Exposé

Negative Seeds and Negative Rules for Boosting the Precision of Relation Extraction within the DARE Framework

Sebastian Krause[1]
Supervisor : Dr. Feiyu Xu[2]
Supervisor: Prof. Dr. Ulf Leser[1]

February 2010

[1] Institut für Informatik, Humboldt-Universität zu Berlin
[2] DFKI GmbH, Berlin

1 Motivation and Background

The work presented here is an extension of an existing machine learning framework called DARE[3], a system for learning rules that can be used for extracting instances of relations with different complexity from natural language texts ([XuUsLi07] and [Xu07]).

In this paper the term "relation", in contrast to the intuitive sense of the word, refers to a set of tuples with a certain arity. These tuples represent facts or events about real-world objects and concepts. An example is the 4-ary relation whose tuples express the award winning event of nobel prize laureates. The following tuple is an instance of this relation:

(Barack Obama , 2009 , Nobel Prize , Peace)
LAUREATE'S_NAME YEAR_OF_AWARDING PRIZE_NAME PRIZE_AREA

LAUREATE'S_NAME, YEAR_OF_AWARDING, PRIZE_NAME *and* PRIZE_AREA *will be later referred to as the arguments of a relation instance.*

Instances of such relations can be found in a variety of natural language texts, for example, mentionings of nobel prize awards can be found in daily newspapers. The process of recognizing those relation instances in a text is referred to as "relation extraction".

The DARE-System has been developed to learn mapping rules between the linguistic structures expressed in natural language texts and target relations. Rules are acquired in a so-called bootstrapping process:

[3] "Domain Adaptive Relation Extraction", see `http://dare.dfki.de` for a demo implementation

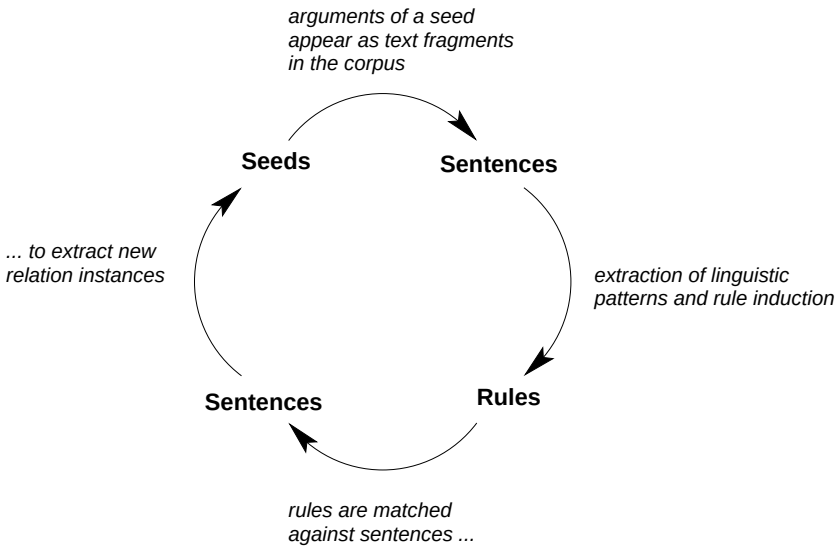

Figure 1. Bootstrapping process

It is initialized by a small set of relation instances ("semantic seeds") from the desired relation. These relation instances are used by the DARE-System to extract rules from a text corpus, which then get applied to the corpus again to extract yet unknown relation instances. The new relation instances are again used as seeds to learn further rules. This process continues until no more relation instances and rules can be found. The results of a run of the DARE-System are therefore two things:

- Instances of the initial seeds' relation that are mentioned in the corpus
- Rules that map the linguistic structure of a sentence to the initial seeds' relation

The goal of this work is to develop strategies to improve the precision of the learned rules and to avoid the extraction of wrong instances. Since the basic DARE-System uses only positive seeds to start the bootstrapping process, we focus on examining the use of negative seeds to derive negative rules. In contrast to [UsXuLi09], where negative seeds are successfully used to initialize a "negative bootstrapping" process, we set our first task to investigate how the truth value of an extracted relation instance can be evaluated *during the online rule learning process*. The second task is then to integrate the evaluation information into the rule ranking process.

