Knowledge-Based Systems 108 (2016) 79–91

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys

Knowledge base population using semantic label propagation

Lucas Sterckx*, Thomas Demeester, Johannes Deleu, Chris Develder

Ghent University – iMinds Technologiepark Zwijnaarde 15, BE-9052 Ghent, Belgium

ARTICLE INFO

Article history:
Received 15 November 2015
Revised 3 May 2016
Accepted 8 May 2016
Available online 10 May 2016

Keywords:
Relation extraction
Knowledge base population
Distant supervision
Active learning
Semi-supervised learning

ABSTRACT

Training relation extractors for the purpose of automated knowledge base population requires the availability of sufficient training data. The amount of manual labeling can be significantly reduced by applying distant supervision, which generates training data by aligning large text corpora with existing knowledge bases. This typically results in a highly noisy training set, where many training sentences do not express the intended relation. In this paper, we propose to combine distant supervision with minimal human supervision by annotating features (in particular shortest dependency paths) rather than complete relation instances. Such feature labeling eliminates noise from the initial training set, resulting in a significant increase of precision at the expense of recall. We further improve on this approach by introducing the Semantic Label Propagation (SLP) method, which uses the similarity between low-dimensional representations of candidate training instances to again extend the (filtered) training set in order to increase recall while maintaining high precision. Our strategy is evaluated on an established test collection designed for knowledge base population (KBP) from the TAC KBP English slot filling task. The experimental results show that SLP leads to substantial performance gains when compared to existing approaches while requiring an almost negligible human annotation effort.

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. Examples include Wikipedia-based KBs (e.g., YAGO [1], DBpedia [2], and Freebase [3]), KBs generated from Web documents (e.g., NELL [4], PROSPERA [5]), or open information extraction approaches (e.g., TextRunner [6], PRISMATIC [7]). Other knowledge bases like ConceptNet [8] or SenticNet [9] collect conceptual information conveyed by natural language and make them easily accessible for systems performing tasks like commonsense reasoning and sentiment analysis[10]. Besides the academic projects, several commercial projects were initiated by major corporations like Microsoft (Satori[1]), Google (Knowledge Graph [11]), Facebook[2], Walmart [12] and others. This is driven by a wide variety of applications for which KBs are increasingly found to be essential, e.g., digital assistants, or for enhancing search engine results with semantic search information.

Because KBs are often manually constructed, they tend to be incomplete. For example, 78.5% of *persons* in Freebase have no known *nationality* [13]. To 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 [11]. 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.

There are two categories of RE systems: (i) *closed*-schema IE systems extract relations from a fixed schema or for a closed set of relations while (ii) *open* domain IE systems extract relations defined by arbitrary phrases between arguments. We focus on the completion of KBs with a fixed schema, i.e., closed IE systems.

Effective approaches for closed schema RE apply some form of supervised or semi-supervised learning [14–19] and generally follow three steps: (i) sentences expressing relations are transformed to a data representation, e.g., vectors are constructed to be used in feature-based methods, (ii) a binary or multi-class classifier is trained from positive and negative instances, and (iii) the model is then applied to new or unseen instances. To review the evolution of these and other natural language processing techniques readers can refer to the article by Cambria and White [20].

Supervised systems are limited by the availability of expensive training data. To counter this problem, the technique of iterative bootstrapping has been proposed [21,22] in which an initial seed

* Corresponding author.
E-mail address: lucas.sterckx@intec.ugent.be (L. Sterckx).

[1] https://blogs.bing.com/search/2013/03/21/understand-your-world-with-bing
[2] http://www.insidefacebook.com/2013/01/14/

http://dx.doi.org/10.1016/j.knosys.2016.05.015
0950-7051/

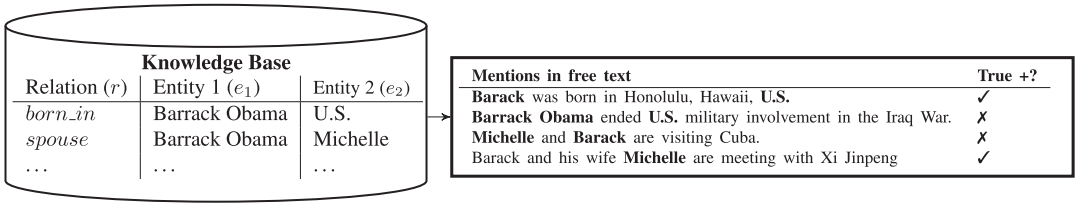

Fig. 1. Illustration of the distant supervision paradigm and errors.

set of known facts is used to learn patterns, which in turn are used to learn new facts and incrementally extend the training set. These bootstrapping approaches suffer from semantic drift and are highly dependent on the initial seed set.

