Distant Supervision from Knowledge Graphs

Alisa Smirnova, Julien Audiffren, Philippe Cudré-Mauroux

Synonyms

Self-supervised relation extraction.

Definitions

Definition 1. A *Knowledge Graph*, or Knowledge Base, is a semantic network defined as a set of triples (s, p, o) specifying that a node s (subject) is connected to another node o (object) by the property p. Sets of such triples form a directed graph, where nodes in the graph represent the subject and object in the triples and labelled edges represent the predicates connecting the subjects to the values. Generally, Knowledge Graphs can be seen as a collection of RDF triplets (see `https://www.w3.org/RDF/`).

Definition 2. *Distant Supervision* is a technique to automatically annotate input data using information contained in the knowledge graph. Given a knowledge graph $\mathscr{G}$ and a text corpus $\mathscr{C}$ (i. e., a collection of texts), the key idea of distant supervision is to align $\mathscr{G}$ to $\mathscr{C}$. More specifically, the idea is to first collect those sentences from the corpus $\mathscr{C}$ that mention the entity pair (e_1, e_2) where both e_1 and e_2 exist in the knowledge graph $\mathscr{G}$. If a sentence mentions (e_1, e_2) and there exists one triple (e_1, r, e_2) in the knowledge graph, then the distant supervision approach *labels* this sentence as an instance (also called *mention*) of relation r. The task of extracting such triples from raw text is called *relation extraction*.

Basic Approach

The idea of using a knowledge graph as a source of labels for training data was first proposed by Mintz et al (2009) and relies on the following assumption:

Assumption 1 *If two entities participate in a relation, any sentence that contains those two entities might express that relation.*

For example, the following sentence:

> South African entrepreneur **Elon Musk** is known for founding **Tesla Motors** and SpaceX.

mentions the tuple *(Elon Musk, Tesla Motors)*. Assuming that the triple *(Elon Mask, created, Tesla Motors)* exists in the knowledge graph, the textual sentence is labeled with the relation *created*, and can be used as training data for subsequent relation extractions. In this context, the set of sentences sharing the same entity pair is typically called a *bag* of sentences.

The relation extraction task can be considered as a classification problem, where the goal is to predict, for every entity pair, the relation it participates in from multiple classes. The classifier needs training data where every entity pair is represented as a *feature vector* (a binary representation of the input) and labeled with a corresponding relation. After learning on the training data, the testing step is performed, where the classifier predicts the relation labels for previously unseen entity pairs.

The transformation of the text corpus $\mathscr{C}$ into a usable training set involves several steps illustrated in Figure 1. The first step is a preprocessing step where classical Natural Language Processing tools are used to identify potential entities from the text.

The second step, Entity Matching, takes as input the textual entities identified in the previous step and tries to match them to one of the instances in the knowledge graph (e.g., it tries to match "William Tell" as found in the text to its corresponding instance in Wikidata, for example instance number "Q30908" (see `https://www.wikidata.org/wiki/Q30908`)).

In the third step, sentences where two entities are correctly matched to the knowledge graph are processed to extract features. Two types of features, lexical and syntactic, were originally proposed by Mintz et al (2009). Similar features were used in many subsequent approaches. Lexical features include the sequence of words between the two entities, flags indicating which entity name comes first in the sentence as well as the words immediately to the left of the first entity and to the right of the second entity. In addition to these features, syntactic features can be taken from a dependency parser – which extracts syntactic relations between words such as the dependency path between two entities.

Finally, the last step is called labelling, i. e., obtaining relation labels corresponding to the entity pair from the knowledge graph.

The test set is obtained as above when generating training data, except for the labelling part, i. e., the first three steps (natural language processing, entity matching and feature extraction) remain the same. After the training and test sets are constructed, a standard classifier can be applied for relation extraction. In this context, Mintz et al (2009) used a multi-class logistic classifier optimized using L-BFGS with Gaussian regularization.