The remainder of this exposé is organized as follows: section 2 explains why using a bootstrapping process with only positive seeds is problematic, section 3 describes our approach to boosting the precision of DARE and section 4 states concrete steps for achieving the named goals.

2 Problem

Across different domains the DARE-System has achieved a maximum precision up to 80%. According to [Xu07] and [UsXuLi09] reasons of incorrect extracted relation instances, apart from erroneous linguistic analysis and false information in the text corpus, are learned rules that do not express the desired relation. Such wrong rules occur because of intersections of the desired relation with other relations, that are also mentioned in the corpus.

For example, all nobel prize laureates were nominated before the award, just like many other non-awarded persons, therefore the *"nobel prize nomination event"*-relation is a superset of the *"nobel prize award event"*-relation. The bootstrapping process leads to the learning of dangerous rules for the *"nobel prize nomination event"*-relation which cover both correct and wrong instances, i. e. instances that are not necessarily part of the *"nobel prize award event"*-relation.

Negation of statements and the occurrence of modality, for example the expression of wishes or opinions, are other causes for the extraction of wrong relation instances.

3 Approach

The present rule ranking mechanism of DARE is based on a few parameters only. The first one is the *frequency* with which a rule fires, i. e. the number of sentences a rule matches. The second parameter is the *number of iterations* since the bootstrapping process started. We consider to take more parameters into account, in order to detect negative relation instances and dangerous rules. These parameters are: *trustworthiness*, *specificity* and *distinctiveness*.

The *trustworthiness* of rules can be determined by validating extracted relation instances against given database knowledge. The basic idea is to put more information into the extraction process to achieve more precise results, while the hope is that even adding a small amount of data might already be sufficient to boost the precision in a satisfying way. In other words the hope is that dangerous rules, i. e rules with a low trustworthiness, extract relation instances whose incorrectness can be inferred by using the given database and thus giving the DARE-System the opportunity to recognize this rule as a dangerous one.

Assume we want to extract instances of the relation with award winning events of nobel prize laureates, which we define as follows:

$$\mathcal{R}_{\text{nobel}} \subseteq \textit{NE-Persons} \times \textit{NE-Years} \times \textit{NE-Prizes} \times \textit{NE-Prize-Areas}$$

where *NE-Persons, NE-Years, NE-Prizes* and *NE-Prize-Areas* are sets of named entities that can in texts and which are recognized by a named entity extraction component. To determine the correctness of an extracted instance x we now have to construct the database of additional knowledge in such a way that it represents a closed-world within the target relation $\mathcal{R}_{\text{nobel}}$. For example we could choose to use all nobel laureates in the category *medicine*:

$$\mathcal{CW}_{\text{nobel-medicine}} = \{x \in \mathcal{R}_{\text{nobel}} \,|x.\text{PRIZE\_AREA} = \texttt{ne\_medicine}\}$$

Given this database, i. e. all the tuples in $\mathcal{CW}_{\text{nobel-medicine}}$, we can try to validate the extracted instance x by reasoning:

- If x is in our closed-world, we will know it is correct.
- If x is not in our closed-world, but it contains an award event from the category *medicine*, we will know it is wrong.
- If x is not in our closed-world and it contains an award event from some other category (e. g. *chemistry*), we cannot infer anything.

In general, we have to construct a database containing a closed-world *for each* target relation we want to use the DARE-System with. Although this sounds very elaborate, it does make sense because on the one hand this closed-world database can be quite small (it only needs to list all instances from the target relation with a certain value in one of the arguments of the relation instance) and is therefore usually easy to create. On the other hand there is the possibility to use the resulting rules of the bootstrapping process for extracting instances of different, maybe less popular domains (see [XuUsLi08]). This means that creating a closed-world database for a difficult domain can be avoided if there exists a related domain with better data properties.

