Knowledge Base Population using Semantic Label Propagation

Lucas Sterckx LUCAS.STERCKX@INTEC.UGENT.BE
Thomas Demeester THOMAS.DEMEESTER@INTEC.UGENT.BE
Johannes Deleu JOHANNES.DELEU@INTEC.UGENT.BE
Chris Develder CHRIS.DEVELDER@INTEC.UGENT.BE
Ghent University - iMinds, Technologiepark Zwijnaarde 15, BE-9052 Gent, Belgium

Keywords: Knowledge Base Population, Relation Extraction, Natural Language Processing

Abstract

We present Semantic Label Propagation (SLP), a method for training of relation extractors for the purpose of automated knowledge base population (KBP). SLP uses minimal human supervision by applying distant supervision and annotating features (in particular shortest dependency paths) rather than complete relation instances. Such feature labeling initially filters too much training data from the training set. We further improve on this approach by introducing SLP, which uses the similarity between low-dimensional representations of candidate training instances to again expand the filtered training set. Evaluation on an established KBP test collection from the TAC English slot filling task shows that SLP substantially outperforms plain distant supervision in terms of recall and precision.

1. Introduction

In recent years we have seen significant advances in the creation of large-scale knowledge bases (KBs), databases containing millions of facts about persons, organizations, events, products, etc. Because KBs are often manually constructed, they tend to be incomplete. To automatically complete a KB we need a knowledge base population (KBP) system that extracts information from various sources of which a large fraction comprises unstructured written text items. A vital component of a KBP system is a relation extractor to populate a target field of the KB with facts extracted from natural language. Relation extraction (RE) is the task of assigning a semantic relationship between (pairs of) entities in text. Important for Supervised RE is the availability of training data. When an existing KB is available, a large set of known facts can be used to bootstrap training data, this procedure is known as distant supervision (DS): training data is obtained by heuristically aligning facts from a KB with an unlabeled corpus containing mentions of the KB's entities. DS suffers from noisy training data, since it may include incorrect examples for the intended relation: sentences mentioning a KB entry's entities do not necessarily express that exact KB relation. In this work we start from a distantly supervised training set and demonstrate how noise can be reduced, even with as few as just 5 minutes of annotations per relation, while obtaining significant improvements in precision and recall of the extracted relations.

Appearing in *Proceedings of Benelearn 2016*. Copyright 2016 by the author(s)/owner(s).

2. Labeling Strategy for Noise Reduction

First, we perform a highly selective form of noise reduction starting from a fully distantly supervised relation extractor by using the feature weights of this initial extractor to guide manual supervision in the feature space. Various questions arise from this. When do we over-constrain the original training set generated by DS? What is the trade-of between the application of DS with highly diverse labeled instances, and the constraining approach of labeling features, with a highly accurate yet restricted set of training data? Our approach comprises the following steps (see Figure 1):

1. An existing KB is used to generate distantly supervised training instances by matching its facts with sentences from a large text corpus.

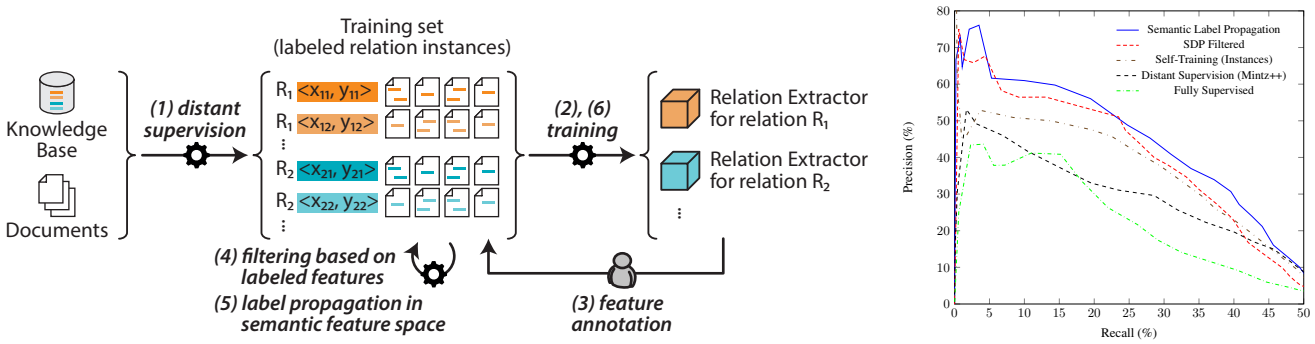


Figure 1. (Left) Overview of the Semantic Label Propagation workflow. (Right) Precision-Recall Graph displaying the output of the TAC KBP evaluation for different systems, for varying classifier decision thresholds.

2. An initial relation extractor is trained using the noisy training data generated in step (1).
3. Confident positive features (we use shortest dependency paths or patterns) learned by this initial classifier are presented to an annotator with knowledge of the semantics of the relation and labeled as true or false positive (see Table 1).
4. The collection of positive training instances is filtered to only patterns labeled as true positives in step (3) and a second classifier is trained.
5. In a *semi-supervised* step, the filtered distantly supervised training data is added to training data by propagating labels from labeled features to distantly supervised instances based on similarity in a semantic vector space of reduced dimension using *word embeddings*. Word embeddings allow for a relaxed semantic matching between the labeled seed patterns and the remaining weakly labeled patterns. Representing small phrases by summing each individual words embedding leads to semantic representations of small phrases that are meaningful for the goal of relation extraction. We represent each relation instance by a single vector by first removing stop-words and averaging the embeddings of the words on the dependency text path. This technique is coined *Semantic Label Propagation* (Sterckx et al., 2016).
6. A final supervised relation extractor is trained on this augmented training set.

3. Results and Conclusions

We evaluate the relation extractors in the context of a Knowledge Base Population system (Sterckx et al., 2015) using the NIST TAC KBP English Slot Filling (ESF) Evaluation from 2012 to 2014. In the end-to-end ESF framework, the input to the system is a given entity (the 'query'), a set of relations, and a collection of articles. The output is a set of slot fillers, where each slot filler is a triple consisting of two entities (including the query entity) and a relation predicted to hold among these entities. As shown in Figure 1, comparing the precision and recall for Semantic Label Propagation with the other approaches for partial supervision for RE shows that SLP outperforms the DS method, semi-supervision using instance labeling, and full supervision methods. In our TAC KBP 2015 submission (Sterckx et al., 2015), the ideas presented in this paper were applied, leading to a 2nd place out of 12 participating systems in the Slot Filling Variant. These results showed the influence of a clean training data and the effectiveness of SLP and self-training.

Table 1. Examples of top-ranked features

Relation	Top Short Dependency Text Pattern	Label
top_employees	PER $\xleftarrow{appos}$ *executive* $\xrightarrow{prep_of}$ ORG	✓
	ORG $\xleftarrow{nn}$ *founder* $\xrightarrow{prep_of}$ PER	✗
children	PER-2 $\xleftarrow{appos}$ *son* $\xrightarrow{prep_of}$ PER-1	✓
	PER-2 $\xleftarrow{nn}$ *grandson* $\xrightarrow{prep_of}$ PER-1	✗

References

Sterckx, L., Demeester, T., Deleu, J., & Develder, C. (2015). Ghent University-IBCN participation in TAC-KBP 2015 cold start task. *Proc. 8th Text Analysis Conf.* Gaithersburg, USA.

Sterckx, L., Demeester, T., Deleu, J., & Develder, C. (2016). Knowledge base population using semantic label propagation. *Knowledge-Based Systems.*