Combining Generative and Discriminative Model Scores for Distant Supervision

Benjamin Roth, Dietrich Klakow
Saarland University
Spoken Language Systems
Saarbrücken, Germany
{benjamin.roth|dietrich.klakow}@lsv.uni-saarland.de

Abstract

Distant supervision is a scheme to generate noisy training data for relation extraction by aligning entities of a knowledge base with text. In this work we combine the output of a discriminative at-least-one learner with that of a generative hierarchical topic model to reduce the noise in distant supervision data. The combination significantly increases the ranking quality of extracted facts and achieves state-of-the-art extraction performance in an end-to-end setting. A simple linear interpolation of the model scores performs better than a parameter-free scheme based on non-dominated sorting.

1 Introduction

Relation extraction is the task of finding relational facts in unstructured text and putting them into a structured (tabularized) knowledge base. Training machine learning algorithms for relation extraction requires training data. If the set of relations is pre-specified, the training data needs to be labeled with those relations.

Manual annotation of training data is laborious and costly, however, the knowledge base may already partially be filled with instances from the relations. This is utilized by a scheme known as distant supervision (DS) (Mintz et al., 2009): text is automatically labeled by aligning (matching) pairs of entities that are contained in a knowledge base with their textual occurrences. Whenever such a match is encountered, the surrounding context (sentence) is assumed to express the relation.

This assumption, however, can fail. Consider the example given in (Takamatsu et al., 2012): If the tuple `place_of_birth(Michael Jackson, Gary)` is contained in the knowledge base, one matching context could be:

Michael Jackson *was born in* ***Gary*** *...*

And another possible context:

Michael Jackson *moved from* ***Gary*** *...*

Clearly, only the first context indeed expresses the relation and should be labeled accordingly.

Three basic approaches have been proposed to deal with noisy distant supervision instances: The *discriminative at-least-one* approach (Riedel et al., 2010), that requires that at least one of the matches for a relation-entity tuple indeed expresses the relation; The *generative* approach (Alfonseca et al., 2012) that separates relation-specific distributions from noise distributions by using hierarchical topic models; And the *pattern correlation* approach (Takamatsu et al., 2012) that assumes that contexts which match argument pairs have a high overlap in argument pairs with other patterns expressing the relation.

In this work we combine 1) a *discriminative at-least-one* learner, that requires high scores for both a dedicated noise label and the matched relation, and 2) a *generative topic model* that uses a feature-based representation to separate relation-specific patterns from background or pair-specific noise. We score surface patterns and show that combining the two approaches results in a better ranking quality of relational facts. In an end-to-end evaluation we set a threshold on the pattern scores and apply the pat-

Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, pages 24–29,
Seattle, Washington, USA, 18-21 October 2013. ©2013 Association for Computational Linguistics

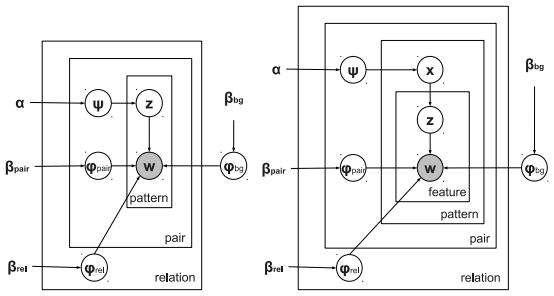

Figure 1: Hierarchical topic models. Intertext model (left) and feature model (right).

terns in a TAC KBP-style evaluation. Although the surface patterns are very simple (only strings of tokens), they achieve state-of-the-art extraction results.

2 Related Work

2.1 At-Least-One Models

The original form of distant supervision (Mintz et al., 2009) assumes all sentences containing an entity pair to be potential patterns for the relation holding between the entities. A variety of models relax this assumption and only presume that *at least one* of the entity pair occurrences is a textual manifestation of the relation. The first proposed model with an at-least-one learner is that of Riedel et al. (2010) and Yao et al. (2010). It consists of a factor graph that includes binary variables for contexts, and groups contexts together for each entity pair. MultiR (Hoffmann et al., 2011) can be viewed as a multi-label extension of (Riedel et al., 2010). A further extension is MIMLRE (Surdeanu et al., 2012), a jointly trained two-stage classification model.

