Using Active Learning and Semantic Clustering for Noise Reduction in Distant Supervision

Lucas Sterckx, Thomas Demeester, Johannes Deleu, Chris Develder
Ghent University - iMinds
Ghent - Belgium
firstname.lastname@intec.ugent.be

Abstract

The use of external databases to generate training data, also known as Distant Supervision, has become an effective way to train supervised relation extractors but this approach inherently suffers from noise. In this paper we propose a method for noise reduction in distantly supervised training data, using a discriminative classifier and semantic similarity between the contexts of the training examples. We describe an active learning strategy which exploits hierarchical clustering of the candidate training samples. To further improve the effectiveness of this approach, we study the use of several methods for dimensionality reduction of the training samples. We find that semantic clustering of training data combined with cluster-based active learning allows filtering the training data, hence facilitating the creation of a clean training set for relation extraction, at a reduced manual labeling cost.

1 Introduction

For the task of extracting relations between entities according to a fixed schema (also known as Knowledge Base Population (KBP)), distantly supervised approaches are currently state-of-the-art [19]. A requisite for the effectiveness of these techniques is the availability of labeled data, which is expensive to obtain. An approach to solve this issue and produce large quantities of training data is distant supervision (DS) [11]. DS creates labeled data using readily available repositories like FreeBase or DBpedia with facts like "*Person* $\rightarrow$ *city-of-residence*$\rightarrow$ *Location*" (for the remainder denoted as "*per:city_of_residence*") and the assumption that every phrase mentioning both entities participating in the relation expresses the corresponding relation from the database. Using this approach, a large quantity of training data can be generated automatically. However, intuitively this assumption only holds for a fraction of the extracted mentions, as two entities may co-occur in one sentence for many alternative reasons. A challenge we address in this paper is to develop strategies to improve the quality of the training data and reduce the amount of noise.

In our participation in the Text Analysis Conference for Knowledge Base Population (TAC-KBP) slot filling track organized by NIST [1], a baseline supervised classification using DS was implemented as described in [6]. In our submissions, we already showed the value of noise reduction based on straightforward human annotation of randomly selected training instances: cleaning based on a classifier (trained on the annotated instances) resulted in 8% higher precision. As this required extra manual annotation of the training samples, we search for an efficient way to query the distantly supervised data and train a classifier using a minimal amount of supervision but an improved noise reduction.

This work contributes by presenting a strategy for noise reduction using a supervised classifier trained using labeled mentions from distantly supervised data. By incorporating semantic relatedness between the mentions we can use an active learning approach which exploits the resulting

clustering of training data. Intelligent querying of training data clusters and assigning labels to similar unknown training examples trains a classifier based on less human supervision while optimizing the capability of separating noisy from true relation contexts.

2 Related Work

The approach of DS was first presented by Mintz et al. [11] for training of binary Support Vector Machines which used a set of lexical and non-lexical features for classification. Since then, several methods for noise reduction of the data have been proposed. For a recent survey we refer to Roth et al. [16]. Models like the one proposed by Riedel et al. [14], MultiR [10] and MIML-RE [20] involve latent variables which model the assumption that at least one generated example for an entity pair and relation is a true positive, or apply a generative model [21]. Our approach is less complex, using a discriminative classifier based on manually annotated examples of true positive and false positive relation mentions within each of the generated training sets. Using this classifier, we filter training data explicitly, independent of the entities involved and for each relation separately, solely based on the surface text.

Recent work has combined DS with small amounts of labeled data, these labels are either included directly in a latent variable model [13] or used in an active learning setting. Active learning was previously performed in relation extraction by Sun and Grishman [18] for extending a relation extraction system to recognize a new type of relation. An approach which uses active learning for DS was recently proposed by Angeli et al. [2] and successfully applied in the top performing system in the TAC slot filling competition [19]: they show that a small number of examples can yield improvements in end-to-end accuracy of the relation extraction using several approaches from active learning literature and a new metric incorporating uncertainty and representativeness. Our work differs from this and others in that we use a cluster based active learning approach, evaluating directly on a set of labeled training examples.

3 Semantic-Cluster-Aware Sampling

Our approach assumes that true positive mentions within each training set are similar to each other, in terms of text and meaning, and tend to cluster together, unlike false positive mentions which are less similar and more diverse. This inspired us for the application of cluster-aware sampling of the training data for the training of the noise-reduction classifier. An active learning approach that exploits cluster structure in data was introduced by Dasgupta and Hsu [4]. This algorithm takes a pool of unlabeled data and performs hierarchical clustering, resulting in a tree structure of data points. The algorithm starts out by randomly querying data points in the vector space and searches for a pruning of the tree that optimizes the pureness of each cluster. Each iteration, a number of data points are sampled in such a way that less pure clusters are more likely to be sampled from and unseen samples receive the label of the majority of known samples in the cluster it belongs to.

