Adaptation of Relation Extraction Rules to New Domains

Feiyu Xu, Hans Uszkoreit, Hong Li

LT Lab

DFKI GmbH

Stuhlsatzenhausweg 3

66123 Saarbrücken, Germany

{feiyu, uszkoreit, lihong}@dfki.de

Niko Felger

University of Cambridge, Computer Laboratory

15 JJ Thomson Avenue, Cambridge CB3 0FD, UK

niko.felger@googlemail.com

Abstract

This paper presents various strategies for improving the extraction performance of less prominent relations with the help of the rules learned for similar relations, for which large volumes of data are available that exhibit suitable data properties. The rules are learned via a minimally supervised machine learning system for relation extraction called DARE. Starting from semantic seeds, DARE extracts linguistic grammar rules associated with semantic roles from parsed news texts. The performance analysis with respect to different experiment domains shows that the data property plays an important role for DARE. Especially the redundancy of the data and the connectivity of instances and pattern rules have a strong influence on recall. However, most real-world data sets do not possess the desirable small-world property. Therefore, we propose three scenarios to overcome the data property problem of some domains by exploiting a similar domain with better data properties. The first two strategies stay with the same corpus but try to extract new similar relations with learned rules. The third strategy adapts the learned rules to a new corpus. All three strategies show that frequently mentioned relations can help in the detection of less frequent relations.

1. Motivation

The goal of Information Extraction (IE) is to detect relevant pieces of information in free texts and to recognize relations among them (Appelt and Israel, 1999). In recent years, many interesting approaches have been developed for automatically learning relation extraction grammars suited for mapping general linguistic analyses to domain-specific semantic relations (e.g., Riloff (1996), Muslea (1999), Agichtein and Gravano (2000), Yangarber (2001), Greenwood and Stevenson (2006), Stevenson and Greenwood (2006), Xu et al. (2006), Xu et al. (2007) and Xu (2007)).

The success and feasibility of these automatic learning approaches are often strongly dependent on the available data sources and their properties. Data redundancy is particularily relevant for minimally supervised or unsupervised methods (e.g., Jones (2005), Xu et al. (2007), Uszkoreit (2007) and Xu (2007)). In practice, the distribution of relation types in the available data such as online newspapers or blogs is often connected with the popularity and prominence of the relation types. Nobel Prize awards, for instance, are reported by most newspapers, thus, the Nobel Prize belongs to the most frequently mentioned prizes.

In contrast to the Nobel Prize, there are also many prizes that are only mentioned once or twice in the same corpus of newspapers. In this paper, we will present an approach to the adaptation of automatically learned grammar extraction rules from domains with benevolent data properties to related but less benevolent domains exhibiting sparse data and less redundancy.

The paper is organized as follows: section 2 gives an overview of the DARE system, which implements the minimally supervised learning method for relation extraction grammars from some semantic seed. Section 3 discusses the role of data property for the boostrapping based learning and extraction framework. Section 4 describes various strategies for improving the performance and overcoming the data limitations. Section 5 reports on the experimental results. Section 6 concludes the paper with a summary.

2. DARE System

In Xu et al. (2007) and Xu (2007) we present a system called DARE, which implements a minimally supervised method for automatically learning relation extraction grammars. Starting from sample relation instances as seeds, the DARE IE system acquires extraction patterns and systematically induces rules of varied complexity. The system works on a collection of free natural language texts without any annotation of domain information. The only provided domain knowledge for the entire process is the seed. The learning algorithm employs a bootstrapping scheme. At each iteration, rules will be learned based on the seed and then relation instances will be extracted after applying the automatically acquired rules. The newly detected relation instances are then applied as the seed for the next iteration of learning. The termination condition is that no new rules and relations can be acquired anymore.

The complexity of the seed determines the complexity of extracted relations. The seed is utilized for identifying the explicit linguistic expressions containing mentions of relation instances or instances of their k-ary projections where $1<k<n$. Because the seed samples are not linguistic patterns, the learning system is not restricted to

a particular linguistic representation and therefore suitable for various linguistic analysis methods and representation formats.

The DARE pattern discovery is bottom-up and compositional, i.e., complex patterns can build on top of simple patterns for projections. The DARE rule representation model supports this strategy.

