

Open Information Extraction with Global Structure Constraints

Qi Zhu¹, Xiang Ren², Jingbo Shang¹, Yu Zhang¹, Frank F. Xu³, Jiawei Han¹

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. Open information extraction (IE) systems mine relation tuples (i.e., entity arguments and a predicate string to describe their relation) from sentences. However, current open IE systems ignore the fact that *global* statistics in a large corpus can be *collectively* leveraged to identify high-quality sentence-level extractions. 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/relation 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, 7] 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 “ NP_0 such as $\{NP_1, NP_2, \dots\}$ ” for hyponymy relation extraction [5]. ReVerb [4] identified relational phrases via part-of-speech-based regular expressions. Besides part-of-speech tags, recent works start to use more linguistic features, like dependency parsing, to induct long distance relationships [7]. Similarly, ClausIE [3] inducted short but coherent pieces of information along dependency paths, 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 efforts achieve high precision and reasonable coverage on extraction results, they all focus on local 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 representations 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 exist multiple entity phrases in one sentence, selecting head and tail entity arguments for relation tuples may suffer from local structural ambiguity.

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

- (1) Do context-dependent phrasal segmentation on target corpus, to generate entity phrases $\mathcal{E}$ and relation phrases $\mathcal{R}$.
- (2) Identify predicate between entity argument pair $(e_h, e_t) \in E_p^+$ and organize sentence-level relation tuples $\mathcal{T}$ based on local segmentation objective Eq. 1, where $\|w_i\| = \|v_h + v_r - v_t\|$ is the global structure constraints and $A_i(e_h, e_t)$ stands for the word sequence probability between the entity argument pair.
- (3) Learn entity and relation representations $\mathcal{V}$ via global translating objective Eq. 2, $E_{h',l,t'}^-$ is composed of training tuples with either h or t replaced.
- (4) Update sentence-level extractions with joint information from local context and global constraints.

$$O_{local} = \sum_{(e_h, e_t) \in E_p^+} \log P(r, e_h, e_t) = \sum_{(e_h, e_t) \in E_p^+} \sum_i \log \|w_i\| A_i(e_h, e_t) \quad (1)$$

$$O_{global} = \sum_{(h,l,t) \in E_{h,l,t}^+} \sum_{(h',l,t') \in E_{h',l,t'}^-} \|v_h + v_l - v_t\| - \|v_{h'} + v_l - v_{t'}\| - \gamma \quad (2)$$

This paper is published under the Creative Commons Attribution 4.0 International (CC BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.

WWW' 18 Companion, April 23–27, 2018, Lyon, France

© 2018 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC BY 4.0 License.

ACM ISBN 978-1-4503-5640-4/18/04.

<https://doi.org/10.1145/3184558.3186927>

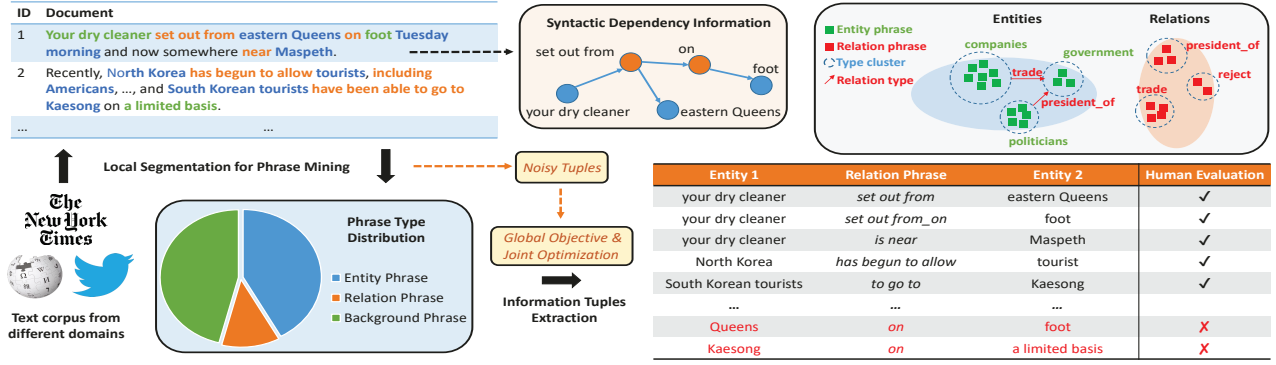


Figure 1: The Overview of the ReMine Framework.

