A Survey of Noise Reduction Methods for Distant Supervision

Benjamin Roth, Tassilo Barth, Michael Wiegand, Dietrich Klakow
Saarland University
Spoken Language Systems
Saarbrücken, Germany
{benjamin.roth|tbarth|michael.wiegand|dietrich.klakow}@lsv.uni-saarland.de

ABSTRACT

We survey recent approaches to noise reduction in distant supervision learning for relation extraction. We group them according to the principles they are based on: at-least-one constraints, topic-based models, or pattern correlations. Besides describing them, we illustrate the fundamental differences and attempt to give an outlook to potentially fruitful further research. In addition, we identify related work in sentiment analysis which could profit from approaches to noise reduction.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing

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) [8]. 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 requires 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)*, [10]: Textual matches of entities from fact tuples are used to automatically generate relation contexts as training instances, Figure 1 shows the basic assumed workflow for distant supervision.

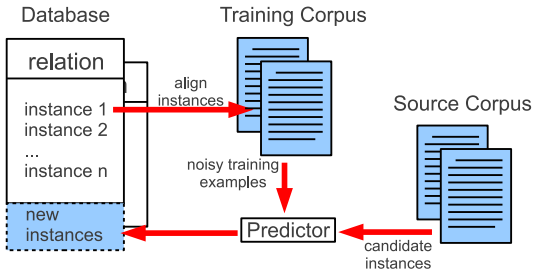

Figure 1: Distant supervision for knowledge base population.

Often only a small fraction of such matches indeed express the relation of the fact tuple. For example, the arguments of the fact tuple *("Barack Obama", born-in, "Honululu")* could match in *true positive* contexts like *"Barack Obama was born in Honululu"*, as well as *false positive* contexts like *"Barack Obama visited Honululu"*.

A number of different approaches have been introduced to automatically determine which training contexts, obtained from relation argument matching, are *true positives*, and which are *false positives*. This paper aims at giving an overview of approaches tackling this problem (cf. Table 1). They are each based on one of the following principles:

- *At-least-one* constraints state that at least one positively classified context is indeed a true positive – and not necessarily all of them (see Section 2.1). We deem it potentially fruitful to further research to contrast the at-least-one principle to other schemes applied in prediction (Section 2.2).
- *Hierarchical topic models* are based on the idea of separating the distributions that generate relation-specific contexts from those generating pair-specific contexts or background text (Section 3).
- *Pattern correlations* are at the heart of an approach which assumes that contexts matching argument pairs for a relation either express that relation, or have a high overlap in argument pairs with other patterns expressing the relation. In other words, they explicitly model the fact that a pattern is matching and exploit this to transfer probability mass to similar patterns (Section 4).

While noise reduction for distant supervision has been

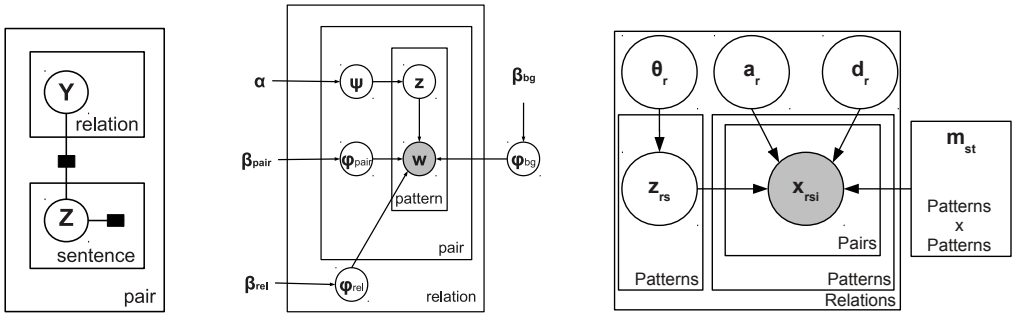

Figure 2: ***Left:*** **MultiR in plate notation. Mention and relation variables are connected by a dedicated factor that is 1 iff the at-least-one assumption is fulfilled.** ***Middle:*** **Hierarchical topic model for DS. Context patterns are generated by either a background, relation-specific or pair-specific distribution.** ***Right:*** **Plate diagram for the Takamatsu model. The hidden variables z_{rs} indicate whether a relation r is expressed by a pattern s. The observed variables x_{rsi} denote which contexts are matched by an argument pair i. θ_r, a_r and d_r are the parameters to be learned, m contains the correlation statistics.**

