica Corpus[2].

Related Work

Since DIPRE (Brin 1999) extracted book information, the relation of *author* and *title*, with pattern matching method, more work of relation extraction were created. Supervised learning were often used after the ACE competition [3] held the relation extraction track. When applying supervised learning to relation extraction problem, a set of training data is required and the problem is formulated as a classification task. When considering a single relation, the problem could be viewed as a binary classification task and aims at deciding whether the relation exists in the given concept pairs. Supervised learning for relation extraction includes feature-based method (Zhou et al. 2005) and kernel-based method (Zelenko, Aone, and Richardella 2003).

With the growing of information, supervised learning was no longer affordable to deal with such huge data. Therefore, one of the semi-supervised learning method – distant supervision was proposed to deal with large number of unlabeled data with the assistance of an exterior knowledge base. Distant supervision provided weakly labeling mechanism by taking knowledge base as labeling heuristic. At first, only hypernym relation was learned by borrowing knowledge from WordNet (Snow, Jurafsky, and Ng 2004). Afterwards, hundreds of relations were extracted from Wikipedia by applying Freebase as assistance (Mintz et al. 2009).

Given distant supervision, the labeling effort is heavily reduced; however, it causes noise when the sources of corpus and knowledge base are not correlated. For example,

[1] ConceptNet: http://conceptnet5.media.mit.edu

[2] Academia Sinica Balanced Corpus of Modern Chinese: http://rocling.iis.sinica.edu.tw/CKIP/engversion/20corpus.htm

[3] Automatic Content Extraction: http://www.itl.nist.gov/iad/mig/tests/ace/

when considering the 2 sentences "Alice was born in Taipei" and "Alice went to Taipei on Saturday", both contain the two entities "Alice" and "Taipei". The former sentence indicates the relation *BornIn* while the latter sentence expresses the relation *WentTo*. The example shows that multiple relations could be conveyed in different sentences that contain the same pair of entities. Thus, it is not reasonable to label the two sentences with the same relation. Riedel considered the cases that the distant supervision assumption is violated (Riedel, Yao, and McCallum 2010). Taking Freebase as the assistant knowledge base for labeling sentences in 2 corpora, Wikipedia and New York Time Corpus, Riedel found that 31% labels for New York Time Corpus violate the assumption while only 13% for Wikipedia. To avoid the unreason of the strong assumption, multiple instance learning is applied to this problem.

Multiple instance learning (MIL) learns a classifier based on a set of training **bags**, where data are collected with some policies (Amores 2013). MIL has been applied to several tasks such as drug discovery, text classification, image classification, and so forth. Relation extraction problem adopts MIL by packing bags by the index representing 2 entities (Bunescu and Mooney 2007a). Entities and mentions could be considered at the same time (Yao, Riedel, and McCallum 2010; Riedel, Yao, and McCallum 2010). The entity pairs and the sentences mentioning both entities are modeled in a conditional probability distribution. Then the unlabeled mentions would be given a probabilistic value deciding the possibility that the relation exists in the sentence. Furthermore, the relation extraction problem could be extended as multiple-instance-multiple-label problem, which models the mentions of pairs, with the labels of relations (Surdeanu et al. 2012). One model could deal with multiple labels. Hence, the method deals with multiple relations simultaneously.

Problem and Framework

Considering the scenario of relation extraction, given a set of entity pairs as seeds indicating a relation, we are going to extract new pairs representing such relations from a corpus.

Notations

First, we let C denote a corpus. Each $s \in C$ is a sentence, which is constructed by words. Given a corpus C, an entity set is defined as $E = \{e \mid e \text{ is a word in } C\}$. Then we let R denotes a relation set. Each $r \in R$ is a relation, corresponding to a seed set $S_r = \{(e_i, e_j) \mid e_i, e_j \in E\}$, $r \in R$. The tuple $(e_i, e_j) \in S_r$ indicates that 2 entities e_i and e_j are semantically connected with the relation r. In this problem, a new pair set is defined as $N_r = \{(e_i, e_j) \mid e_i, e_j \in E; (e_i, e_j) \notin S_r\}$, $r \in R$.

Problem Definition

Given a corpus C and a seed set S_r, the relation extraction system will create a new pair set N_r. The pairs in N_r are extracted from C and excluded from S_r.

- **Input**: a corpus C, a seed set $S_r = \{(e_i, e_j) \mid e_i, e_j \in E\}$, $r \in R$
- **Output**: a set of new pairs $N_r = \{(e_i, e_j) \mid r \in R; e_i, e_j \in E; (e_i, e_j) \notin S_r\}$, $r \in R$

Framework

The overall framework of the relation extraction system is shown on Figure 1, and the process is defined as Algorithm 1. The framework is separated into 3 parts: **bag generator**, **relation predictor** and **pair evaluator**.

