Feature-Based Models for Improving the Quality of Noisy Training Data for Relation Extraction

Benjamin Roth
Saarland University
Spoken Language Systems
Saarbrücken, Germany
benjamin.roth@lsv.uni-saarland.de

Dietrich Klakow
Saarland University
Spoken Language Systems
Saarbrücken, Germany
dietrich.klakow@lsv.uni-saarland.de

ABSTRACT

Supervised relation extraction from text relies on annotated data. Distant supervision is a scheme to obtain noisy training data by using a knowledge base of relational tuples as the ground truth and finding entity pair matches in a text corpus. We propose and evaluate two feature-based models for increasing the quality of distant supervision extraction patterns. The first model is an extension of a hierarchical topic model that induces background, relation-specific and argument-pair specific feature distributions. The second model is a perceptron, trained to match an objective function that enforces two constraints: 1) an *at-least-one* semantics, i.e. at least one training example per relational tuple is assumed to be correct; 2) high scores for a dedicated NIL label that accounts for the noise in the training data. For both algorithms, neither explicit negative data nor the ratio of negatives has to be provided. Both algorithms give improvements over a maximum likelihood baseline as well as over a previous topic model without features, evaluated on TAC KBP data.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing; I.2.7 [**Artificial Intelligence**]: Natural language processing—*text analysis*

General Terms

Algorithms, Experimentation

Keywords

Information Extraction, Machine Learning, Pattern Learning, Topic Models, Distant Supervision

1. INTRODUCTION

Relation extraction can be formulated as the task of turning unstructured text into tabularized information. Two relation extraction paradigms can be distinguished: 1) open information extraction, the unsupervised clustering of entity-context tuples [2], and 2) relation extraction for a fixed relation inventory, which is also known as knowledge-base population (KBP) [5]. While open information extraction does not require annotated data, it may not always provide the most useful granularity or partitioning for a specific task. In contrast, relation extraction for a pre-specified relation inventory may be better tailored for a specific task, but needs labeled training data; however, textual annotation is costly.

Databases with fact tuples such as *(PERSON, born-in, CITY)* are often readily available. However, there is usually no or only very little text annotated according to whether it expresses a relation (e.g. *born-in*) between particular entities (e.g. of types *PERSON* and *CITY*). This is used by the paradigm of *distant supervision (DS)*, [6]: Textual matches of entities from fact tuples are used to automatically generate relation contexts as training instances. For example, the arguments of the fact tuple *("Barack Obama", born-in, "Honululu")* could match in contexts like *"Barack Obama was born in Honululu"* and *"Barack Obama visited Honululu"*. Often only a small fraction of such matches indeed express the relation of the fact tuple.

The task of this work is to increase the quality of semi-automatically extracted training data for a given relation inventory. We are considering relation-specific surface patterns (e.g. *"[ARG1] was born in [ARG2]"*) that are obtained from noisy DS training data and applied to free text (together with a type checker). The aim is to find a scoring algorithm which computes a ranking of noisy patterns that correlates best with the precision of facts extracted by them. We propose and evaluate two feature-based models for estimating the probability of noisy DS patterns expressing a given relation. The first is an extension of a recently proposed hierarchical topic model [1], which is based on the principle that patterns are generated either by a background distribution, a relation-specific distribution or an argument-pair-specific distribution, depending on which value a hidden topic variable takes on. The second model is a multi-class perceptron learner based on the principle that a DS pattern for a particular relation has a high probability of either expressing no relation between two entities or the relation of the corresponding database fact. Additionally, the perceptron learner assumes that for each relational tuple with a textual match, the training data contains at least one context that in fact expresses the relation. Experiments show that both approaches significantly improve the quality of extracted relational patterns.

CIKM'13, Oct. 27–Nov. 1, 2013, San Francisco, CA, USA.

http://dx.doi.org/10.1145/2505515.2507850.