Xu (2007) presents examples of the DARE rule representation in the prize award domain. The example relation in the domain contains four arguments representing an event that a person or an organization won a particular prize in a specific area in a certain year, see (1):

(1)

<recipient, award, area, year>

(2)

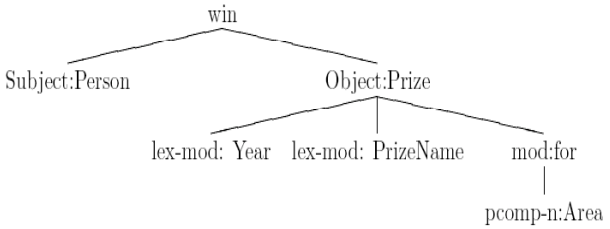
<Mohamed ElBaradei, Nobel, Peace, 2005>

(3)

Mohamed ElBaradei won the 2005 Nobel Prize for Peace on Friday for his efforts to limit the spread of atomic weapons.

(2) is an example relation instance of (1), referring to an event mentioned in the sentence (3). DARE learns three rules from the dependency tree (4): (5), (6) and (7).

(4)



(5) extracts the semantic argument area from a prepositional phrase, while (6) extracts three arguments *year*, *prize* and *area* from the complex noun phrase and calls the rule (5) for the argument *area*.

(5)

Rule name:: area_1

Rule body:: [head [pos preposition; lex-form "for"]; daughters ⟨ [pcomp-n [head 1 Area]] ⟩]

Output:: ⟨1 *Area*⟩

(6)

Rule name:: year_prize_area_1

Rule body:: [head [pos noun; lex-form "prize"]; daughters ⟨ [lex-mod [head 1 Year]], [lex-mod [head 2 Prize]], [mod [head [pos preposition; lex-form "for"]]; rule area_1:: ⟨3 *Area*⟩] ⟩]

Output:: ⟨1 *Year*, 2 *Prize*, 3 *Area*⟩

(7) is the rule extracting all four arguments from the verb phrase dominated by the verb "win" and makes use of (6) to handle the arguments embedded in the linguistic argument "object".

(7)

Rule name:: recipient_prize_area_year_1

Rule body:: [head [pos verb; mode active; lex-form "win"]; daughters ⟨ [subject [head 1 Person]], [object [head [pos noun; lex-form "prize"]]; rule year prize area 1:: ⟨4 *Year*, 2 *Prize*, 3 *Area*⟩] ⟩]

Output:: ⟨1 *Recipient*, 2 *Prize*, 3 *Area*, 4 *Year*⟩

3. Data Properties and Performance

The DARE system has been applied to two domains, Nobel Prize awards and management succession events (Xu, 2007). The Nobel Prize award corpus contains data from various newspapers, thus possessing more redundant data than the MUC-6 data for the management succession that only contains reports from one source, i.e., the New York Times.

If we look more closely at the seed-driven bootstrapping of pattern learning and instance extraction, the entire process can be represented in a bipartite graph where the nodes are either instances or patterns, and connectivity between instances and patterns is detected by systems such as DARE. To achieve good performance, the DARE system is expected to find seed instances leading to many patterns and patterns leading to many instances thus serving as hubs in the learning process, following the duality principle (Brin (1998), Agichtein and Gravano (2000), and Yangarber (2001)). This means that the hub instances or hub patterns build relevant nodes in

the graph. The interplay between the seeds and rules exhibits a Zipf distribution (Zipf, 1932). This implies that a few rules extract most instances.

The evaluation of the Nobel Prize and management succession corpora (Xu, 2007) shows the influence of the data properties on the system performance. In comparison with the management succession domain, the skewed degree distribution can be shown for both patterns and instances in the Nobel Prize domain. Therefore most nodes in the graph can be reached in a few steps. Thus, even with a single instance as seed, the DARE system performs well in this domain. The performance of the DARE system for the Nobel Prize domain is very promising. The connectivity behavior in the management succession is completely different from the Nobel Prize domain. Its patterns and instances have a very low degree of connectivity. Thus, we need more instances as seed to discover enough patterns. The overall performance is comparably poor.

It is clear that the distribution of mentions to events in the Nobel Prize domain will more likely show a Zipf distribution than the data in the management succession domain. Therefore, DARE performs better in he former domain.

