Open Information Extraction with Global Structure Constraints

Qi Zhu 1, Xiang Ren 2, Jingbo Shang 1, Yu Zhang 1, Frank F. Xu 3, Jiawei Han 1
1University of Illinois at Urbana-Champaign, IL, USA
2University of Southern California, CA, USA
3Shanghai Jiao Tong University, Shanghai, China

1{qiz3, shang7, yuz9, hanj}@illinois.edu 2xiangren@usc.edu 3frankxu@sjtu.edu.cn

ABSTRACT

Extracting entities and their relations from text is an important task for understanding massive text corpora. In this paper, we propose a novel open IE system, called ReMine, which integrates local context signal and global structural signal in a unified framework with distant supervision. The new system can be efficiently applied to different domains as it uses facts from external knowledge bases as supervision; and can effectively score sentence-level tuple extractions based on corpus-level statistics. Specifically, we design a joint optimization problem to unify (1) segmenting entity/rela-
tion phrases in individual sentences based on local context; and (2) measuring the quality of sentence-level extractions with a translating-
based objective. Experiments on real-world corpora from different domains demonstrate the effectiveness and robustness of ReMine when compared to other open IE systems.

1 INTRODUCTION

Massive corpora are emerging worldwide in different domains and languages. The sheer size of such data and the fast pace of new data generation make manual curation unscalable and infeasible. Information extraction (IE), i.e., entity and relation extraction, is a key step towards automated knowledge acquisition. Recently, researchers proposed several open-domain IE systems [1–4, 8] that do not require much human curation.

Open domain information extraction has first been introduced by Banko et al. [2]. Most of the existing work follows two lines of work, that is, pattern based methods or clause based methods. Pattern based information extraction can be as early as Hearst patterns like “NP0 such as {NP1, NP2, ...}” for hyponymy relation extraction [5]. ReVerb [4] identified relational phrase via part-of-speech-based regular expression. Besides part-of-speech tags, recent works start to use more linguistic features, like dependency parsing, to induct long distance relationships [6, 8]. Similarly, ClausIE [3] inducted short but coherent piece of information along dependency path, which is typically subject, predicate and optional object with complement. Angeli et al. adopt a clause splitter using distant training and mapped predicates to a known relation schema statistically [1]. Although previous effort achieves comparable high precision and reasonable coverage on extraction result, they all focus on local linguistic context.

In this paper, we study open IE problem from a unified perspective for massive text corpora, as shown in Fig. 1. First, ReMine will identify entity and relation phrases from local context. Then, ReMine jointly mines relation tuples and measure extraction with global translating objective. Local consistent text segmentation may generate noisy tuples, such as <your dry cleaner, set out from, eastern Queens> and <eastern Queens, on, foot>. However, from the global cohesiveness view, we may infer the second tuple as a false positive. Entity phrases like “eastern Queens” are seldom linked by relation phrase “on” in extracted tuples. Overall, ReMine will iteratively refine extracted tuples and learn entity and relation representation at corpus level.

2 THE REMINE FRAMEWORK

ReMine aims to jointly address two sub problems, that is, extracting entity & relation phrases and mining relation tuples. There are two challenges respectively, first, distant supervision may contain “false” entities and relation seeds, robust quality score need to be assigned on every phrase. Second, there exists multiple entity phrases in one sentence, selecting head and tail entity arguments for relation tuples may suffer from local structure ambiguity.

Framework Overview. We proposed a framework that integrates both local context and global statistics called ReMine (see also Fig. 1) as follows:

1) Do context-agnostic phrasal segmentation on target corpus, to generate entity phrases $E$ and relation phrases $R$. Apply random forest to obtain phrase type and quality from partially labeled training data $D_L$ with distant supervision.

2) Identify predicate $I$ between entity argument pair $(h,t) \in E^+_p$ and organize sentence-level relation tuples $T$ based on local segmentation objective Eq. 1, where $P(S,C_{(h,t)},T)$ generates phrase segments between two token span $C_{(h,t)}$.

3) Learn entity and relation representations $V$ via global translating objective Eq. 2. $F_{h,t,l,t'}$ is composed of training tuples with either $h$ or $t$ replaced.

4) Update sentence-level extractions with joint information from local context and global statistics, reduce local error by checking its global cohesiveness.

\[
O_{local} = \sum_{(e,e') \in E_p} \log P(S,C_{(e,e')},T) \quad (1)
\]

\[
O_{global} = \sum_{(h,l,t) \in E^+_{h,l,t}} \sum_{(h',l',t') \in F_{h',l',t'}} \gamma + \|h + l - t\| - \|h' + l - t'\| \quad (2)
\]

Overall updating schema To satisfy both local and global constraints, we have following objective for open IE

\[
O = O_{local} + O_{global} \quad (3)
\]

To maximize above unified open IE objective, we first initialize positive entity pairs $E^+_p$ given sentence segmentation result. Then we perform local segmentation between entities, which leads to
information tuples. Note that, at the first round, there are no global representation, so we preserve every relation phrase between entity arguments. Later we update global phrase semantic representation via global objective. With both global semantic information and local segmentation result, ReMine updates Positive Pairs and iteratively updating local and global objective until convergence, which will lead to a stable positive entity pairs. In experiments, we discover empirically, iteration can be stopped after second round.

3 EXPERIMENTS

We use two datasets in our experiments: (1) NYT [7] The training corpus consists of 236k sentences from 1987-2007 New York Times news articles. (2) Twitter. The dataset consists of 1.4 million tweets in Los Angeles with entities and or noun phrases collected from 2014.08.01 to 2014.11.30.

We compare ReMine with the following state-of-the-art information extraction methods including both pattern-based and clause-based methods: (1) OLLIE [8] (2) ClausIE [3] (3) Stanford OpenIE [1]

Open IE system can extract information tuples from open domain corpora. We compared ReMine with three Open IE systems mentioned above. We manually labeled the extractions got from ReMine and other three baseline extractors. Each extraction was labeled by two independent annotators for 2 rounds. Both annotators are highly proficient and literate in English.

Among all the Open IE system described above, ReMine and OLLIE extract a relatively small number of tuples. For example, for the first 100 sentences in the NYT test set, both ReMine and OLLIE get about 300 tuples. In contrast, Stanford OpenIE returns more than 1,000 tuples. We sample top 300 tuples for each Open IE system to plot the curves. The results are shown in Figure 2 and Table 1. According to the curve in Figure 2, ReMine achieves the best performance among all Open IE systems. In the NYT dataset, OLLIE, Stanford OpenIE and ClausIE actually have similar overall precision (i.e. P@300). But OLLIE has a “higher” curve since most tuples obtained by Stanford OpenIE and ClausIE will be assigned score 1. In contrast, the scores of different tuples obtained by OLLIE and ReMine are usually distinct from each other. In Table 1, ReMine also consistently performs the best according to the rank-based measures. In the Twitter dataset, ClausIE has a rather low score since there are lots of non-standard language usages and grammatical errors in tweets. Therefore clause-based methods may not achieve a satisfying performance. In contrast, ReMine shows its power in dealing with short and noisy text.

REFERENCES