META-DARE:
Monitoring the Minimally Supervised ML of Relation Extraction Rules

Hong Li **Feiyu Xu** **Hans Uszkoreit**

German Research Center for Artificial Intelligence (DFKI), LT-Lab
Alt-Moabit 91c, 10559 Berlin, Germany
{lihong,feiyu,uszkoreit}@dfki.de
http://www.dfki.de/lt/

Abstract

This paper demonstrates a web-based online system, called META-DARE[1]. META-DARE is built to assist researchers to obtain insights into seed-based minimally supervised machine learning for relation extraction. META-DARE allows researchers and students to conduct experiments with an existing machine learning system called DARE (Xu et al., 2007). Users can run their own learning experiments by constructing initial seed examples and can monitor the learning process in a very detailed way, namely, via interacting with each node in the learning graph and viewing its content. Furthermore, users can study the learned relation extraction rules and their applications. META-DARE is also an analysis tool which gives an overview of the whole learning process: the number of iterations, the input and output behaviors of each iteration, and the general performance of the extracted instances and their distributions. Moreover, META-DARE provides a very convenient user interface for visualization of the learning graph, the learned rules and the system performance profile.

1 Introduction

Seed-based minimally supervised machine learning within a bootstrapping framework has been widely applied to various information extraction tasks (e.g., (Hearst, 1992; Riloff, 1996; Brin, 1998; Agichtein and Gravano, 2000; Sudo et al., 2003; Greenwood and Stevenson, 2006; Blohm and Cimiano, 2007)). The power of this approach is that it needs only a small set of examples of either patterns or relation instances and can learn and discover many useful extraction rules and relation instances from unannotated texts. Within this framework, Xu et al. (2007) develop a learning approach, called DARE, which learns relation extraction rules for dealing with relations of various complexity by utilizing some relation examples as semantic seed in the initialization and has achieved very promising results for the extraction of complex relations. In the recent years, more and more researchers are interested in understanding the underlying process behind this approach and attempt to identify relevant learning parameters to improve the system performance.

Xu (2007) investigates the role of the seed selection in connection with the data properties in a careful way with our DARE system. Xu (2007) and Li et al. (2011) describe the applications of DARE system in different domains for different relation extraction tyes, for example, the Nobel-Prize-Winning event, management succession relations defined in MUC-6, marriage relationship, etc. Uszkoreit et al. (2009) describe a further empirical analysis of the seed construction and its influence on the learning performance and show that size, arity and distinctiveness of the seed examples play various important roles for the learning performance. Thus, the system demonstrated here, called META-DARE, serves as a monitoring and analysis system for conducting various experiments with seed-based minimally supervised machine learning. META-DARE is also aimed to assist researchers to understand the DARE algorithm and its rule representation and the interaction between rule learning and relation instance extraction. It allows users to construct different seed sets with respect to size, arity and specificity to start experiments on the example domains. Moreover, it provides a detailed survey of all learning iterations including the learned rules and extracted instances and their respective properties. Finally, it delivers a qualitative analysis of the learning per-

[1] http://dare.dfki.de/

Proceedings of Recent Advances in Natural Language Processing, pages 378–384,
Hissar, Bulgaria, 12-14 September 2011.

formance.

As a web service, it offers a very user-friendly visualization of the learning graph and allows users to interact with the learning graph and study the interaction between learning rules and extracted relation instances. Each rule and extracted instance is presented in a feature structure format. Furthermore, the wrong instances extracted by DARE are visually extra marked so that users can investigate them and learn lessons from them. As a side effect, META-DARE is a very useful and effective tool for teaching information extraction.

The paper is organized as follows: Section 2 outlines the overall architecture, while Section 3 explains the experiment corpus. Section 4 describes the DARE system and the learning algorithm. In Section 5, we introduce the seed selector. Section 6 reports the visualization functions of META-DARE. Section 7 gives a conclusion and discusses future ideas.

2 META-DARE: Overall Architecture

Figure 1 depicts the overall architecture of the META-DARE system.

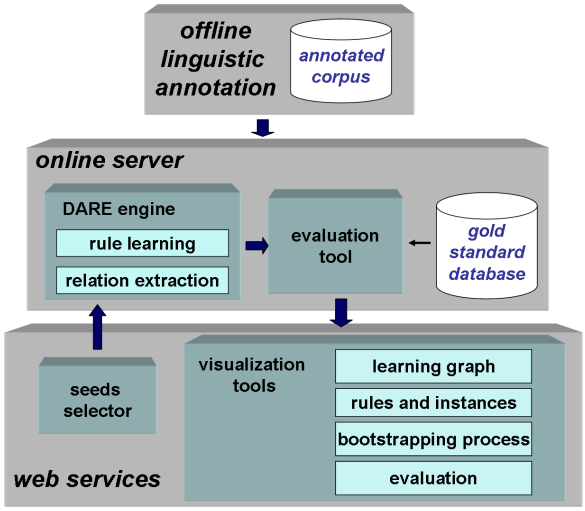

Figure 1: META-DARE: Overall architecture

META-DARE contains three major parts:

- **Online server**: This module is responsible for learning, extracting and evaluation. Its core component is the *DARE engine* for *rule learning* and *relation extraction*. The *evaluation tool* is responsible for validation of the extracted instances against our gold standard databases.
- **Offline linguistic annotation**: This component automatically annotates the corpus texts with named entity information and dependency tree structures using standard NLP tools. All annotations are stored in XML format.
- **Web services**: This part is responsible for user interaction and visualization of learning, extraction and evaluation results. The component *Seeds Selector* allows users to choose their own initial seed set for their experiments. The *visualization tools* present the learning graph and allow users to view learned rules, extracted instances and their interactions. Furthermore, evaluation results of the extracted instances are presented in tabular form.