Algorithm 1 Overall process of relation extraction

Input: a set of seeds $S_r^{(1)}$, a corpus C, a set of entities E, maximal iteration number M
Output: a set of new pairs N_r

```
generate an unlabeled pair set U = {(e_i, e_j) | e_i, e_j ∈ E} from C
for t = 1 to M do
    generate a labeled bag set B_label^(t) from C and S_r^(t) with Bag Generator
    generate an unlabeled bag set B_unlabel^(t) from C and U with Bag Generator
    train a model Relation Predictor with B_label^(t)
    with the Relation Predictor, predict labels for all data in B_unlabel^(t)
    select positive pairs from B_unlabel^(t) as N_r^(t)
    generate new seed set S_r^(t+1) from N_r^(t) by Pair Evaluator
end for
return N_r^(1) ∪ N_r^(2) ∪ ... ∪ N_r^(M)
```

Bag Generator

The bag generator aims at mapping pairs to bags. As the example in Figure 2, a bag of the pair (**Taipei**, **Taiwan**) consists of sentences from the corpus mentioning **Taipei** and **Taiwan**. The bag generator not only groups sentences in a bag, but also transforms sentences to feature vectors. Given any pair (e_i, e_j), e_i, $e_j \in E$ and a corpus C, a bag b is generated with sentence s mentioning e_i and e_j. Thus, $b = \{v \mid v = f(e_i, e_j, s)\}$, where f is the function transforming the sentence with entity pair to feature vector. The input and output of a bag generator is defined as following:

- **Input**: a corpus C, an entity pair (e_i, e_j), $e_i, e_j \in E$
- **Output**: a bag b associated with (e_i, e_j)

With the bag generator, a set of seeds will be mapped to a set of bags. The label of seeds will be brought to the corresponding bags.

Relation Predictor

The relation predictor is used for generating new pairs from the corpus as a standard machine learning process. With labeled bag set B_{label} and an algorithm $\mathcal{A}$, the predictor is created to predict the label of each bag $b \in B_{unlabel}$.

In this work, the algorithm $\mathcal{A}$ is a multiple instance learning algorithm due to the restriction of the problem. The input and output of the relation predictor is defined as following:

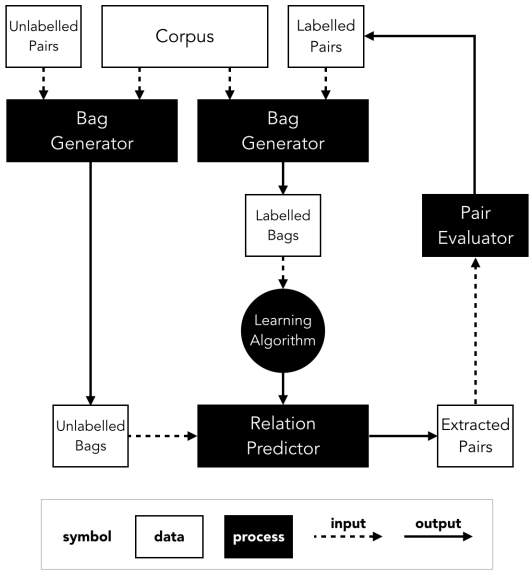

Figure 1: Framework of the relation extraction system, including the 3 components: bag generator, relation predictor, and pair evaluator

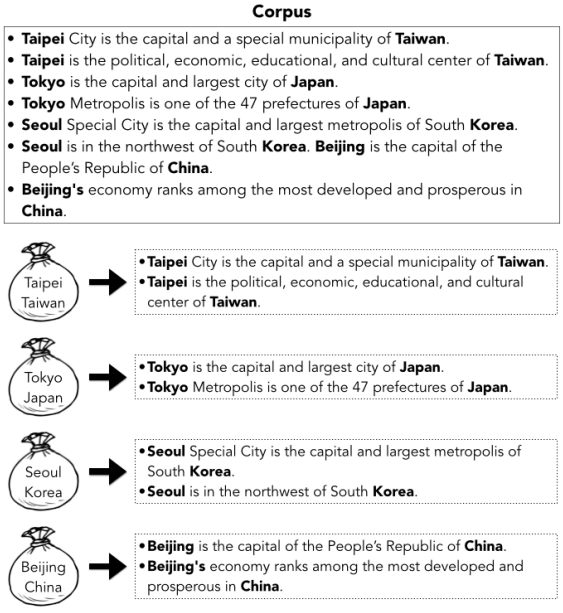

Figure 2: The corpus on the top contains sentences, which are selected from Wikipedia. The four bags represent four pairs. Each bag includes the sentences from corpus, which are illustrated on the right side.