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^{+0} 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 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 [6] The training corpus consists of 236k sentences from ~294k 1987-2007 New York Times news articles. (2) Twitter: The dataset consists of 1.4 million tweets in Los Angeles 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 [7] (2) ClausIE [3] (3) Stanford OpenIE [1]

Table 1: Performance of different methods on both datasets.

Dataset	Methods	P@100	P@200	MAP	NDCG@100	NDCG@200	MRR
NYT	ClausIE	0.520	0.475	0.534	0.567	0.632	0.027
	Stanford	0.660	0.585	0.630	0.655	0.726	0.023
	OLLIE	0.670	0.640	0.683	0.684	0.775	0.028
	ReMine	0.740	0.685	0.726	0.757	0.776	0.027
Twitter	ClausIE	0.280	0.290	0.292	0.309	0.522	0.024
	Stanford	0.520	0.495	0.493	0.468	0.620	0.016
	OLLIE	0.580	0.510	0.525	0.519	0.626	0.017
	ReMine	0.590	0.625	0.593	0.585	0.657	0.022

Open IE system can extract information tuples from open domain corpus. 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.

Among all the Open IE system described above, ReMine and OLLIE extract a relatively small number of tuples. For example,

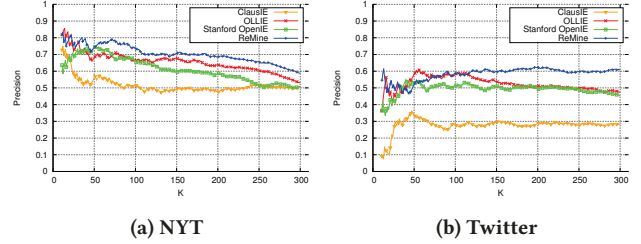


Figure 2: Precision@K on both datasets for the first 100 sentences in the NYT test set, both ReMine and OLLIE get about 300 tuples. In contrast, Stanford OpenIE returns more than 1,000 tuples. We sample top 300 tuples for each Open IE system to plot the curves. The results are shown in Figure 2 and Table 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 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 grammatic 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

- [1] Gabor Angeli, Melvin Johnson Premkumar, and Christopher D Manning. 2015. Leveraging linguistic structure for open domain information extraction. In *ACL '15*. ACL, 344–354.
- [2] Michele Banko, Michael J Cafarella, Stephen Soderland, Matthew Broadhead, and Oren Etzioni. 2007. Open Information Extraction from the Web.. In *IJCAI*, Vol. 7. 2670–2676.
- [3] Luciano Del Corro and Rainer Gemulla. 2013. Clausie: clause-based open information extraction. In *WWW'13*. ACM, 355–366.
- [4] Anthony Fader, Stephen Soderland, and Oren Etzioni. 2011. Identifying relations for open information extraction. In *EMNLP'11*. ACL, 1535–1545.
- [5] Marti A Hearst. 1992. Automatic acquisition of hyponyms from large text corpora. In *ACL '92*. ACL, 539–545.
- [6] Sebastian Riedel, Limin Yao, Andrew McCallum, and Benjamin M Marlin. 2013. Relation Extraction with Matrix Factorization and Universal Schemas. In *HLT-NAACL '13*. ACL, 74–84.
- [7] Michael Schmitz, Robert Bart, Stephen Soderland, Oren Etzioni, et al. 2012. Open language learning for information extraction. In *EMNLP-CoNLL '12*. ACL, 523–534.