	Type	**Baseline**	**KB** (relations)	**Ground Truth**	**Corpus**
Riedel 2010, Yao 2010	At-least-one	plain distant supervision, Joint model w/o at-least-one	Freebase (430)	Freebase, Human ratings	New York Times
Hoffmann 2011	At-least-one	Riedel 2010	Riedel 2010	Freebase, Human ratings	New York Times
Surdeanu 2012	At-least-one	plain distant supervision, Riedel 2010, Hoffmann 2011	Riedel 2010, Wikipedia infoboxes (42)	Freebase, TAC key (2010+2011)	New York Times, TAC KBP corpus (2010+2011)
Alfonseca 2012	Topic Model	MLE of $P(rel\|pattern)$	Freebase (3)	Human ratings	Web news articles
Takamatsu 2012	Pattern Correlation	plain distant supervision, MLE, MultiR (Hoffmann)	Freebase (24)	Freebase, Human ratings	Wikipedia
Roth 2013	At-least-one, Topic Model	MLE, Alfonseca 2012	Seeds for TAC relations (42)	TAC key (2009-2011)	TAC KBP corpus (2009-2011)

Table 1: Overview over the experimental settings of the approaches covered in this survey.

mostly studied for relation extraction, it may also be of interest to other areas in which training data is generated by using sets of easily available seeds. We will briefly point to some related work in the field of sentiment analysis in section 5.

2. AT-LEAST-ONE MODELS

Normally, distant supervision assumes all sentences containing an entity pair to be potential patterns for the relation holding between the entities. As found by [13], this assumption quickly becomes untenable when dealing with text data not directly associated with the knowledge base the facts are taken from. In the following, we describe approaches implementing a relaxing constraint which only presumes that at least one of the entity pair occurrences is a textual manifestation of the relation (*at-least-one* assumption).

2.1 Existing Models

Formally, the at-least-one assumption states that *"If two entities participate in a relation, at least one sentence that mentions these two entities might express that relation"* [13]

Various models are fundamentally based on this idea ([13], [18], [7], [15], [14]). Relation classification models are trained with an objective function that includes this constraint. Typically, at-least-one models are multi-class models over a set of relations, including a special NIL label to indicate that none of relations in the knowledge base is expressed by a context.

While the underlying idea regarding noise reduction is the same for all of those models, they differ in other assumptions about dependencies in the data, at what point the at-least-one constraint is used, and in their inference algorithms. The first proposed model with an at-least-one learner is that of Riedel et al. [13]. It consists of a factor graph that includes binary variables for contexts, and groups contexts together for each entity pair. An entity pair is associated with a variable that can take on a relation value or NIL. A global objective function penalizes the violations of at-least-one constraints, and SampleRank is used to infer the model.

MultiR [7] (see Figure 2) can be viewed as a multi-label extension of [13]. Given an entity pair, the model can predict multiple ("overlapping") relations simultaneously; MultiR uses a perceptron training scheme. A further extension is MIMLRE [15], a jointly trained two-stage classification model. MIMLRE, on one layer, makes multi-class predic-

tions for contexts. The predictions of this layer are used by a collection of binary per-relation classifiers to predict the labels for an entity pair. The at-least-one semantics is brought into the model by a special feature in the per-relation classifiers. Most at-least-one approaches require dedicated negative training data to estimate enough probability mass for the NIL class. An at-least-one model that does not require negative training data while enforcing additional ranking constraints for NIL is used in [14].

2.2 Connection to Redundancy Models

Many relation extraction systems decide whether a fact is extracted or not at prediction time according to the following simple rule: A fact is extracted if and only if there is a positive decision for at least one context. This decision rule is mirrored on the training side for at-least-one-context training. A straightforward continuous generalization of this rule is to assign a score by noisy-or [9]. MIMLRE [15] for example, uses an at-least-one-context scheme for training, but noisy-or for prediction. At-least-one-context and noisy-or schemes are simple examples of redundancy models, i.e. models that combine scores for several instances to an overall prediction. While at-least-one-context models have been extensively studied for training – equivalent to at-least-one prediction – less work has been done on noisy-or training (Takamatsu et. al [17] use noisy-or in their correlation calculation).

Both views (at-least-one and noisy-or) do not consider the number of contexts for a fact triple scored low by the model, instead such objective functions tend to only consider, for each candidate triple, the contexts that are given a high model score. The overall number of contexts for a candidate tuple is not included in the model – large numbers of contexts that are given a low probability for the relation do not influence the score negatively. This has been identified as a problem for prediction by Downey et al. [4], it led to the development of the probabilistic URNS model which expects particular minimal ratios of *true* and *false* contexts, depending on the number of contexts for a fact. We assume similar models could be beneficial during training by relaxing the at-least-one constraint for singleton tuples and requiring more positive instances for frequently matching tuples[1].

To summarize, we believe that only the most simple redundancy model, at-least-one, has been extensively applied to training with distant supervision data. Redundancy models with more connections to probability theory – such as noisy-or or URNS – remain largely unexplored.