- **Input**: a labeled bag set B_{label}, an unlabeled bag set $B_{unlabel}$, a learning algorithm $\mathcal{A}$
- **Output**: a set of new pairs N

Pair Evaluator

To iteratively learn new pairs from the corpus, we update the seed set for each iteration. To avoid using the false positive pairs as seeds in the next iteration, the result should be evaluated by another mechanism. Here we use human intelligence as the evaluator.

Given the new pair set $N_r^{(t)}$ generated in the t^{th} iteration, we ask human to evaluate the correctness and generate another set $S_r^{(t+1)}$, which is the seed set in the next iteration. The input and output of the pair evaluator is defined as following:

- **Input**: a set of candidate pairs $N_r^{(t)}$
- **Output**: a set of confident pairs $S_r^{(t+1)}$

Features

When generating the training data, plain texts were transformed to features. We followed the features of existing relation extraction work. The features are categorized as textual, part-of-speech (POS) tag, and syntactic features (Zhou et al. 2005; Mintz et al. 2009). Textual features consider the words in the sentence, including the entities, words between entities, words before and after the entities. POS tag features also take words into account, by using the POS tags of the words, which are marked in the corpus. For example, POS tag of entities, POS tag of words between, before and after entities are regarded as features. Syntactic features utilize parse tree and dependency tag in the sentence, which are obtained from Stanford Parser [4].

Automatic Labeling

Distant supervision is adopted for relation extraction in order to reduce labeling effort. Instead of describing a relation with a *sentence*, here we use a *bag* to represent a relation. Given any seed $(e_i, e_j) \in S_r$ and a bag of sentences mentioning e_i and e_j, at least one sentence in the bag might express r.

A seed set $S_r = S_r^+ \cup S_r^-$, where S_r^+ contains entity pairs with relation r and S_r^- contains entity pairs without relation r. Any entity pair (e_i, e_j) in the seed set S_r may correspond to a bag of sentences $b \subset C$, where each sentence $s \in b$ contains the 2 enties e_i and e_j. If $(e_i, e_j) \in S_r^+$, then the label of the bag $y = +1$. Otherwise, if $(e_i, e_j) \in S_r^-$, then $y = -1$.

Multiple Instance Learning

One of the naive algorithms of multiple instance learning (MIL) learns with instances in the bags. The instances used for training are labeled according to the bags they belong

[4] Stanford Parser: http://nlp.stanford.edu/software/lex-parser.shtml

to. Comparing with MIL, this approach is named as "Single Instance Learning (SIL)" (Ray and Craven 2005). Without violating the assumption of MIL, instances in negative bags are certainly labeled as negative. However, instances in positive bags are all regarded as positive, which causes the negative instances to be mislabeled as positive.

Sparse MIL (sMIL) (Bunescu and Mooney 2007b) modifies the constraint of SIL. sMIL assumes that few instances in positive bags are really positive, so it favors the situation that few positive instances exist in positive bags. sMIL also models for large bags. It looses the constraint when bag size is large because it is not easy to find a positive instance for a sparse positive bag. sMIL is equivalent to SIL when there is only one instance in the positive bag.

The transductive SVM modifies the standard SVM to a constrained version, where the decision boundary is assumed as far from the unlabeled data as possible. In the problem of MIL, instances in positive bags could be viewed as unlabeled instances since the assumption "at least one instance in the positive bag is positive" indicates that the labels in positive bags are unsure. Sparse Transductive MIL (stMIL) replaces the original SVM with transductive SVM. In our work, stMIL is adopted for predicting the relation for sentences, because the positive bags used for training are sparse positive.

Experiment

In the experiment, the seeds are pairs from Chinese ConceptNet, in which there are 15 pre-defined relations. The corpus for generating labeled data and extracting new pairs is Sinica Corpus, which is separated as about 600,000 sentences and each sentence is segmented as words. Each 2 words in the sentence may convey one or no relation.

Multiple instance learning learns from bags, which are collections of instances. In this work, an instance is a sentence, and a bag is generated by sentences containing a specific entity pair. The size of bag depends on the occurrence of pairs. Given the seed of a relation in ConceptNet, the number of instances about the seed is decided by the frequency of the seed in Sinica Corpus. Most seeds in ConceptNet occur rarely in Sinica Corpus and the bag size is decided as 10.

We extracted new pairs of three relations: *AtLocation*, *CapableOf*, and *HasProperty* with the tool misvm (Doran and Ray 2014). The relations are defined as following and the extracted pairs of the three relations are sampled in Table 1.

- *AtLocation*(A,B): B is the location of A.
- *CapableOf*(A,B): A is able to do B.
- *HasProperty*(A,B): A has B as its property.