4. Adaptation Experiments

As discussed above, data properties are the crucial factor for DARE performance. The management succession domain has a relatively low recall suffering from poor redundancy: nearly all events are just mentioned once, since the data is from a single newspaper, namely, the New York Times. In Xu and Uszkoreit (2007) and Uszkoreit (2007), several strategies have been proposed to circumvent the lack of the required data property. A general and direct approach is to utilize the web to increase redundancy, as also independently proposed by Blohm and Cimiano (2007).

Another strategy is to enlarge the domain or utilize some prominent sibling domains as carrier domains. This requires the modelling of relevant ontological relationships between different domains. For example, the lesser known *Albert Lasker Award* domain belongs to the Prize award domain, having the Nobel Prize award as its prominent sibling domain. In this paper, we present the application of this strategy.

The Nobel Prize is one of the most prominent prizes with extensive media coverage leading to the desired high degree of redundancy in mentions. Patterns learned for the Nobel Prize should be generic enough to extract relations for other prizes and awards too. Indeed, these patterns turn out to be especially helpful to detect less prominent and less mentioned prizes and awards.
Three experiments are designed to determine whether the learned patterns are applicable for the extraction of additional prize winning events and similar relations.
In the first experiment, we apply the patterns to the original Nobel Prize corpus in order to acquire other award events.

In the second experiment, we remove the entity restriction of the "prize name" in the corresponding pattern slots and allow the prize name slot to be filled with any noun phrases, even if they are not recognized as prize or award names. The motivation is to detect prizes and awards that are not discovered by the entity recognition system.

In the third experiment, we apply the learned patterns to a domain corpus on music and musicians with the aim of extracting music award events and to learn new pattern rules with the DARE system.

5. Experimental Results

In the first scenario, a list of Prize winning events has been extracted. The most frequently detected prize is the *Pulitzer Prize*. We have detected 97 *Pulitzer Prize* winning event instances. Among them 95 are correct. Similar to the Nobel Prize, the prize winners obtain the Pulitzer Prize in some special area in literature, e.g., poetry. The precision of the Pulitzer Prize detection is 97%. We also find award events for the following prizes that are recognized by the NE system:

albert lasker award
pritzker prize
turner prize
prix_de_rome

The event instances of the these prizes are mentioned very rarely in the corpus. Only one to three instances for each prize were found.

In the second scenario (which we call *fuzzy extraction*), we find more awards, even less well-known ones, and also other events of winning, e.g., of winning money or praise, as shown in Table 1. Here, the precision of our extraction task is 73%.

Prize and Award	Other
Academy Award	$ 1 million
Cannes Film Festival's Best Actor award	about $ 226,000
American Library Association Caldecott Award	acclaim
American Society	discovery
Blitzker	doctorate
Emmy	election
feature photography award	game
the first Caldecott Medal	master's degree
Francesca Primus Prize	presidency reelection
gold (gold metal)	scholarship
National Book Award	.
Oscar	.
P.G.A	.
PEN/Faulkner Award	.
prize	
reporting (the investigative reporting award)	
Tony (Tony Award)	
U.S. Open	

Table 1: fuzzy extraction

In the third scenario, we are concerned with entertainment awards and musicians that have won these. As opposed to the Nobel domain, this domain exhibits an additional degree of freedom, considering a whole range of different awards, each providing its own set of prize areas. Each of these prizes can be more or less well-known, as can be the artists.

To create an appropriate domain corpus for this scenario, we first conduct a survey of web sites in order to find useful text sources for the relevant relations in the music awards domain. We select the top 100 (well-known) and the bottom 100 (less well-known) musicians from a music database. We combine the musician names (NAME) with some relevant keywords into queries such as "NAME news", "NAME music news", "NAME award", "NAME prize", and "NAME winner". It turns out that the top musicians are most frequently distributed in some general information websites such as Wikipedia. A larger proportion of the lesser-known musicians are mentioned on blogging and fan websites such as blogspot.com or myspace.com. This means that the syntactic constructions potentially used to encode relations concerning popular artists differ from those used to encode relations concerning less popular ones. This result can be exploited when he task consists of the detection of rising stars.

Initial extraction experiments (Felger, 2007) have been carried out on a corpus containing 8.920 documents (324.479 sentences) retrieved from the BBC news website[1]. Due to the relative difficulty of the task compared to the Nobel domain – indicated below – we concentrated our initial experiments on a set of three very popular artists and twelve awards, shown in Table 2. The search queries for retrieving the documents that constitute the corpus consisted of various combinations of artist and award names, as well as the names of prize area for the respective awards.