2.2 Hierarchical Topic Model

The hierarchical topic model (*HierTopics*) by Alfonseca et al. (2012) models the distant supervision data by a generative model. For each corpus match of an entity pair in the knowledge base, the corresponding surface pattern is assumed to be typical for either the entity pair, the relation, or neither. This principle is then used to infer distributions over patterns of one of the following types:

1. For every entity pair, a pair-specific distribution.
2. For every relation, a relation-specific distribution.
3. A general background distribution.

The generative process assumes that for each argument pair in the knowledge base, all patterns are generated by first choosing a hidden variable z which can take on three values, B for background, R for relation and P for pair. Corresponding vocabulary distributions ($\phi_{bg}, \phi_{rel}, \phi_{pair}$) for generating the context patterns are chosen according to the value of z. The Dirichlet-smoothed vocabulary distributions are shared on the respective levels. Figure 1 shows the plate diagram of the HierTopics model.

3 Model Extensions and Combination

3.1 Generative Model

We use a feature-based extension (Roth and Klakow, 2013) of Alfonseca et al. (2012) to include bigrams for a more fine-grained representation of the patterns. For including features in the model, the model is extended with a second layer of hidden variables. A variable x represents a choice of B, R or P for every pattern, i.e. there is one variable x for every pattern. Each feature is generated conditioned on a second variable $z \in \{B, R, P\}$, i.e. there are as many variables z for a pattern as there are features for it. First, the hidden variable x is generated, then all z variables are generated for the corresponding features (see Figure 1). 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 features at indices i are mapped to the corresponding pattern indices by a function $j(i)$; p_{same} is set to .99 to enforce the correspondence between pattern and feature topics. [1]

3.2 Discriminative Model

As a second feature-based model, we employ a perceptron model that enforces constraints on the labels for patterns (Roth and Klakow, 2013). The model consists of log-linear factors for the set of relations

[1] The hyper-parameters used for the feature-based topic model are $\alpha = (1, 1, 1)$ and $\beta = (.1, .001, .001)$.

Algorithm 1 At-Least-One Perceptron Training

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

$\mathcal{R}$ as well as a factor for the *NIL* label (no relation). Probabilities for a relation r given a sentence pattern s are calculated by normalizing over log-linear factors 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 learner is directed by the following semantics: First, for a sentence s that has a distant supervision match for relation r, relation r should have a higher probability than any other relation $r' \in \mathcal{R} \setminus r$. As extractions are expected to be noisy, high probabilities for NIL are enforced by a second constraint: NIL must have a higher probability than any relation $r' \in \mathcal{R} \setminus r$. Third, at least one DS sentence for an argument pair is expected to express the corresponding relation r. For sentences s for an entity pair belonging to relation r, this can be written as the following constraints:

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

The violation of any of the above constraints triggers a perceptron update. The basic algorithm is sketched in Algorithm 1.[2]

3.3 Model Combination

The per-pattern probabilities $P(r|pat)$ are calculated as in Alfonseca et al. (2012) and aggregated over all pattern occurrences: For the topic model, the number of times the relation-specific topic has been sampled for a pattern is divided by $n(pat)$, the number of times the same pattern has been observed. Analogously for the perceptron, the number of times a pattern co-occurs with entity pairs for r is multiplied by the perceptron score and divided by $n(pat)$.

[2]The weight vectors are averaged over 20 iterations.

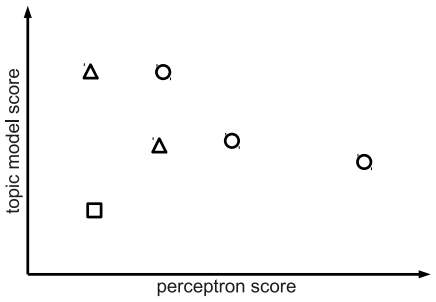

Figure 2: Score combination by non-dominated sorting: Circles indicate patterns on the Pareto-frontier, which are ranked highest. They are followed by the triangles, the square indicates the lowest ranked pattern in this example.