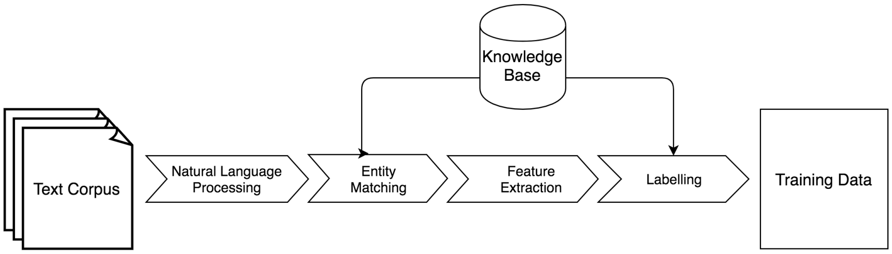

Fig. 1 The pipeline of preparing training data with distant supervision

Shortcomings of Distant Supervision

Automatically labelled training data are typically noisy, i. e., they contain false positives and false negatives, and this is also the case for distant supervision. On one hand, false negatives are mostly caused by the incompleteness of the knowledge graph. Two entities can be related in reality but their relation might be missing in the knowledge graph, hence their mentions will be wrongly labeled as negative examples by distant supervision. On the other hand, two entities may appear in the same sentence because they are related to the same topic, but not necessarily because the sentence is expressing the relation. Such examples might yield false positives using distant supervision.

Distant Supervision Improvements

This section lists several extensions aimed at improving the basic distant supervision method.

At-Least-One Principle

Riedel et al (2010) replace Assumption 1 with the following:

Assumption 2 *If two entities participate in a relation, at least one sentence that mentions those two entities will express that relation.*

This relaxed assumption, while more intuitive, comes at the cost of a more challenging classification problem. To tackle this problem the authors propose an undirected graphical model that solves both the task of predicting a relation between two entities and the task of predicting which sentence expresses this relation. The model is developed on top of a *factor graph*, a popular probabilistic graphical network representation where an undirected bipartite graph connects relation *variables* with *factors* representing their joint probabilities.

The original model cannot capture the case when two entities participate in more than one relation. In our above example, the relation between *Elon Musk* and *Tesla Motors* is not only *co-founder_Of* but also *CEO_Of*. Hoffmann et al (2011) and Surdeanu et al (2012) propose undirected graphical models, called MultiR and MIML − RE respectively, to perform *multi-instance multi-label* classification. Both models infer a relation expressed by a particular mention, thus the models are able to predict more than one relation between the two entities. The `MultiR` model is a conditional probability model that learns a joint distribution of both mention-level and bag-level assignments. The `MIML − RE` model contains two layers of classifiers. The first-level classifier assigns a label to a particular mention. The second level has k binary classifiers, where k is the number of known relation labels for the particular pair of entities. Each bag-level classifier y_i decides if the relation r_i holds for the given entity pair, using the mention-level classifications as input.

Negative Labels

In distant supervision, the knowledge graph only provides positive labels. Thus, negative training data is produced synthetically, e. g., by sampling the sentences containing two entities which are not related in the knowledge graph. The standard method to generate negative training data is to label mentions of unrelated entity pairs as negative samples. However, since the knowledge graph is incomplete, the absence of any relation label does not necessarily mean that the corresponding entities are not related. Hence, such a method potentially brings additional errors to the training data.

Several extensions of this model are proposed to overcome this problem. Min et al (2013) propose an additional layer to `MIML − RE` in order to model the bag-level label noise. Ritter et al (2013) extend the `MultiR` model to handle missing data. Xu et al (2013) explore ideas to enhance the knowledge graph, i. e., to infer entity pairs that are likely to participate in a relation even if they are unlabelled in the knowledge graph, and add them as positive samples.

Fan et al (2014) propose an alternative approach to tackle the problems of both erroneous labelling and incompleteness of the knowledge graph. They formulate the relation extraction task as a *matrix completion* problem.