Eminem	55
Madonna	56
U2	57

Academy Awards (Oscars)	3
American Music Awards	11
Billboard Awards	12
Brit Awards	14
Golden Globe Awards	7
Golden Raspberry Awards	14
Grammy Awards	36
Meteor Ireland Music Awards	17
MTV Video Music Awards	32
MTV European Music Awards	15
MTV Movie Awards	4
The Source Awards	3

1983	1	1996	2
1986	1	1997	6
1987	4	1998	9
1988	4	1999	6
1989	8	2000	17
1990	4	2001	21
1991	2	2003	26
1992	4	2004	4
1993	4	2005	5
1994	2	2006	10
1995	4	2007	1

Table 2: Sample artists and awards

Structurally, the music awards domain is similar to the Nobel domain, since there also is a *prize* that is awarded to a *winner* in a specific *area* in a certain *year*. However, several factors make this task harder.

Firstly, the reporting style in our corpus is very different for these awards. Typically, a single document reports the simultaneous awarding of several awards in different areas to a number of winners in one award ceremony, or it names a number of awards that a specific artist has won. It is rare that a news article reports the awarding of a single prize in a single area, as it is common with the Nobel Prize. Consequently, the data is less redundant. This may be due to the fact that entertainment awards usually have more areas than other prizes, and that it is not uncommon for an artist to be awarded prizes in several areas at the same awarding ceremony.

[1] news.bbc.co.uk

Secondly, while Nobel Prize winners are usually persons – a type of named entity for which recognizers perform reasonably well – the linguistic entities referring to music award winners may be true person names, stage names, or band names, which are often much harder to recognize (e.g. band names such as *Get cape. Wear cape. Fly.* or *We Are Scientists*). Little research has been conducted so far on detecting such types of names. This creates several corollary problems, e.g. reduced performance of Named Entity Recognition, as well as a loss of the ability to generalize from parse tree patterns, when the parser is confused by the internal structure of the names.

Thirdly, artist, prize and prize area names exhibit a higher degree of synonymy variation than other prize or person names; for example, *Eminem*, *Marshall Mathers*, and *Slim Shady* all refer to the same artist, and *Oscar* and *Academy Award* refer to the same award. We dealt with this issue by manually specifying a list of surface variants for each of the chosen awards, artists, and prize areas.

Among the 324.479 sentences in our corpus, 14.076 contain a mention of one of the artists or awards we consider, according to our Named Entity Recognition system, while 1993 sentences contain mentions of two or more entity types suitable to fill different slots of the award relation. 152 sentences contain mentions of three or more such types and 6 contain mentions of all four types. Obviously, entity mentions alone are not sufficient for assuming an instance of the relation or its respective *n*-ary projection. We find 933 sentences that mention two entities that form a subset of the entities in an instance of our Ideal Table, 43 sentences to mention three such entities, and 4 sentences to mention all entities of an instance in the Ideal Table. These figures, however, only provide a rather optimistic estimate of the number of instances actually contained in the corpus. The event description may, e.g., be in the scope of a negation.

We compare the performance of two system configurations. In one configuration the DARE system is applied to the music awards domain in the way it has been applied to the Nobel domain, using the tuple *<Eminem, 2001, Grammy Award, Best Rap Album>* as the seed instance. In the second configuration, the extraction grammar obtained on the Nobel domain is additionally used as a seed of rules. This grammar consists of 717 rules, of which 131 are 4-ary, i.e. fill four slots of the relation template, 452 are 3-ary, and 134 are 2-ary.

	Instances	Precision	Extracted rules 2-/3-/4-ary
without Nobel rules	19	78.95%	108/7/–
with Nobel rules	38	73.68%	154/3/–

Table 3: Performance of the plain DARE system compared to the adapted system

It is not very surprising, that in this initial evaluation, our

system does not perform as well on the music awards domain as it does on the Nobel domain. Table 3 summarizes the performance results of our experiments. The DARE system extracts 19 relation instances of varying complexity at a precision of 78.95% with learned 115 rules. Using the rules obtained in the Nobel Prize domain in addition, the number of extracted instances rises to 38, at a precision of 73.68%. Of these instances, 26 are discovered by rules learned in the Nobel Prize domain. The number of additional rules learned, 157, is also higher than when using the plain system.