As stated in the original paper [4], the algorithm is most effective when pure clusters are found at the top of the hierarchical tree. Thus, when applied for noise reduction, this approach benefits from relation contexts that are clustered according to the meaning or relation they express. The simple bag-of-words representation results in a high dimensionality of the relation contexts with only few ways of clustering contexts with similar meaning. We need a transformation of the contexts into a vector space of reduced dimension, with those having a similar expression of relation being transformed into similar representations. This is exactly what semantic clustering achieves, i.e. clustering contexts according to meaning.

Semantic clustering of relations has been performed on several occasions in the context of Open Information Extraction to cluster output having similar meaning and is related to the task of paraphrase and synonym detection [9, 23, 24, 15]. We use a transformation based on a simple composition of the words participating in the context. While much research has been directed at ways of constructing distributional representations of individual words, for example co-occurrence based representations and word embeddings, there has been far less consensus regarding the representation of larger constructions such as phrases and sentences from these representations. Blacoe et al. [3] show that a simple composition like addition or multiplication of the distributional word representations is competitive with more complex operations like the Recursive Neural Networks proposed by Socher et

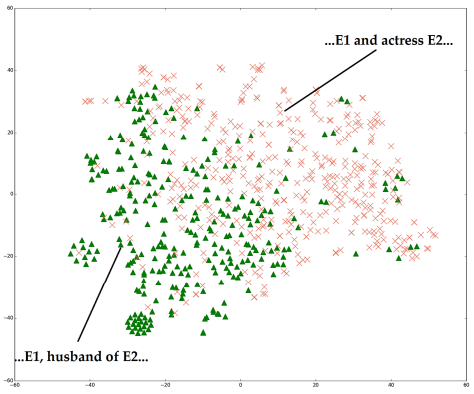

Figure 1: Visualization of relation contexts in a semantic vector space for relation "*per:spouse_of*".

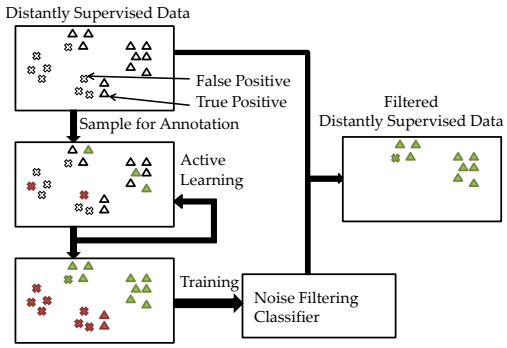

Figure 2: Methodology for filtering of noisy mentions.

al. [17] for detection of paraphrases and synonyms. We chose to ignore word order and sum all distributional representations from words participating in the surrounding context of the mention and normalize them.

4 Experiments and Results

We use FreeBase [7] as our source of fact relations by first matching the schema from the TAC-KBP to fields from FreeBase. The participating entities are matched in phrases from articles from the English GigaWord corpus [8]. As part of a participation in the TAC-KBP slot filing competition, a team of students was asked to assign 2.000 training samples with a *True* or *False* label with respect to the 2014 TAC-annotation guidelines for a selection of 12 relations with a large quantity of training data. As these samples were selected at random, some of the relations contained very pure or highly noisy training sets. Phrases were filtered for duplicates and entity names were removed from the surface text.

Effective representations should be able to separate true examples of a relation being expressed from false examples. We visualize this in Figure 1 for the relation *per:spouse_of*. Words in between the subject and object entities of the relation are transformed to their semantic vectors using word embeddings which are summed and normalized. In our experiments we use the GloVe word-embeddings with 100 dimensions trained on Wikipedia text [12], which are made available from the authors' website.[1] The resulting sentence representations are clustered and represented in a two-dimensional space using the t-SNE algorithm [22] in Figure 1. True examples of the relation are represented in Figure 1 as dark triangles, while false examples are the lighter crosses. The resulting figure shows that this basic transformation alone is able to capture some of the semantic meaning of relations.