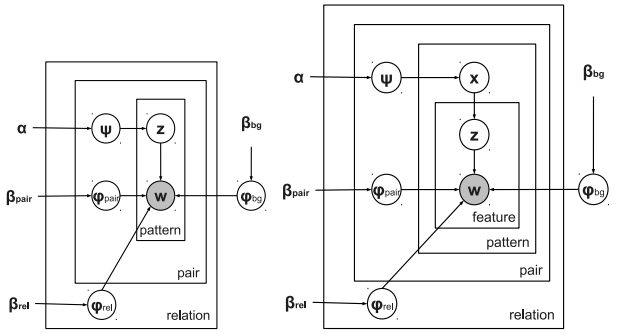

Figure 1: Hierarchical topic models. Intertext model (left) and feature model (right).

2. HIERARCHICAL TOPIC MODELS FOR DISTANT SUPERVISION

2.1 Intertext Topic Model

We build on the hierarchical topic model used by [1] that aims at extracting relation specific patterns. We refer to this model as the original *intertext* model, as relational patterns are defined only by the text between the arguments and are not broken up into features. This model is the hierarchical topic model for multi-document summarization of [3] with the following correspondences: Pairs of arguments are assumed to form documents, with the surface patterns as their words. Pairs of arguments are grouped together according to the relation they stand in.

The generative process assumes that for each argument pair of a particular relation, all the patterns (intertext contexts of DS matches) are generated by first generating a hidden variable z at a position i, depending on a pair-specific distribution ψ (with Dirichlet hyper parameters α). The variable z can take on three values, B for background, R for relation and P for pair. Corresponding vocabulary distributions $(\phi_{bg}, \phi_{rel}, \phi_{pair})$ are chosen for generating the context pattern at position i. The vocabulary distributions are smoothed by Dirichlet hyper parameters $\beta_{bg}, \beta_{rel}, \beta_{pair}$ and shared on the respective levels. See Figure 1 for a plate diagram of the basic model. Gibbs sampling is used to infer the topics of the document collection.

2.2 Feature Topic Model

The intertext topic model can only treat patterns as a whole. While this may be sufficient for very frequent patterns, for the long tail of infrequent patterns evidence on pattern-level may be to sparse to do meaningful inference. We therefore extend the model to include common structural elements on sub-patterns level (i.e. bigrams) that may be indicative and are shared among patterns.

In order to include features in the model, we propose a model with two layers of hidden variables. A variable x represents a choice of B, R or P for every pattern. Each feature is generated conditioned on a second variable $z \in \{B, R, P\}$ (the same way as in the simple intertext model). The features now range over index i, and patterns over index j. For a pattern at index j, first one hidden variable x is generated, then all z variables are generated for the corresponding features at indices i (see Figure 1).

Algorithm 1 At-Least-One Perceptron Learner with NIL

$\theta \leftarrow 0$
for $r \in \mathcal{R}$ **do**
  for $pair \in$ kb_pairs(r) **do**
    for $s \in$ sentences$(pair)$ **do**
      for $r' \in \mathcal{R} \setminus r$ **do**
        if $P(r|s,\theta) \leq P(r'|s,\theta)$ **then**
          $\theta \leftarrow \theta + \phi_i(s,r) - \phi_i(s,r')$
        if $P(NIL|s,\theta) \leq P(r'|s,\theta)$ **then**
          $\theta \leftarrow \theta + \phi_i(s,NIL) - \phi_i(s,r')$
    if $\forall_{s \in sentences(pair)} : P(r|s,\theta) \leq P(NIL|s,\theta)$ **then**
      $s^* = \arg\max_s \frac{P(r|s,\theta)}{P(NIL|s,\theta)}$
      $\theta \leftarrow \theta + \phi_i(s^*,r) - \phi_i(s^*,NIL)$

In the following formula a function $j(i)$ is used to denote the mapping from a feature index i to the index j of the corresponding pattern. The values B, R or P of z depend on the corresponding x by a transition distribution:

$$P(Z_i = z | X_{j(i)} = x) = \begin{cases} p_{same}, & \text{if } z = x \\ \frac{1-p_{same}}{2}, & \text{otherwise} \end{cases}$$

where p_{same} is set to .99 to enforce the correspondence between pattern and feature topics. While the original work reports hyper parameters $\alpha = (15, 1, 15)$, we found a uniform prior $\alpha = (1, 1, 1)$ to work slightly better, which we use for the feature-based experiments.