6. Conclusions

Although the DARE rule representation is very expressive and can ideally cover all linguistic constructions that can be utilized as pattern rules, the discrepancy in the system performance between the Nobel Prize award domain and the management succession domain points out that the DARE system, or more generally, the bootstrapping framework, is more suitable for some types of data than for others. Even within the Nobel Prize corpus, the performance improves when the data size increases. All three experiments described in Section 4 and 5 confirm the carrier function of a more fertile sibling domain. The patterns learned by the Nobel Prize domain are generic enough to be applicable to other awards. In particular, a prominent sister domain helps to extract more instances than could be extracted by learning from the actual target domain. More research is needed to investigate the differences of domains and tasks with respect to the DARE learning approach, especially concerning the properties of patterns, redundancies and confusion candidates.

Acknowledgement

The work presented here was partially supported by a research grant from the German Federal Ministry of Education, Science, Research and Technology (BMBF) to the DFKI project Hylap (FKZ: 01 IW F02) and the international project RASCALLI supported by the European Commission Cognitive Systems Programme (IST-27596-2004).

References

Agichtein, E. and Gravano, L. (2000). *SNOWBALL: Extracting relations from large plain-text collections*. In Proceedings of the Fifth ACM International Conference on Digital Libraries.

Appelt, D. and D. Israel (1999). *Introduction to Information Extraction Technology.*

Blohm, S. and P. Cimiano (2007). Using the Web to Reduce Data Sparseness in Pattern-based Information Extraction. *Proceedings of the 11th European Conference on Principles and Practice of Knowledge Discovery in Databases.*

Brin, S. (1998). Extracting patterns and relations from the world wide web. In *WebDB Workshop at 6th International Conference on Extending Database Technology, EDBT'98.*

Felger, N. (2007). Portierung eines Relationsextraktions-systems auf eine neue Domäne. B.Sc. Thesis, Saarland University, Saarbrücken, Germany.

Greenwood, M. A. and M. Stevenson (2006). Improving semi-supervised acquisition of relation extraction patterns. In *Proceedings of the Workshop on Information Extraction Beyond the Document*, Sydney, Australia, pp. 29--35. ACL.

Jones, R. (2005). Learning to Extract Entities from Labeled and Unlabeled Text. Ph. D. thesis, University of Utah.

Muslea, I. (1999). Extraction patterns for information extraction tasks: A survey. In *AAAI Workshop on Machine Learning for Information Extraction*, Orlando, Florida.

Riloff E. (1996). Automatically generating extraction patterns from untagged text. In *Proceedings of Thirteenth National Conference on Artificial Intelligence (AAAI-96)*, pp. 1044--1049. The AAAI Press/MIT Press.

Stevenson, M. and M. A. Greenwood (2006). Comparing information extraction pattern models. In *Proceedings of the Workshop on Information Extraction Beyond The Document*, Sydney, Australia, pp. 12--19. ACL.

Uszkoreit, H. (2007). Invited talk: Methods and Applications for Relation Detection: Potential and Limitations of Automatic Learning in IE. *2007 IEEE International Conference on Natural Language Processing and Knowledge Engineering (IEEE NLP-KE 2007)*.

Xu, F. and H. Uszkoreit (2007). Minimally supervised learning of relation extraction rules using semantic seeds. A seminar talk at the National Center for Text Mining (NaCTeM).

Xu, F., H. Uszkoreit, and H. Li (2006). Automatic event and relation detection with seeds of varying complexity. In *Proceedings of AAAI 2006 Workshop Event Extraction and Synthesis*, Boston.

Xu, F., H. Uszkoreit, and H. Li (2007). A seed-driven bottom-up machine learning framework for extracting relations of various complexity. *Proceedings of ACL07*, 584--591.

Xu, F. (2007). Bootstrapping Relation Extraction from Semantic Seeds. PhD thesis, Saarland University, Saarbrücken, Germany, 2007.

Yangarber, R. (2001). Scenarion Customization for Information Extraction. Dissertation, Department of Computer Science, Graduate School of Arts and Science, New York University, New York, USA.

Zipf, G.K. (1932). Selected Studies of the Principle of Relative Frequency in Language. Cambridge, MA.: Harvard University Press.