3. HIERARCHICAL TOPIC MODELS

The hierarchical topic model (*HierTopics*) presented by [1] is a generative model. It assumes that a context pattern matching an entity pair in the knowledge base for a particular relation is either typical for the entity pair, the relation, or neither. This principle is then used to infer distributions of one of the following types:

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

The model is the hierarchical topic model for multi-document summarization of [6]. Pairs of arguments are assumed to form documents, with the surface patterns as their words. Also, the pairs are grouped together according to the relation they stand in.

The generative process assumes that for each argument pair of a particular relation, all patterns (surface strings or dependency paths between arguments from distant supervision matches) are generated by first choosing 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 to generate 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 2 for a plate diagram of the HierTopics model. Gibbs sampling is used to infer the topics of the document collection.

The HierTopics model aims at separating out the relation vocabulary in an efficient and theoretically appealing way. Compared to at-least-one models it allows for a desired degree of freedom: The amount of positively labeled contexts per entity pair is dependent on the vocabulary and not on fixed ratios or minimal numbers. Moreover, it does not require dedicated negative training data. On the other hand, at-least-one learning has been demonstrated to be an apt building block (e.g. as a dedicated node in a factor graph) in more complex models using effective discriminative training schemes.

The original hierarchical topic model [1] treats patterns as a whole. In [14] the model is extended by a second layer of hidden variables in order to include bigram features to improve estimates for the long tail of infrequent patterns. A comparison to an at-least-one perceptron learner shows that while the simple surface-pattern topic model version is better than a maximum likelihood baseline, it is not as good as the at-least-one model. By including features in the hierarchical topic model its performance comes very close to that of the perceptron.

4. PATTERN CORRELATIONS

While *HierTopics* models the generative process of the distant supervision corpus and then obtains information about relevance of patterns as a by-product, Takamatsu et al. [17] aim more directly at modeling whether a pattern expresses a relation or not. The underlying idea is that contexts matching argument-pairs for a relation either express that relation, or have a high overlap in argument pairs with other patterns expressing the relation (or, none of the two, which is covered by an additional constant probability). The arguments of patterns that express a relation may still frequently co-occur with other patterns that do not express the relation.

To give an example, given some patterns *s="[ARG1] and [ARG2]"* and *t="[ARG1] is the wife of [ARG2]"*, if there is a context

[1] One anonymous reviewer pointed out that when the mention variables are marginalised out, many low-probability mentions alone would even increase the relation probability in an at-least-one setting: The more low-probability mentions there are, the higher the probability that at least one of them is 'active' and that the relation variable is true.

"[Michelle Obama] and [Barack Obama]" = s([MO], [BO])

the context (or, rather its pattern s) can be labeled negative for a relation *spouse_of* if pattern t is labeled positive and $P(pair \in s|pair \in t)$ is high. Note that it is not necessary that the actual context

"[Michelle Obama] is the wife of [Barack Obama]" = t([MO], [BO])

is present in the training data. This is a major difference to at-least-one training schemes. To give a different example, a negative label for

"[Michelle Obama] is the wife of [Barack Obama]" = t([MO], [BO])

could not be explained by a positive label for *"[ARG1] and [ARG2]"* if $P(pair \in t|pair \in s)$ is small. The pattern co-occurrence probabilities are calculated prior to inference based on the overlap of sets of entity pairs matched by the patterns.

A probabilistic graphical model (see Figure 2) is learned that contains a hidden variables z_{rs} indicating whether a pattern s indeed expresses a relation r. The topology of the model is different from *HierTopics*: Although the observed variables are tuples of patterns and argument pairs in both cases, Takamatsu et al. group the contexts by patterns and do not consider repeated occurrences of contexts.

The rationale behind the probabilistic process is the following: If a tuple of a relational pattern s and argument pair i is observed, and argument pair i is in the knowledge base, then this can have one of the following causes:

1. Pattern s expresses relation r, i.e. $z_{rs} = true$.
2. Pattern s does not express relation r – however, some other pattern t expresses r and arguments of t are often arguments of s, i.e. $z_{rt} = true$ and $P(pair \in s|pair \in t)$ is high.
3. Pattern s does not express relation r – however, the existence of fact i in the knowledge base is explained by some other process not captured by the model.

That is, the model deals separately with case one, when the underlying variable for the pattern directly expresses the fact in the knowledge base (relation r holds for the argument pair), and cases two and three, when the argument pair is in the knowledge base but the pattern does not express r. For case two, it is not necessary that another pattern occurs with argument pair i, as it would be the case in an at-least-one setting. This way, the model can hypothesize whether an entity pair i could have been generated by another pattern t expressing r, even if t and i have never been observed together in the corpus.