3. AT-LEAST-ONE PERCEPTRON LEARNER WITH NIL

As a second feature-based model, we propose a perceptron model with an objective function that enforces certain constraints. The model includes log-linear factors for all relations (from set of relations $\mathcal{R}$) as well as a factor for the *NIL* label, expressing noise or no relation between the entities. The estimated probabilities for a relation r given a sentence pattern s are calculated by normalizing over log-linear factors:

$$P(r|s,\theta) = \frac{f_r(s)}{\sum_{r' \in \mathcal{R} \cup NIL} f_{r'}(s)}$$

The factors are defined as:

$$f_r(s) = exp\left(\sum_i \phi_i(s,r)\theta_i\right)$$

with $\phi(s,r)$ the feature vector for sentence s and label assignment r, and θ_r the feature weight vector.The final scoring used to rank the patterns is obtained by:

$$\frac{P(r|s,\theta)}{P(r|s,\theta) + P(NIL|s,\theta)}$$

As we do not use annotated data for the NIL assignment and only have noisy DS data for the relation assignment, the learner is directed by the following semantics: First, for a DS sentence s that has a textual match for relation r, relation r should have a higher probability than any other relation $r' \in \mathcal{R} \setminus r$. Second, as extractions are noisy, we also expect a high probability for NIL. We therefore introduce the constraint that NIL has a higher probability than any

relation $r' \in \mathcal{R} \setminus r$ for which s is not a DS sentence. Third, at least one DS sentence for an argument pair is expected to express the corresponding relation r. For sentences s_i for an entity pair belonging to relation r, this can be written in the form of the following constraints:

$$\forall_{i,r'} : P(r|s_i) > P(r'|s_i) \wedge P(NIL|s_i) > P(r'|s_i)$$

$$\exists_i : P(r|s_i) > P(NIL|s_i)$$

Hence, the model is a multi-class learner to predict either the label r or the label NIL with the final decision constrained by the at-least-one semantics. An important point to note is that the NIL class is learned from candidates of all relations and not just a negative per-relation classification decision. It is therefore roughly corresponding to the background vocabulary in the topic model. The violation of any of the above constraints triggers a perceptron update. Similar to [4], the update corresponding to a violated *at-least-one* constraint is applied only to the one sentence that already has the highest score for the correct label. The basic algorithm is sketched in Algorithm 1. The actual implementation additionally uses averaging over all updates, weighting of updates by the model score of wrong labels and iterating several times over the data (we use 20 iterations in our experiments).

4. EVALUATION

We measure the ranking quality of the patterns by the ranking quality of their extractions. We use all TAC KBP queries from the years 2009-2011 and the TAC KBP 2009-2011 corpora[1]. The queries consist of 298 query entities with types *PERSON* or *ORGANIZATION*; there are 42 relations to be considered. First, candidate sentences are retrieved from the corpora in which the query entity and a second entity with the correct type for a sought relation is contained. Those candidate sentences are then used to provide answer candidates if one of the patterns – extracted from the training data – matches. An answer candidate is assigned the score of the matching pattern. If several patterns match, the score of the highest scored pattern is assigned. This ranking is then evaluated using the TAC KBP gold annotations[2]. The basis for pattern extraction is the (noisy) DS training data of a top-3 ranked system in TAC KBP 2012 [7]. We also use the retrieval components of this system to obtain sentence and answer candidates. The basis of evaluation consists of 38,939 response candidates from pattern matches, ranked according to their respective pattern scores. 951 of the response candidates are correct according to the gold annotation, 38 (out of 42) relations have at least one correct response candidate. Evaluation results are reported as averages over per-relation results.