The new rule ranking mechanism could use some scoring method which takes into account the result of the validation of the extracted relation instance. For example, the score of a rule which was created using relation instances that are mostly correct and partly with unknown correctness may be higher than the score of one which only extracts relation instances whose correctness can not be determined.

Note that because the scoring system needs to be flexible, as not all extracted relation instances can be validated, we have the benefit that small inaccuracies in the closed-world database, e. g. a missing laureate in medicine, will not affect the extraction result dramatically, as long as we have chosen a prominent (i. e. one that is mentionend often in the used corpus) subpart of the relation as our closed-world.

Another aspect of the scoring method should deal with the *specificity* of relation instances, that is the number of relation arguments that are actually filled with a value. The specificity of a rule is defined by the specificity of the relation instances it extracts. Depending on the relation, an underspecified relation instance that lacks certain arguments can still be useful or rather worthless. In the *"nobel prize award event"*-relation a necessary argument is LAUREATE'S_NAME, while YEAR_OF_AWARDING and PRIZE_AREA are not that important.

Imagine for example the following relation instance with full arity:

```
(Barack Obama, 2009, nobel prize, peace)
```

This relation instance is obviously correct and the following one, which is underspecified, is therefore correct, too:

```
(Barack Obama, --- , nobel prize, ---)
```

Nevertheless, this one could appear in a sentence like this:

"Barack Obama appreciated the Nobel Prize."

The bootstrapping process (see Figure 1) would use this sentence to create a rule that would not express the desired relation, as persons that appreciate the nobel prize not necessarily were also awarded with it. Underspecified relation instances should therefore obtain a lower score than relation instances with full arity.

At last, the *distinctiveness* of seeds should be taken into account. A seed that is not part of many relations is called distinctive. For example, a seed containing only Obama and the Nobel Prize is surely a non-distinctive one, as Barack Obama most likely occurs in a lot of sentences together with a mentioning of a nobel prize without being involved in his own award winning event. Rule ranking should therefore adapt to the distinctiveness of seeds to avoid the change-over to a different relation in the bootstrapping process. Determining the distinctiveness of relation instances during the bootstrapping is tricky, adding information about the approximate distinctiveness of relation instances to the database may help.

4 Work plan

At first a scoring method will be developed, which will supersede the present rule ranking. Afterwards the impact of the size of the added database on the increase in precision will be investigated. All experiments will be conducted on a corpus with sentences from the nobel prize award domain. The results will then be verified on a corpus from the domain of personal relations between celebrities.

In my preparation work, appropriate database knowledge for both corpora has already been acquired, as well as a corpus with texts about celebrities. The corpus with sentences about nobel prize laureates is already available for the further research and experiments.

References

[XuUsLi07] Feiyu Xu, Hans Uszkoreit, Hong Li: *A Seed-driven Bottom-up Machine Learning Framework for Extracting Relations of Various Complexity.* Proceedings of ACL 2007, 45th Annual Meeting of the Association for Computational Linguistics (2007)

[Xu07] Feiyu Xu: *Bootstrapping Relation Extraction from Semantic Seeds.* PHD-Thesis, Saarland University (2007)

[UsXuLi09] Hans Uszkoreit, Feiyu Xu, Hong Li: *Analysis and Improvement of Minimally Supervised Machine Learning for Relation Extraction.* 14th International Conference on Applications of Natural Language to Information Systems. NLDB-09, Springer (2009)

[XuUsLi08] Feiyu Xu, Hans Uszkoreit, Hong Li, Niko Felger: *Adaptation of Relation Extraction Rules to New Domains.* Proceedings of the Poster Session of the Sixth International Conference on Language Resources and Evaluation, LREC (2008)