Topic Models

Another widely used approach to improve text analysis and text classification is topic models. In this context, topics represent clusters of terms (words or patterns) which often co-occur in the documents together. The goal of topic models is to assign a topic to every term.

Topic models can be applied to distance supervision by mapping a *document* to a sentence mentioning an entity pair, and a *topic* to a relation. Words are represented by lexical and syntactic features, such as POS-tags or dependency paths between the entities. While the mention-level classifiers described above assign a relation to every sentence individually, the topic model classifiers are capable to capture more general dependencies between textual patterns and relations, which in practice can lead to improved performance.

Yao et al (2011) propose a series of generative probabilistic models in that context. Though the presented models are designed for unsupervised, open relation extraction (i. e., the set of the target relations is not pre-specified), the authors also show how the detected clusters can improve distantly supervised relation extraction. All their proposed models are based on *latent Dirichlet allocation* (LDA). The models differ in the set of features used and thus in their ability to cluster patterns. The advantage of generative topic models is that they are capable to 'transfer' information from known patterns to unseen patterns, i. e., to associate a relation expressed by an already observed pattern to a new pattern.

Alfonseca et al (2012) distinguish patterns that are expressing the relation from the ones that are not relation-specific and can be used across relations. For instance, the pattern "was born in" is relation-specific while the pattern "lived in" can be used across different relations, i. e., *mayorOf* or *placeOfDeath*. Their proposed models are also based on LDA. Different submodels have been suggested to capture three subsets of patterns:

- general patterns that appear for all relations;
- patterns that are specific for entity pairs and not generalisable across relations;
- patterns that are observed across most pairs with the same relation (i. e., relation-specific patterns).

Embedding-Based Methods

The methods discussed above are based on a large variety of lexical and syntactic features. We discuss below approaches that map the textual representation of relations and entity pairs onto a vector space. A mapping from discrete objects, such as words, to vectors of real number is called an embedding in this context. Embeddings-based approaches for relation extraction do not require extensive feature engineering and natural language processing, but they still require NER tags to link entities in the knowledge graph with their textual mentions.

Riedel et al (2013) apply matrix factorization for the relation extraction task. The authors compute a low-rank factorization of the probability matrix M, where rows correspond to entity pairs and columns correspond to relations. The elements of the matrix correspond to the probability that a relation holds between two entities. This factorization provides an embedding of entity pairs and of relations, which is then used to predict the probability of new triplets using a logistic function and one of the following feature model:

1. The latent feature model (F) defines $\theta_{r,t}$ as a measure of compatibility between relation r and tuple t via the dot product of their embeddings a_r and v_t respectively.
2. The neighbourhood model (N) defines $\theta_{r,t}$ as a set of weights $w_{r,r'}$, corresponding to a directed association strength between relation r and r', where both relations are observed for tuple t.
3. The entity model (E) also measures a compatibility between tuple t and relation r, but it provides latent feature vectors for every entity and every argument of relation r and thus can be used for n-ary relations. Entity types are implicitly embedded into the entity representation.
4. The combined model (NFE) defines $\theta_{r,t}$ as a sum of the parameters defined by the three above models.

Neural Networks

Another direction in text classification is applying Convolutional Neural Networks (CNNs). CNNs are able to find specific *n*-grams in the text and classify the relations expressed in the sentences based on these *n*-grams.

Zeng et al (2015) perform relation extraction via Piecewise Convolutional Neural Networks (PCNNs). The learning procedure of PCNNs is similar to standard CNNs. It consists of four parts: *Vector Representation, Convolution, Piecewise Max Pooling* and *Softmax Output*. In contrast to the embeddings-based approaches above, the proposed model does not build word embeddings itself, but uses pre-trained word vectors instead. Additionally, the model encodes a position of the word in the sentence with *position features* introduced by Zeng et al (2014). Position features represent the relative distance between the current word in the sentence and both entities of interest e_1 and e_2. The vector representation of a word is then the concatenation of word embedding and a position feature. In their work, the authors consider that the model predicts a relation for a bag (i. e., for an entity pair) if and only if a positive label is assigned to at least one entity mention.