For the patterns of the form ***[ARG1]*** *context* ***[ARG2]***, we compute the following scores:

- **Maximum Likelihood (MLE)**: $\frac{n(pat,r)}{n(pat)}$
- **Topic Model**: $\frac{n(pat,topic(r))}{n(pat)}$
- **Perceptron**: $\frac{n(pat,r)}{n(pat)} \cdot \frac{P(r|s,\theta)}{P(r|s,\theta)+P(NIL|s,\theta)}$
- **Interpolation**: $\frac{0.5 \cdot n(pat,topic(r))}{n(pat)} + \frac{0.5 \cdot n(pat,r) \cdot P(r|s,\theta)}{n(pat) \cdot (P(r|s,\theta)+P(NIL|s,\theta))}$

The topic model and perceptron approaches are based on plausible yet fundamentally different principles of modeling noise without direct supervision. It is therefore an interesting question how complementary the models are and how much can be gained from a combination. As the two models do not use direct supervision, we also avoid tuning parameters for their combination.

We use two schemes to obtain a combined ranking from the two model scores: The first is a ranking based on non-dominated sorting by successively computing the Pareto-frontier of the 2-dimensional score vectors (Borzsony et al., 2001; Godfrey et al., 2007). The underlying principle is that all data points (patterns in our case) that are not dominated by another point[3] build the frontier and are ranked highest (see Figure 2), with ties broken by linear

[3]A data point h_1 dominates a data point h_2 if $h_1 \geq h_2$ in all metrics and $h_1 > h_2$ in at least one metric.

combination. Sorting by computing the Pareto-frontier has been applied to training machine translation systems (Duh et al., 2012) to combine the translation quality metrics BLEU, RIBES and NTER, each of which is based on different principles. In the context of machine translation it has been found to outperform a linear interpolation of the metrics and to be more stable to non-smooth metrics and non-comparable scalings. We compare non-dominated sorting with a simple linear interpolation with uniform weights.

4 Evaluation

4.1 Ranking-Based Evaluation

Evaluation is done on the ranking quality according to TAC KBP gold annotations (Ji et al., 2010) of extracted facts from all TAC KBP queries from 2009-2011 and the TAC KBP 2009-2011 corpora. First, candidate sentences are retrieved in which the query entity and a second entity with the appropriate type are contained. Candidate sentences are then used to provide answer candidates if one of the extracted patterns matches. The answer candidates are ranked according to the score of the matching pattern.

The basis for pattern extraction is the noisy DS training data of a top-3 ranked system in TAC KBP 2012 (Roth et al., 2012). The retrieval component of this system is used to obtain sentence and answer candidates (ranked according to their respective pattern scores). Evaluation results are reported as averages over per-relation results of the standard ranking metrics mean average precision (*map*), geometric map (*gmap*), precision at 5 and at 10 (*p@5*, *p@10*).

The maximum-likelihood estimator (*MLE*) baseline scores patterns by the relative frequency they occur with a certain relation. The hierarchical topic (*hier orig*) as described in Alfonseca et al. (2012) increases the scores under most metrics, however the increase is only significant for p@5 and p@10. The feature-based extension of the topic model (*hier feat*) has significantly better ranking quality. Slightly better scores are obtained by the at-least-one perceptron learner. It is interesting to see that the model combinations both by non-dominated sorting *perc+hier (pareto)* as well as uniform interpolation *perc+hier (itpl)* give a further increase in ranking quality. The simpler interpolation scheme generally works best. Figure 3 shows the Precision/Recall curves of the basic models and the linear interpolation. On the P/R curve, the linear interpolation is equal or better than the single methods on all recall levels.

method	map	gmap	p@5	p@10
MLE	.253	.142	.263	.232
hier orig	.270	.158	.353*	.297*
hier feature	.318$^{\dagger *}$	.205$^{\dagger *}$	.363*	.321*
perceptron	.330$^{\dagger *}$	.210$^{\dagger *}$	.379*	.337*
perc+hier (pareto)	.340$^{\dagger *}$	**.220**$^{\dagger *}$	.400*	.340*
perc+hier (itpl)	**.344**$^{\dagger *}$	**.220**$^{\dagger *}$	**.426**$^{\dagger *}$	**.353**$^{\dagger *}$

Table 1: Ranking quality of extracted facts. Significance (paired t-test, $p < 0.05$) w.r.t. *MLE*(*) and *hier orig*(†).