The active learning strategy is performed on 70% of the DS-data, 30% is set aside to evaluate classification. The general methodology for filtering distantly supervised data is shown in Figure 2. Previously described active learning iteratively only queries a number of DS-examples, but results in a fully labeled distantly supervised data set (each unknown sample then receives the label of the majority of its cluster). The resulting fully labeled DS-data is used to train a logistic regression classifier using only word count vectors as features in a basic text classification setting to filter the training data. At most two words before the entity first mentioned, the words in between the entities and at most two words following the entity mentioned last are included. We compare our cluster-based active learning in the semantic vector space, with uniform sampling, clustering using Bag-of-Words vectors and clustering after transformation using Latent Semantic Indexing (LSI) [5] (also for 100 dimensions). This process is repeated 20 times using stratified cross-folds.

[1] `http://nlp.stanford.edu/projects/glove/`

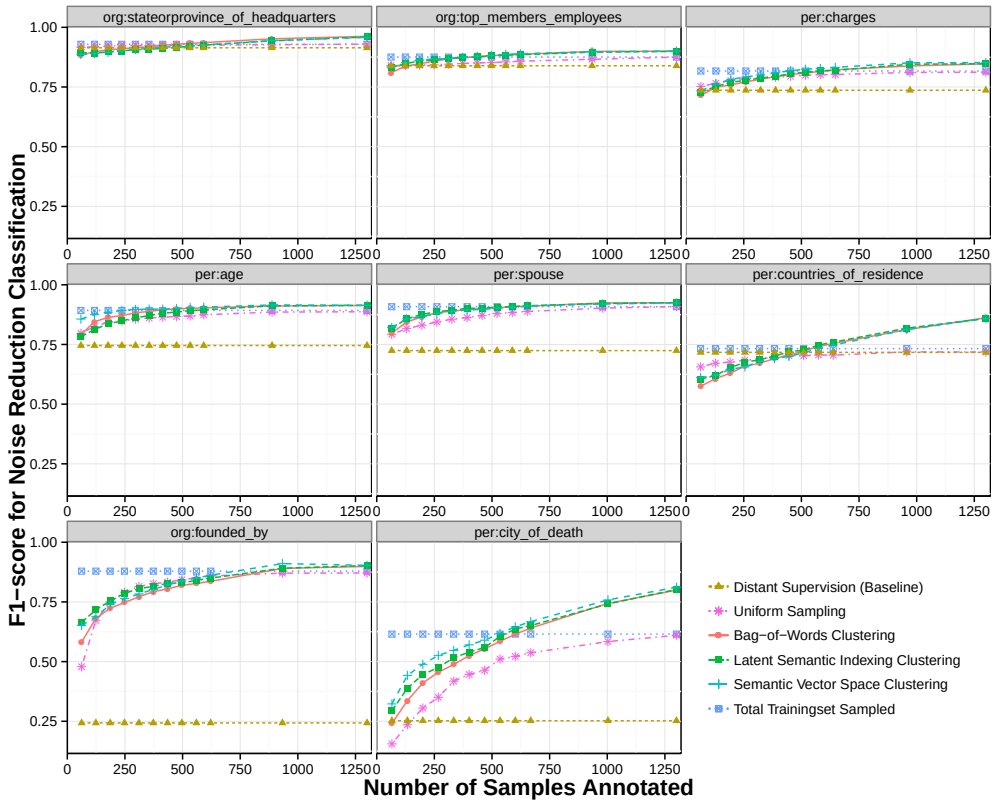

Figure 3: Performance of cluster-based active learning approach.

Table 1: Macro-average filter performance using 70 labeled distantly supervised training examples

	Precision	Recall	F1
Distant Supervision (Baseline)	51.9	100.0	60.8
Random Sampling	72.0	**72.8**	66.0
Bag-of-Words Clustering	73.4	65.2	66.6
Latent Semantic Indexing Clustering	73.7	68.5	68.3
Semantic Vector Space Clustering	**74.6**	71.4	**71.2**

Using all of the labeled data, supervised noise reduction is able to increase average fraction of true positive of the DS training set from 47% to 84% while maintaining a recall of 88%. Noise reduction using active learning for a selection of relations is presented in Figure 3. Separately for each relation, after each increase of 5% sampled data we calculate the averaged F1 score of the classification for each of the strategies. Performance of the noise reduction is highly dependent on the relation. For relatively pure training sets, as is the case for relation "*org:state_or_province_of_headquarters*", with more than 85% of the training data being positive examples, are hard to filter. For these relations supervised filtering appears ineffective or even detrimental, others need a minimum amount of samples to benefit like "*per:countries_of_residence*". Cluster aware active learning is an effective strategy for a majority of the noisy relations, converging faster to the optimal performance of filtering. Overall, performance using semantic clustering of contexts is slightly better than using LSI clustering, while with very few samples and relations like "*per:age*" and "*per:city_of_death*" performance increase is larger. Another observation is that, because the algorithm also provides the test samples with a label (based on the majority of the labels in the same cluster as the test sample), classification performance surpasses that of a fully labeled training set while approximately only

half is sampled. Table 1 shows macro average precision, recall and F1 using a minimal amount of only 70 samples for noisy relations (fraction of true positives less than 85%).