Lin et al (2016) extend the PCNN approach. The authors explore different methods to overcome the wrong labelling problem. More specifically, the model learns weights of the sentences in a bag in order to select sentences that are true relation mentions. Two ways of defining the weights are explored:

- Average, where every sentence has the same weight;
- Selective Attention, where sentence weight depends on the query-based function which scores how well the sentence expresses a given relation.

Leveraging Auxiliary Information for Supervision

A number of distant supervision improvements use additional knowledge to enhance the model. Angeli et al (2014) and Pershina et al (2014) study the impact of extending training data with manually annotated data, which is generally-speaking more reliable than the labels obtained from the knowledge graph. Pershina et al (2014) propose to perform feature selection to generalize human labeled data into training guidelines and include them into the $\mathtt{MIML-RE}$ model. Angeli et al (2014) propose to initialize the $\mathtt{MIML-RE}$ model with manually annotated data, since the original model is sensitive to initialization. As manual annotation is very expensive, carefully selecting which sentences to annotate is of utmost importance. The authors study several criteria to pick the most valuable sentences for manual annotation.

Another explored direction is improving entity identification, i. e., extract sentences mentioning entities not only by their canonical names, but also by abbreviated mentions, acronyms, paraphrases and even pronouns. Given the following sentences:

> Elon Musk is trying to redefine transportation on earth and in space.
>
> Through Tesla Motors – of which he is cofounder, CEO and Chairman – he is

> aiming to bring fully-electric vehicles to the mass market.

The second sentence contains a relation mention (Elon Musk, cofounderOf, Tesla Motors) that cannot be extracted without *co-reference resolution*, which refers to the task of clustering mentions of the same entity together, typically within a single sentence or document.

Augenstein et al (2016) explore a variety of strategies for making entity identification tools more robust across domains. Moreover, they perform co-reference resolution for extracting relations across sentence boundaries. The approach is designed to perform relation extraction on very heterogeneous text corpuses, such as those extracted from the Web.

Koch et al (2014) use entity types and co-reference resolution for a more accurate relation extraction. Two sets of entity types are available: coarse types (PERSON, LOCATION, ORGANIZATION and MISC) come from the named entity recognizer (NER) while fine-grained types come from the knowledge graph. To bring type-awareness into the system, the authors build separate relation extractors on top of the `MultiR` model for each pair of coarse types, e. g., (PERSON, PERSON), (PERSON, LOCATION), and combine the extractions from the extractors of every type signature. Candidate mentions with incompatible entity types are then discarded, which improves precision at the cost of a slight drop in recall.

Chang et al (2014) propose a tensor decomposition approach that also uses entity type information and can be stacked with other models. In particular, it helps to overcome the fact that the matrix factorization model does not share information between rows, i. e., between entity pairs (including pairs containing the same entity).

Type-LDA is a topic model proposed by Yao et al (2011) that considers fined-grained entity types based on latent Dirichlet allocation. In contrast to the previous approaches, it learns fine-grained entity types directly from the text.

Finally, a few relation extraction approaches integrate logical connections or constraints between the relations. As an example, the relation *capitalOf* between two entities directly implies another relation *cityOf*. Furthermore, many relation pairs are mutually exclusive, e. g., *spouseOf* and *parentOf*.

Rocktäschel et al (2015) propose to inject logical formulae into the relations and entity pairs embeddings. Their model is based on the matrix factorization approach from Riedel et al (2013). They explore two methods, namely pre-factorization inference and joint optimization, and choose to focus on direct relation implication.