The hierarchical topic model has originally been evaluated against maximum likelihood estimation by comparison of precision/probability curves [1]. However, note that in theory the precision values at probability thresholds can be increased (at the expense of recall) also by methods that generally lower relation probabilities without improving the overall ranking quality. While we include a precision/probability evaluation (Figure 2, left diagram), we focus on comparison of ranking measures (Table 1) and P/R curves (Figure 2, right diagram) to consider recall.

[1]For more information on TAC KBP see: `http://www.nist.gov/tac/tracks/index.html`

[2]Note that those annotations are a result of pooling and therefore incomplete and under-estimating precision. However, they allow for a relative comparison of ranking quality.

method	map	gmap	p@5	p@10
MLE	.253	.142	.263	.232
hier orig	.270	.158	.353*	.297*
hier feat	.312$^{\dagger **}$	.199$^{\dagger **}$	.347*	.303*
hier orig +burn	.286	.181	**.379****	.300*
hier feat +burn	.318$^{\dagger\dagger **}$	.205$^{\dagger\dagger **}$	.363**	.321**
perceptron	**.330**$^{\dagger\dagger **}$	**.210**$^{\dagger\dagger **}$	**.379****	**.337****

Table 1: Ranking quality on TAC KBP extractions. Significance (paired t-test) is marked w.r.t. *MLE* (*p<0.05, **p<0.01) and w.r.t. *hier orig* (†p<0.05, ††p<0.01).

The first baseline we use is the maximum-likelihood estimator (*MLE*), which scores patterns by the relative frequency they occur with a certain relation. For the following methods, the relative frequency is weighted by the score from the respective models that a certain pattern actually expresses a particular relation. A more advanced base-line is the hierarchical topic model (*hier orig*) as described in [1]. Additional substantial improvements were obtained by averaging counts over the last 10 training iterations, which is known as burn-in (*hier orig +burn*). Performance was improved under most measures by including features (bigrams) into the hierarchical model (*hier feat +burn*) as described in Section 2.2. The overall best results were obtained by the perceptron learner (*perceptron*) as described in Section 3, using bi-gram features.

The two proposed models clearly show the most significant improvement under the metrics. Mean average precision (*map*), geometric map (*gmap*) and interpolated precision at recall levels provide metrics over the entire ranking quality of all of the evaluation set. Precision at 5 and at 10 (*p@5*, *p@10*) are included for reference – however, these metrics disregard most of the evaluation set and give a coarser picture.

5. RELATED WORK

While there is a large body of work on relation extraction, knowledge-base population and distant supervision, an overview over the state of the art can be found in [5].

Feature-based topic models have been used in the context of relation extraction for unsupervised clustering of contexts by [10]. While those clusters have then been used as features in a DS context, the problem of noisy DS candidates has not been tackled by this work. [1] introduced a topic model for improving extraction of candidates, which part of this paper is an extension of so that it is possible to include features.

Employing an at-least-one semantics for relation extraction has been introduced by [11], which use SampleRank [9] and Gibbs-sampling, guided by an at-least-one objective function. [8] use an at-least-one feature to predict aggregate level relation labels. [4] use an at-least-one learner similar to ours to predict the relation of sentence aggregates. In the above cases explicit negative training data is used to increase the probability of the NIL label. Instead of using negative training data, in our work we introduced ranking