5 Conclusion

In this paper we presented a novel approach for filtering a distantly supervised training set by building a binary classifier to detect true relation mentions, the classifier is trained using a cluster based active learning strategy. We show that clustering of relation mentions and adding semantic information reduces human effort and makes this a promising approach more feasible to filter a wide variety of relations. For future work we suggest the use of more sophisticated methods which take into account composition for transforming context to a semantic vector space.

References

[1] Task Description for English Slot Filling at TAC-KBP. 2014.

[2] Gabor Angeli, Julie Tibshirani, JY Wu, and CD Manning. Combining Distant and Partial Supervision for Relation Extraction. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2014.

[3] William Blacoe and Mirella Lapata. A Comparison of Vector-based Representations for Semantic Composition. *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, 2012.

[4] Sanjoy Dasgupta and Daniel Hsu. Hierarchical sampling for active learning. In *Proceedings of the 25th international conference on Machine learning*, 2008.

[5] Scott C. Deerwester, Susan T Dumais, Thomas K. Landauer, George W. Furnas, and Richard A. Harshman. Indexing by latent semantic analysis. volume 41, pages 391–407, 1990.

[6] Matthias Feys, Lucas Sterckx, Laurent Mertens, Johannes Deleu, Thomas Demeester, and Chris Develder. Ghent University-IBCN Participation in TAC-KBP 2014 Slot Filling and Cold Start Tasks. 2014.

[7] Google. Freebase data dumps. `https://developers.google.com/freebase/data`, 2014.

[8] David Graff, Junbo Kong, Ke Chen, and Kazuaki Maeda. English gigaword. *Linguistic Data Consortium, Philadelphia*, 2003.

[9] Takaaki Hasegawa, Satoshi Sekine, and Ralph Grishman. Discovering relations among named entities from large corpora. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, 2004.

[10] Raphael Hoffmann, Congle Zhang, Xiao Ling, Luke Zettlemoyer, and Daniel 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. Association for Computational Linguistics, 2011.

[11] Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL*, pages 1003–1011. Association for Computational Linguistics, 2009.

[12] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2014.

[13] Maria Pershina, Bonan Min, Wei Xu, and Ralph Grishman. Infusion of labeled data into distant supervision for relation extraction. In *Proceedings of the 2014 Conference of the Association for Computational Linguistics (ACL 2014)*, Baltimore, US, June 2014. Association for Computational Linguistics.

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

[15] Lorenza Romano, Milen Kouylekov, Idan Szpektor, Ido Dagan, and Alberto Lavelli. Investigating a generic paraphrase-based approach for relation extraction. 2006.

[16] Benjamin Roth, Tassilo Barth, Michael Wiegand, and Dietrich Klakow. A survey of noise reduction methods for distant supervision. In *Proceedings of the 2013 Workshop on Automated Knowledge Base Construction*, AKBC '13, pages 73–78, New York, NY, USA, 2013. ACM.

[17] Richard Socher, Brody Huval, Christopher D. Manning, and Andrew Y. Ng. Semantic compositionality through recursive matrix-vector spaces. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, 2012.

[18] Ang Sun and Ralph Grishman. Active learning for relation type extension with local and global data views. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, CIKM '12, pages 1105–1112, New York, NY, USA, 2012. ACM.

[19] Mihai Surdeanu and Heng Ji. Overview of the English Slot Filling Track at the TAC2014 Knowledge Base Population Evaluation. 2014.

[20] Mihai Surdeanu, Julie Tibshirani, Ramesh Nallapati, and Christopher D Manning. Multi-instance multi-label learning for relation extraction. pages 455–465, 2012.

[21] Shingo Takamatsu, Issei Sato, and Hiroshi 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*, pages 721–729. Association for Computational Linguistics, 2012.

[22] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 2008.

[23] Wei Wang, Romaric Besanon, Olivier Ferret, and Brigitte Grau. Semantic clustering of relations between named entities. In Adam Przepirkowski and Maciej Ogrodniczuk, editors, *Advances in Natural Language Processing*, volume 8686 of *Lecture Notes in Computer Science*, pages 358–370. Springer International Publishing, 2014.

[24] Limin Yao, Aria Haghighi, Sebastian Riedel, and Andrew McCallum. Structured relation discovery using generative models. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, 2011.