The approach proposed by Han and Sun (2016) is also based on the idea of using indirect supervision. The authors rely on a Markov Logic Network as a representation language and use:

- the consistency of relation labels, i. e., the inter-dependencies between relations that were mentioned previously (implications and mutual exclusion);
- the consistency between relation and arguments, i. e., entity type information retrieved from the knowledge graph is used to filter inconsistent candidates;
- the consistency between neighbouring instances: the authors define the

similarity function between two relation candidates as the cosine similarity of their feature vectors.

Acknowledgement

This work was supported by the Swiss National Science Foundation under grant number 407540_167320 (Tighten-it-All).

References

Alfonseca E, Filippova K, Delort JY, Garrido G (2012) Pattern learning for relation extraction with a hierarchical topic model. In: Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2, Association for Computational Linguistics, pp 54–59

Angeli G, Tibshirani J, Wu J, Manning CD (2014) Combining distant and partial supervision for relation extraction. In: EMNLP, pp 1556–1567

Augenstein I, Maynard D, Ciravegna F (2016) Distantly supervised web relation extraction for knowledge base population. Semantic Web 7(4):335–349

Chang KW, Yih SWt, Yang B, Meek C (2014) Typed tensor decomposition of knowledge bases for relation extraction

Fan M, Zhao D, Zhou Q, Liu Z, Zheng TF, Chang EY (2014) Distant supervision for relation extraction with matrix completion. In: ACL (1), Citeseer, pp 839–849

Han X, Sun L (2016) Global distant supervision for relation extraction. In: Thirtieth AAAI Conference on Artificial Intelligence

Hoffmann R, Zhang C, Ling X, Zettlemoyer L, Weld DS (2011) Knowledge-based weak supervision for information extraction of overlapping relations. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1, Association for Computational Linguistics, pp 541–550

Koch M, Gilmer J, Soderland S, Weld DS (2014) Type-aware distantly supervised relation extraction with linked arguments. In: Proceedings of EMNLP, Citeseer

Lin Y, Shen S, Liu Z, Luan H, Sun M (2016) Neural relation extraction with selective attention over instances. In: ACL (1)

Min B, Grishman R, Wan L, Wang C, Gondek D (2013) Distant supervision for relation extraction with an incomplete knowledge base. In: HLT-NAACL, pp 777–782

Mintz M, Bills S, Snow R, 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, Association for Computational Linguistics, pp 1003–1011

Pershina M, Min B, Xu W, Grishman R (2014) Infusion of labeled data into distant supervision for relation extraction. In: ACL (2), pp 732–738

Riedel S, Yao L, McCallum A (2010) Modeling relations and their mentions without labeled text. Machine learning and knowledge discovery in databases pp 148–163

Riedel S, Yao L, McCallum A, Marlin BM (2013) Relation extraction with matrix factorization and universal schemas

Ritter A, Zettlemoyer L, Etzioni O, et al (2013) Modeling missing data in distant supervision for information extraction. Transactions of the Association for Computational Linguistics 1:367–378

Rocktäschel T, Singh S, Riedel S (2015) Injecting logical background knowledge into embeddings for relation extraction. In: HLT-NAACL, pp 1119–1129

Surdeanu M, Tibshirani J, Nallapati R, Manning CD (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, Association for Computational Linguistics, pp 455–465

Xu W, Hoffmann R, Zhao L, Grishman R (2013) Filling knowledge base gaps for distant supervision of relation extraction. In: ACL (2), pp 665–670

Yao L, Haghighi A, Riedel S, McCallum A (2011) Structured relation discovery using generative models. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, pp 1456–1466

Zeng D, Liu K, Lai S, Zhou G, Zhao J, et al (2014) Relation classification via convolutional deep neural network. In: COLING, pp 2335–2344

Zeng D, Liu K, Chen Y, Zhao J (2015) Distant supervision for relation extraction via piecewise convolutional neural networks. In: EMNLP, pp 1753–1762