5. DISTANT SUPERVISION IN SENTIMENT ANALYSIS

In this section, we take a brief look at another domain that has employed distant supervision in order to give further evidence to the general importance of this paradigm. Sentiment analysis is a domain that most heavily makes use of distant supervision. Note that since the term "distant supervision" was not coined before Mintz in 2009 [10], the early works in sentiment analysis prior to that date do not explicitly refer to this methodology as "distant supervision". Distant supervision is so popular in sentiment analysis due to the textual source on which it is most frequently applied, namely social media, which contains special properties that can be harnessed for acquiring training data. Most tasks in sentiment analysis differ from the previously mentioned works in that no specific relations are extracted but a text (be it a document, a sentence or phrase) is classified with regard to subjectivity, polarity or even more fine-grained distinctions. Therefore, the "distant supervision" methodology is notably different: A common subtask is the distinction of positive and negative polarity. Early works applied this task on movie reviews from the web written by common users. A very popular method to acquire the labels for the pertaining training data is by using the scores the reviewers assigned to their reviews as a proxy [11, 3]. Typically, a scale of 5 points/stars is employed where 1 point/star is the lowest score and 5 points/stars is the highest score that a reviewer may assign. From that information, one can derive that reviews with 1 or 2 points/stars are negative while 4 or 5 points/stars are positive.

In the more recent subtask emotion classification, tweets have been used as a textual source that may serve as training data for this classification task [5, 12, 16]. Tweets comprise emoticons and hashtags that heavily correlate with certain types of emotions that one wants to automatically predict. For instance, "anger" usually highly correlates with the hashtag `#angry` or "happiness" usually highly correlates with `:-)`.

Even though these applications of distant supervision in sentiment analysis are pretty simple methods, they are very effective. So far, special methods tailored to reduce noise have not been employed, so it is unclear what their impact would be.

6. CONCLUSION

Distant supervision allows for cheap creation of large amounts of training data and has recently been extensively studied in the context of relation extraction. As the training data obtained is inherently noisy, the most challenging problem in this context is to improve the quality of the training data by reducing the amount of noise. In this paper, we have presented a survey of several approaches that have been undertaken to this end. The approaches are based on different underlying ideas: First, a big class of proposed models is based on the principle that it is necessary and sufficient to assume that at least one context expresses a fact in the knowledge base. Second, hierarchical topic models estimate different distributions for background, relation-specific, and pair-specific contexts. A third approach employs argument correlations between patterns. Further work could explore extensions to models, e.g., redundancy models, as well as transferring successful approaches to new applications (sentiment analysis). We hope that this survey gives a due representation of the current state of the art in distant supervision and inspires further research.

Acknowledgements

Benjamin Roth is supported by a Google Europe Fellowship in Natural Language Processing. Tassilo Barth was supported in part by IARPA contract number W911NF-12-C-0015 and Michael Wiegand by the German Federal Ministry of Education and Research (BMBF) under grant no. "01IC10S01".

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 Linguistics: Short Papers-Volume 2*, pages 54–59. Association for Computational Linguistics, 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] J. Blitzer, M. Dredze, and F. Pereira. Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *Annual Meeting-Association For Computational Linguistics*, volume 45, page 440, 2007.

[4] D. Downey, O. Etzioni, and S. Soderland. A probabilistic model of redundancy in information extraction. In *IN IJCAI*, pages 1034–1041, 2005.

[5] A. Go, R. Bhayani, and L. Huang. Twitter sentiment classification using distant supervision. *CS224N Project Report, Stanford*, pages 1–12, 2009.

[6] 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*, pages 362–370. Association for Computational Linguistics, 2009.

[7] 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*, volume 1, pages 541–550, 2011.

[8] 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*, volume 1, pages 1148–1158, 2011.

[9] W. Lin, R. Yangarber, and R. Grishman. Bootstrapped learning of semantic classes from positive and negative examples. In *Proceedings of ICML-2003 Workshop on The Continuum from Labeled to Unlabeled Data*, volume 1, page 21, 2003.

[10] 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: Volume 2-Volume 2*, pages 1003–1011. Association for Computational Linguistics, 2009.

[11] B. Pang, L. Lee, and S. Vaithyanathan. Thumbs up?: sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*, pages 79–86. Association for Computational Linguistics, 2002.

[12] M. Purver and S. Battersby. Experimenting with distant supervision for emotion classification. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 482–491. Association for Computational Linguistics, 2012.

[13] S. Riedel, L. Yao, and A. McCallum. Modeling relations and their mentions without labeled text. In *Machine Learning and Knowledge Discovery in Databases*, pages 148–163. Springer, 2010.

[14] B. Roth and D. Klakow. 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, 2013.

[15] 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*, pages 455–465. Association for Computational Linguistics, 2012.

[16] J. Suttles and N. Ide. Distant supervision for emotion classification with discrete binary values. In *Computational Linguistics and Intelligent Text Processing*, pages 121–136. Springer, 2013.

[17] S. Takamatsu, I. Sato, and H. Nakagawa. 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, 2012. Association for Computational Linguistics.

[18] 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*, pages 1013–1023. Association for Computational Linguistics, 2010.