Overcoming Data Sparsity for Relation Detection in German Novels

Markus Krug
markus.krug@uni-wuerzburg.de
University of Wuerzburg, Germany

Isabella Reger
isabella.reger@uni-wuerzburg.de
University of Wuerzburg, Germany

Fotis Jannidis
fotis.jannidis@uni-wuerzburg.de
University of Wuerzburg, Germany

Lukas Weimer
lukas.weimer@uni-wuerzburg.de
University of Wuerzburg, Germany

Nathalie Madarász
nathalie.madarasz@stud-mail.uni-wuerzburg.de
University of Wuerzburg, Germany

Frank Puppe
frank.puppe@uni-wuerzburg.de
University of Wuerzburg, Germany

Introduction

Within the context of social network analysis (SNA) for literary texts the automatic detection of family relations and similar social relations between characters in novels would be an important step for any macroscopic analysis. Manual labeling is rather inefficient since the text snippets that explicitly describe a relation are sparse within the long text documents; therefore we combine two techniques, active learning and distant supervision, which are often used to overcome data sparsity.

Inspired by distant supervision which uses a high quality information resource to support information extraction from other data, we used expert summaries of literary texts, German novels mainly from the 19th century, since relevant text snippets are much more frequent in summaries. Then we applied an uncertainty-based active learning strategy labeling selected sentences from the novels and the complete summaries. The results show that training on summaries and evaluating on data derived from novels yields reasonable results with high precision and low recall similar to humans solving this task.

After a brief discussion of related work in the next section, the data set and the necessary preprocessing for this work are explained in section three. Section four describes our method in detail and shows strengths and weaknesses.

Related work

The challenge of training an algorithm capable of generalizing from a small set of manually labeled data has created a multitude of approaches like active learning and distant supervision. A good survey on active learning is given in (Finn et al., 2003). Usually it starts with a seed set of manually annotated data. A classifier is then trained and new instances that appear to be very different from the current training data are proposed for manual labeling until the quality of the classifier stops improving. Successful algorithms include Multi-instance Multi-label Relation Extraction (Surdeanu et al., 2012).

Another method specifically used for relation detection in newspapers is distant supervision (Mintz et al., 2009): Given some facts (e.g. Michelle Obama is the wife of Barack Obama), usually stored in a database, the aim is to match those facts to the text (e.g. every sentence containing Michelle and Barack Obama indicates that they are married). The training of the classifier is then performed on the pseudo gold data. Even though the idea appears to be simplistic, the results are comparable to those obtained by active learning.

Jing et al. (Jing et al. 2007) successfully applied relation extraction for SNA in an end-to-end manner and reported that most problems were caused by coreference resolution.

Data and preprocessing

We created three datasets from 213 expert summaries, available from Kindler Literary Lexicon Online, and 1700 novels derived from project Gutenberg and annotated relations between characters:

- We split 500 novels into sentences and applied an uncertainty-based active learning strategy (explained below) to iteratively select new examples (in this case full sentences) using a MaxEnt classifier. In total, about 1100 sentences were labeled in this way. This was labeled by annotator 1. (From now on, we refer to this as the novel data set)

- We split our summaries into sentences and applied the same active learning strategy to select new examples, thereby generating about 1300 labeled sentences. They were labeled by annotator 1. (From now on, this is called summaries I)
- Each of the 213 summaries has been manually labeled with all character references, the co-reference chains amongst them and relation annotations for pairs of entities that are explicitly mentioned to be in a relation. They have been labeled by annotator 2. (From now on, we call this summaries II)

The applied active learning strategy started by manually selecting about 20 seed training sentences which were manually labeled with information about relations between character references. The seed examples were chosen by matching a wordlist containing indicative expressions (such as “mother”, “father”, “servant” or “loves”) to the text, to enable the classifier to learn relations from different relation types in an unbiased fashion (which usually changes during training because the underlying distribution of relations is heavily biased towards family relations). On those seed examples we trained a binary Maximum Entropy classifier which was applied to thousands of unlabeled sentences. The sentences were then ranked by uncertainty of the classifier. Uncertainty for a sentence, in our case, was defined by extracting all pairs of character references first, applying the classifier to every pair and then assigning the minimum probability to the sentence. The classifier was retrained on command of the user and the ranking of the unlabeled sentences restarted. By applying this strategy, we observed that the average certainty of a sentence rises with every iteration and decided to stop the manual labeling once there was no sentence with a classifier probability below 60% for the novels and 70% for the summaries (this does not mean we reached saturation in classification gain).

For the labeling, we used a total of 57 hierarchically ordered relation labels, inspired by (Massey et al., 2015) (see figure 1). All these labels relate person-entities with each other, such as “motherOf” or “loves”.

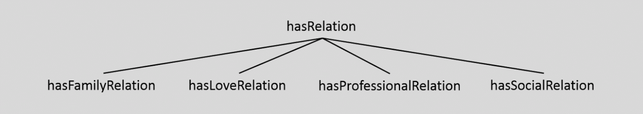


Figure 1: The four main relation types which are further differentiated in 57 relation types in total

The inter-annotator-agreement (IAA) between summary data set I and summary data set II was measured in two ways:

1. A true positive appears when both annotators mark the correct span of the annotation as well as the correct label and the correct arc direction where a correct arc links the two entities in the direction as it is expressed in the text (labeled inter-annotator agreement).
2. A true positive appears when both annotators mark the correct span and arc direction of the relation (unlabeled inter-annotator agreement).