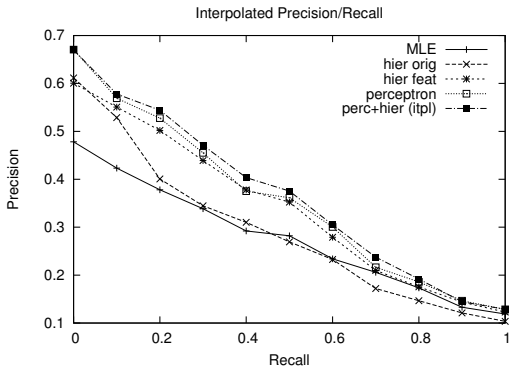

Figure 3: Precision at recall levels.

4.2 End-To-End Evaluation

We evaluate the extraction quality of the induced *perc+hier (itpl)* patterns in an end-to-end setting. We use the evaluation setting of (Surdeanu et al., 2012) and the results obtained with their pipeline for MIMLRE and their re-implementation of MultiR as a point of reference.

In Surdeanu et al. (2012) evaluation is done using a subset of queries from the TAC KBP 2010 and 2011 evaluation. The source corpus is the TAC KBP source corpus and a 2010 Wikipedia dump. Only those answers are considered in scoring that are contained in a list of possible answers from their candidates (reducing the number of gold answers from 1601 to 576 and thereby considerably increasing the value of reported recall).

For evaluating our patterns, we take the same

queries for testing as Surdeanu et al. (2012). As the document collection, we use the TAC KBP source collection and a Wikipedia dump from 07/2009 that was available to us. From this document collection, we use our retrieval pipeline of Roth et al. (2012) and take those sentences that contain query entities and slot filler candidates according to NE-tags. We filter out all candidates that are not contained in the list of candidates considered in (Surdeanu et al., 2012), and use the same reduced set of 576 gold answers as the key. We tune a single threshold parameter $t = .3$ on held-out development data and take all patterns with higher scores. Table 2 shows that results obtained with the induced patterns compare well with state-of-the-art relation extraction systems.

method	Recall	Precision	F1
MultiR	.200	.306	.242
MIMLRE	.314	.247	.277
perc+hier (itpl)	.248	.401	.307

Table 2: TAC Scores on (Surdeanu et al., 2012) queries.

4.3 Illustration: Top-Ranked Patterns

Figure 4 shows top-ranked patterns for **per:title** and **org:top_members_employees**, the two relations with most answers in the gold annotations. For maximum likelihood estimation the score is 1.0 if the patterns occurs only with the relation in question – this includes all cases where the pattern is only found once in the corpus. While this could be circumvented by frequency thresholding, we leave the long tail of the data as it is and let the algorithm deal with both frequent and infrequent patterns.

One can see that while the maximum likelihood patterns contain some reasonable relational contexts, they are less prototypical and more prone to distant supervision errors. The patterns scored high by the proposed combination generalize better, variation at the top is achieved by re-combining elements that carry relational meaning (*"is an"*, *"vice president"*, *"president director"*) or are closely correlated to the particular relation.

per:title, MLE
[ARG1] *, a singing* ***[ARG2]***
****[ARG1]*** *Best film : Capote (as* ***[ARG2]***
[ARG1] *Nunn (born October 7 , 1957 in Little Rock , Arkansas) is an American jazz* ***[ARG2]***
****[ARG2]*** *Kevin Weekes , subbing for a rarely rested* ***[ARG1]***
[ARG1] *Butterfill FRICS (born February 14 , 1941 , Surrey) is a British* ***[ARG2]***

per:title, perc+hier (itpl)
[ARG1] *, is a Canadian* ***[ARG2]***
[ARG1] *Hilligoss is an American* ***[ARG2]***
[ARG1] *, is an American film* ***[ARG2]***
[ARG1] *, is an American film and television* ***[ARG2]***
****[ARG1]*** *for Best* ***[ARG2]***