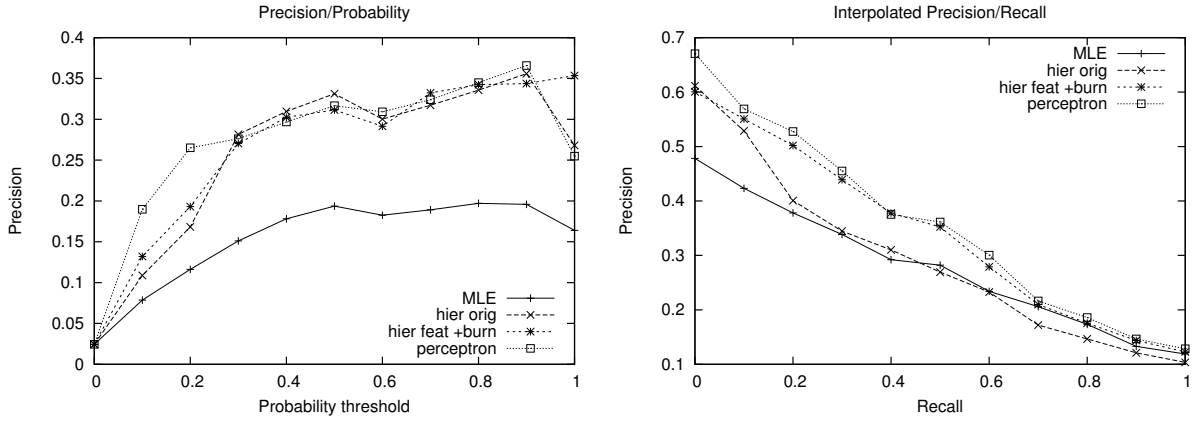

Figure 2: Precision at probability thresholds (left) and precision at recall levels (right).

constraints that – together with an *at-least-one* semantics – boost probability estimates for both NIL and the relation labels in order to learn which of the noisy extractions are positive and which negative.

6. CONCLUSION

We introduced two feature-based models for increasing the quality of noisy distant supervision relation extraction patterns. The first model is an extension of a hierarchical topic model using background, relation and argument-pair specific feature distributions. The second model is a perceptron trained to match an objective function that enforces an at-least-one semantics and high scores for a shared NIL-label. For both algorithms, neither explicit negative data nor the ratio of negatives has to be provided. Both algorithms give significant improvements over a maximum likelihood baseline as well as over a previous topic model without features.

Acknowledgements

Benjamin Roth is a recipient of the Google Europe Fellowship in Natural Language Processing, and this research is supported in part by this Google Fellowship.

7. REFERENCES

[1] E. Alfonseca, K. Filippova, J.-Y. Delort, and G. Garrido. Pattern learning for relation extraction with a hierarchical topic model. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistic*, 2012.

[2] M. Banko, M. J. Cafarella, S. Soderland, M. Broadhead, and O. Etzioni. Open information extraction from the web. In *International Joint Conference on Artificial Intelligence*, 2007.

[3] A. Haghighi and L. Vanderwende. Exploring content models for multi-document summarization. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, 2009.

[4] R. Hoffmann, C. Zhang, X. Ling, L. Zettlemoyer, and D. S. Weld. 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*, 2011.

[5] H. Ji and R. Grishman. Knowledge base population: Successful approaches and challenges. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 2011.

[6] M. Mintz, S. Bills, R. Snow, and D. Jurafsky. 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*, 2009.

[7] B. Roth, G. Chrupala, M. Wiegand, M. Singh, and D. Klakow. Generalizing from freebase and patterns using distant supervision for slot filling. In *Proceedings of the Text Analysis Conference (TAC)*, 2012.

[8] M. Surdeanu, J. Tibshirani, R. Nallapati, and C. D. Manning. 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*, 2012.

[9] M. Wick, K. Rohanimanesh, K. Bellare, A. Culotta, and A. McCallum. Samplerank: Training factor graphs with atomic gradients. In *Proceedings of ICML*, 2011.

[10] L. Yao, A. Haghighi, S. Riedel, and A. McCallum. Structured relation discovery using generative models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2011.

[11] L. Yao, S. Riedel, and A. McCallum. Collective cross-document relation extraction without labelled data. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, 2010.