Table 1 gives an overview of the IAA results.

Method	**Precision**	**Recall**	**F1**
Unlabeled IAA	75.6%	43.7%	55.4%
Labeled IAA	60.9%	35.2%	44.6%

Table 1: IAA results, the comparison assumed summaries II as gold and compared summaries I to it.

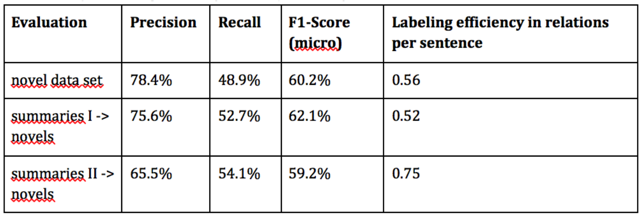
Additionally, we determined 55.5% as the normalized Cohen’s Kappa between our annotators. The results for the IAA are surprisingly low (amount of labeled relations in summaries I compared to the relations in summaries II). The reasons are yet unclear and have to be investigated; we assume that one of them is the high variance of possibilities to express social relations. Labeling the complete summaries may also be more difficult because the annotator needs to read the text completely and might use background knowledge to annotate relations which are only implicit in the text.

Method and evaluation

To compare the transfer from summaries to novels, we trained a classifier, specifically a maximum entropy classifier based on boolean features generated from rule templates because previous work has shown that this classifier is superior in classification accuracy compared to kernel machines, pure rule based approaches or other supervised classifiers such as support vector machines (Krug et al., 2017). Training was done on a data set using the annotations as features and the classifier was applied either to test data from the same set or to a different data set resulting in three evaluations:

- A 5-fold cross evaluation within the novel data set.
- Training on the snippets of the summaries (summaries I or summaries II) and evaluation on the novel data set.

Table 2 shows the result of this experiment for the in-data and cross data evaluation of the relation detection component.



Evaluation	Precision	Recall	F1-Score (micro)	Labeling efficiency in relations per sentence
novel data set	78.4%	48.9%	60.2%	0.56
summaries I -> novels	75.6%	52.7%	62.1%	0.52
summaries II -> novels	65.5%	54.1%	59.2%	0.75

Table 2: The results of a 5-fold in-data set evaluation for both of the data sets and the results for a cross-data set evaluation. Each number represents a micro-average score, i.e. we count every true-positive, false-positive and false-negative in a document and calculate the average scores based on these quantities. We choose the micro score since the label set is rather unbalanced between classes. The efficiency of an approach is measured by calculating number of relations / number of sentences.

Results that are very similar to working directly on the novel (60.2% F1) are achieved by using the model trained on the extracted sentences from the summaries to retrieve information about character relations in the novels (62.1% resp. 59.2%). Since our test data is generated by active learning and only the most difficult examples were chosen for labeling, we expect our results to be a lower bound compared to data in complete novels.

If we use a model trained on the complete summaries, we experience a drop in precision. This drop was to be expected, since the amount of additional labeled relations in the novels is high according to the IAA results (this manifests in the low recall in table 1) as well as can be seen in the labeling efficiency. Altogether, the quality and efficiency of using a classifier trained on summaries are comparable to training on the novels directly based on our data.

Summary

We presented an approach to increase labeling efficiency for relation detection in German novels by transferring knowledge from summaries to novels. It could be shown that using the summaries as trainings data will achieve similar results to using the novels, but the summaries are much shorter and relevant sentences are much more frequent. The inter-annotator agreement for this task is also relatively low which may point to an explanation for the comparatively low results of the automatic approach.

Bibliography

Finn, A., and Kushmerick, N. (2003). “Active learning selection strategies for information extraction.” *Proceedings of the International Workshop on Adaptive Text Extraction and Mining (ATEM-03).*

Jing, H., Kambhatla, N. and Roukos, S. (2007). “Extracting social networks and biographical facts from conversational speech transcripts.” *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*, Prague: ACL, pp. 1040-48.

Krug, M., Wick, C., Jannidis, F., Reger, I. Weimer, L., Madarász, N. and Puppe, F. (2017). “Comparison of Methods for Automatic Relation Extraction in German Novels.” 4. *Tagung Digital Humanities im deutschsprachigen Raum.* Bern: DHd, pp. 223-26.

Massey, P., Xia, P., Bamman, D. and Smith, N. A. (2015). “Annotating character relationships in literary texts.” arXiv preprint: arXiv:1512.00728.

Mintz, M., Bills, S., Snow, R. and Jurafsky, D. (2009). “Distant supervision for relation extraction without labeled data.” *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*. Singapore: ACL, pp. 1003-11.

Surdeanu, M., Tibshirani, J., Nallapati, R. and Manning, C. D. (2012). “Multi-instance multi-label learning for relation extraction.” *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning.* Jeju Island, Korea: ACL, pp. 455-65.