When an existing KB is available, a much larger set of known facts can be used to bootstrap training data, a procedure known as distant supervision (DS). DS automatically labels its own training data by heuristically aligning facts from a KB with an unlabeled corpus. The KB, written as D, can be seen as a collection of relational tables $r(e_1, e_2)$, in which $r \in R$ (R is the set of relation labels), and $< e_1, e_2 >$ is a pair of entities that are known to have relation r. The corpus is written as C.

The intuition underlying DS is that any sentence in C which mentions the same pair of entities (e_1 and e_2) expresses a particular relationship $\hat{r}$ between them, which most likely corresponds to the known fact from the KB, $\hat{r}(e_1, e_2) = r(e_1, e_2)$, and thus forms a positive training example for an extractor of relation r. DS has been successfully applied in many relation extraction tasks [23,24] as it allows for the creation of large training sets with little or no human effort.

Equally apparent from the above intuition is the danger of finding incorrect examples for the intended relation. The heuristic of accepting each co-occurrence of the entity pair $< e_1, e_2 >$ as a positive training item because of the KB entry $r(e_1, e_2)$ is known to generate noisy training data or false positives [25], i.e., two entities co-occurring in text are not guaranteed to express the same relation as the field in the KB they were generated from. The same goes for the generation of negative examples: training data consisting of facts missing from the KB are not guaranteed to be false since a KB in practice is highly incomplete. An illustration of DS generating noisy training data is shown in Fig. 1.

Several strategies have been proposed to reduce this noise. The most prominent make use of latent variable models, in which the assumption is made that each known fact is expressed at least once in the corpus [25–27]. These methods are cumbersome to train and are sensitive to initialization parameters of the model.

An active research direction is the combination of DS with partial supervision. Several recent works differ in the way this supervision is chosen and included. Some focus on active learning, selecting training instances to be labeled according to an uncertainty criterion [23,28], while others focus on annotations of surface patterns and define rules or guidelines in a semi-supervised learning setting [29]. Existing methods for fusion of distant and partial supervision require thousands of annotations and hours of manual labor for minor improvements (4% in F_1 for 23,425 annotations [28] or 2,500 labeled sentences indicating true positives for a 3.9% gain in F_1 [29]). In this work we start from a distantly supervised training set and demonstrate how noise can be reduced, requiring only 5 min of annotations per relation, while obtaining significant improvements in precision and recall of the extracted relations.

We define the following research questions:

RQ 1. How can we add supervision most effectively to reduce noise and optimize relation extractors?

RQ 2. Can we combine semi-supervised learning and dimension reduction techniques to further enhance the quality of the training data and obtain state-of-the-art results using minimal manual supervision?

With the following contributions, we provide answers to these research questions:

1. In answer to RQ 1, we demonstrate the effectiveness and efficiency of filtering training data based on high-precision trigger patterns. These are obtained by training initial weak classifiers and manually labeling a small amount of features chosen according to an active learning criterion.
2. We tackle RQ 2 by proposing a semi-supervised learning technique that allows extending an initial set of high-quality training instances with weakly supervised candidate training items by measuring their similarity in a low-dimensional semantic vector space. This technique is called Semantic Label Propagation.
3. We evaluate our methodology on test data from the English Slot Filling (ESF) task of the knowledge base population track at the 2014 Text Analysis Conference (TAC). We compare different methods by using them in an existing KBP system. Our relation extractors attain state-of-the-art effectiveness (a micro averaged F_1 value of 36%) while depending on a very low manual annotation effort (i.e., 5 min per relation).

In Section 2 we give an overview of existing supervised and semi-supervised RE methods and highlight their remaining shortcomings. Section 3 describes our proposed methodology, with some details on the DS starting point (Section 3.1), the manual feature annotation approach (Section 3.2), and the introduction of the semantic label propagation method (Section 3.3). The experimental results are given in Section 4, followed by our conclusions in Section 5.

2. Related work

The key idea of our proposed approach is to combine DS with a minimal amount of supervision, i.e., requiring as few (feature) annotations as possible. Thus, our work is to be framed in the context of supervised and semi-supervised relation extraction (RE), and is related to approaches designed to minimize the annotation cost, e.g., active learning. Furthermore, we use compact vector representations carrying semantics, i.e., so-called word embeddings. Below, we therefore briefly summarize related work in the areas of (i) supervised RE, (ii) semi-supervised RE, (iii) evaluations of RE, (iv) active learning and (v) word embeddings.

2.1. Supervised relation extraction

Supervised RE methods rely on training data in the form of sentences tagged with a label indicating the presence or absence of the considered relation. There are three broad classes of supervised RE: (i) methods based on manual feature engineering, (ii) kernel based methods, and (iii) convolutional neural nets.

Methods based on feature-engineering [17,30] extract a rich list of manually designed structural, lexical, syntactic and semantic features to represent the given relation mentions as sparse vectors.