To evaluate the effectiveness of iterative learning, Figure 3 shows the precision from the 1^{st} to 6^{th} iteration. For each relation, we evaluate the precision of the top 50 candidates. After being labeled, the candidates are fed as the seeds of next iteration. *HasProperty* has great improvement in the 2^{nd} and 3^{rd} iteration. *AtLocation* performs only 50% at first, but steadily grows afterwards. Although *CapableOf* is not outstanding, it reaches 72% as the best, by adding 14% from the first iteration. The results shows that iterative learning is helpful for relation extraction.

AtLocation	HasProperty	CapableOf
捷運,台北 (metro,Taipei)	範圍,廣 (range,wide)	人,表現 (people,represent)
政治,台灣 (politics,Taiwan)	體積,小 (volume,low)	業者,推出 (dealer,release)
產品,市場 (product,market)	壓力,大 (pressure,strong)	政府,舉辦 (government,hold)
教授,台大 (professor,NTU)	聲音,大 (sound,loud)	學生,使用 (student,use)
活動,學校 (activity,school)	頻率,高 (frequency,high)	政府,採取 (government,adopt)

Table 1: Example of selected pairs of relation *AtLocation*, *HasProperty*, and *CapableOf*

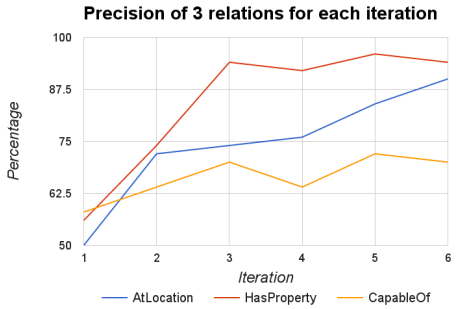

Figure 3: Comparison of the precision by iterations

Conclusion

This work develops a framework for extracting concept pairs from corpus based on the existing pairs in a knowledge base. Distant supervision with multiple instance learning is adopted to avoid costly human labeling work. MIL learns with bags and guarantees that positive bags contain at least one positive instance while negative bags contain all negative instances. In the experiment, concept pairs in Chinese ConceptNet are applied as seeds and sentences in Sinica Corpus serve as the source. Although not all relations could be efficiently extracted at the first iteration, the faults could be corrected manually and fed as the seeds in next iteration, which helps enhance the performance of relation extraction. To sum up, we proposed a iteratively learning framework, which requires little human efforts and generates nearly correct new pairs related to several relations from a corpus.

References

Amores, J. 2013. Multiple instance classification: Review, taxonomy and comparative study. *Artificial Intelligence* 201:81–105.

Brin, S. 1999. Extracting patterns and relations from the world wide web. In *The World Wide Web and Databases*. Springer. 172–183.

Bunescu, R., and Mooney, R. 2007a. Learning to extract relations from the web using minimal supervision. In *Annual meeting-association for Computational Linguistics*, volume 45, 576.

Bunescu, R. C., and Mooney, R. J. 2007b. Multiple instance learning for sparse positive bags. In *Proceedings of the 24th international conference on Machine learning*, 105–112. ACM.

Doran, G., and Ray, S. 2014. A theoretical and empirical analysis of support vector machine methods for multiple-instance classification. *Machine Learning* 97(1-2):79–102.

Mintz, M.; Bills, S.; Snow, R.; and Jurafsky, D. 2009. 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*, 1003–1011. Association for Computational Linguistics.

Ray, S., and Craven, M. 2005. Supervised versus multiple instance learning: An empirical comparison. In *Proceedings of the 22nd international conference on Machine learning*, 697–704. ACM.

Riedel, S.; Yao, L.; and McCallum, A. 2010. Modeling relations and their mentions without labeled text. In *Machine Learning and Knowledge Discovery in Databases*. Springer. 148–163.

Snow, R.; Jurafsky, D.; and Ng, A. Y. 2004. Learning syntactic patterns for automatic hypernym discovery. *Advances in Neural Information Processing Systems 17*.

Surdeanu, M.; Tibshirani, J.; Nallapati, R.; and Manning, C. D. 2012. Multi-instance multi-label learning for relation extraction. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, 455–465. Association for Computational Linguistics.

Yao, L.; Riedel, S.; and McCallum, A. 2010. Collective cross-document relation extraction without labelled data. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, 1013–1023. Association for Computational Linguistics.

Zelenko, D.; Aone, C.; and Richardella, A. 2003. Kernel methods for relation extraction. *The Journal of Machine Learning Research* 3:1083–1106.

Zhou, G.; Su, J.; Zhang, J.; and Zhang, M. 2005. Exploring various knowledge in relation extraction. In *Proceedings of the 43rd annual meeting on association for computational linguistics*, 427–434. Association for Computational Linguistics.