org:top_members_employees, MLE
[ARG2] *remained chairman of* ***[ARG1]***
****[ARG2]*** *asks the ball whether he and* ***[ARG1]***
[ARG2] *was chairman of the* ***[ARG1]***
****[ARG1]*** *, Joe Lieberman and* ***[ARG2]***
****[ARG1]*** *'s responsibility to pin down just how the government decided to front $ 30 billion in taxpayer dollars for the Bear Stearns deal , " Chairman* ***[ARG2]***

org:top_members_employees, perc+hier (itpl)
[ARG2] *, Vice President of the* ***[ARG1]***
[ARG1] *Vice president* ***[ARG2]***
[ARG1] *president director* ***[ARG2]***
[ARG1] *vice president director* ***[ARG2]***
[ARG1] *Board member* ***[ARG2]***

Figure 4: Top-scored patterns for maximum likelihood (MLE) and the interpolation (perc+hier itpl) method. Inexact patterns are marked by *.

5 Conclusion

We have combined two models based on distinct principles for noise reduction in distant supervision: a feature-based extension of a hierarchical topic model, and an at-least-one perceptron. Interpolation increases the quality of extractions and achieves state-of-the-art extraction performance. A combination scheme based on non-dominated sorting, that was inspired by work on combining machine translation metrics, was not as good as a simple linear combination of scores. We think that the good results motivate research into more integrated combinations of noise reduction approaches.

Acknowledgment

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.

References

Enrique Alfonseca, Katja Filippova, Jean-Yves Delort, and Guillermo Garrido. 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*, pages 54–59. Association for Computational Linguistics.

S Borzsony, Donald Kossmann, and Konrad Stocker. 2001. The skyline operator. In *Data Engineering, 2001. Proceedings. 17th International Conference on*, pages 421–430. IEEE.

Kevin Duh, Katsuhito Sudoh, Xianchao Wu, Hajime Tsukada, and Masaaki Nagata. 2012. Learning to translate with multiple objectives. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, pages 1–10. Association for Computational Linguistics.

Parke Godfrey, Ryan Shipley, and Jarek Gryz. 2007. Algorithms and analyses for maximal vector computation. *The VLDB JournalThe International Journal on Very Large Data Bases*, 16(1):5–28.

Raphael Hoffmann, Congle Zhang, Xiao Ling, Luke Zettlemoyer, and Daniel S Weld. 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, pages 541–550.

Heng Ji, Ralph Grishman, Hoa Trang Dang, Kira Griffitt, and Joe Ellis. 2010. Overview of the tac 2010 knowledge base population track. In *Third Text Analysis Conference (TAC 2010)*.

Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. 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*, pages 1003–1011. Association for Computational Linguistics.

Sebastian Riedel, Limin Yao, and Andrew McCallum. 2010. Modeling relations and their mentions without labeled text. In *Machine Learning and Knowledge Discovery in Databases*, pages 148–163. Springer.

Benjamin Roth and Dietrich Klakow. 2013. Feature-based models for improving the quality of noisy training data for relation extraction. In *Proceedings of the 22nd ACM International Conference on Information and Knowledge Management (CIKM)*. ACM.

Benjamin Roth, Grzegorz Chrupala, Michael Wiegand, Mittul Singh, and Dietrich Klakow. 2012. Generalizing from freebase and patterns using distant supervision for slot filling. In *Proceedings of the Text Analysis Conference (TAC)*.

Mihai Surdeanu, Julie Tibshirani, Ramesh Nallapati, and Christopher D Manning. 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*, pages 455–465. Association for Computational Linguistics.

Shingo Takamatsu, Issei Sato, and Hiroshi Nakagawa. 2012. Reducing wrong labels in distant supervision for relation extraction. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers - Volume 1*, ACL '12, pages 721–729, Stroudsburg, PA, USA. Association for Computational Linguistics.

Limin Yao, Sebastian Riedel, and Andrew McCallum. 2010. Collective cross-document relation extraction without labelled data. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 1013–1023. Association for Computational Linguistics.