3 Experiment Corpus

In META-DARE, we use the standard Nobel-Prize corpus described in (Xu et al., 2007), which contains mentionings of the Nobel Prize award events. The target relation for our experiment domain is a quaternary tuple about a person obtaining Nobel Prize in a certain year and in a certain area, described as follows:

$$\langle Person, Prize, Area, Year \rangle \ .$$

There are 3312 domain related documents (18MB) from online newspapers such as NYT, BBC and CNN. To facilitate our learning, the corpus is preprocessed with several NLP tools (see component "offline linguistic annotation"). We utilize the named entity recognize tool **SProUT** to annotate seven types of named entities: *Person, Location, Organization, Prize, Year, PrizeArea* (Drozdzynski et al., 2004). Furthermore, we apply the dependency parser **MiniPar** for obtaining grammatical functions (Lin, 1998). Users can access the annotations via the system web page where the named entities are highlighted and the dependency structures are presented in a tree format.

4 DARE: Bootstrapping Relation Extraction with Semantic Seed

The core engine in META-DARE is DARE (**D**omain **A**daptive **R**elation **E**xtraction), a minimally supervised machine learning framework for extracting relations of various complexity (Xu et

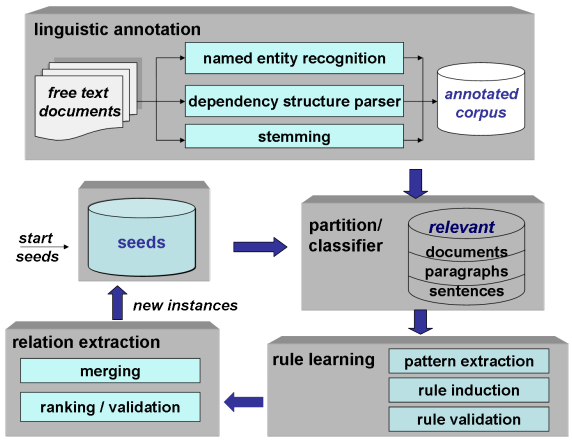

Figure 2: DARE system architecture

al., 2007). Figure 2 illustrates the DARE system architecture.

DARE learns rules from un-annotated free texts, taking some relation instances as examples in the initialization. The learned extraction rules are then applied to the texts for detection of more relation and event instances. The newly discovered relation instances become new seeds for learning more rules. The learning and extraction processes interact with each other and are integrated in a bootstrapping framework. The whole algorithm works as follows:

1. **Input**:
   - A set of un-annotated natural language texts, preprocessed by named entity recognition and dependency parser
   - A trusted set of relation instances, initially chosen ad hoc by the users, as *seeds*.
2. **Partition/Classifier**: Apply seeds to the documents and divide them into relevant and irrelevant documents. A document is relevant if its text fragments contain a minimal number of the relation arguments of a seed and the distance among individual arguments does not exceed the defined width of the textual window.
3. **Rule learning**:
   - **Pattern extraction**: Extract linguistic patterns which contain seed relation arguments as their linguistic arguments and compose the patterns to relation extraction rules.
   - **Rule induction**: Induce relation extraction rules from the set of patterns using compression and generalization methods.
   - **Rule validation**: Rank and validate the rules based on their domain relevance and the trustworthiness of their origin.
4. **Relation extraction**: Apply induced rules to the corpus, in order to extract more relation instances. The extracted instances will be merged and validated.
   - **Merging**: Merge the compatible instances.
   - **Ranking and validation**: Rank and validate the new relation instances.
5. **Stop** if no new rules and relation instances can be found, else repeat step 2 to step 4 with the new seeds resulted from the current step 4.

DARE learns rules basically from the dependency tree structures and proposes a novel compositional rule representation model which supports bottom-up rule composition. A rule for a n-ary relation can be composed of rules for its projections, namely, rules that extract a subset of the n arguments. Furthermore, it defines explicitly the semantic roles of linguistic arguments for the target relation.

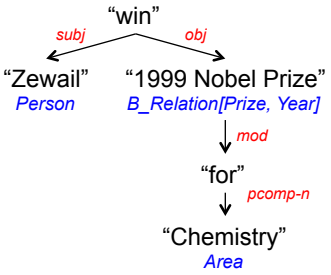

Figure 3: dependency tree example

Let us look at the following example in our experiment domain. Given the following example (1) as our seed which describes a person *Ahmed Zewail* won the *Nobel* Prize in the area of *Chemistry* in the year of *1999*, all four arguments occur in the following sentence (2) in our experiment corpus. The dependency tree structure of sentence (2) is showed in Figure 3.

(1) ⟨*Ahmed Zewail, Nobel, Chemistry, 1999*⟩

(2) *Ahmed Zewail* won the *1999 Nobel* Prize for *Chemistry*.

The rule extracted from example (2) is illustrated in Figure 4, headed by the verb "win". This rule extracts all four arguments for the target relation, where the two arguments *Prize* and *Year* are extracted by its binary projection rule specified as the value of the feature *HEAD* belonging to the grammar function *OBJ* (object). The binary rule detects the *Prize* and *Year* arguments in a complex NP such as "the 1999 Nobel Prize".

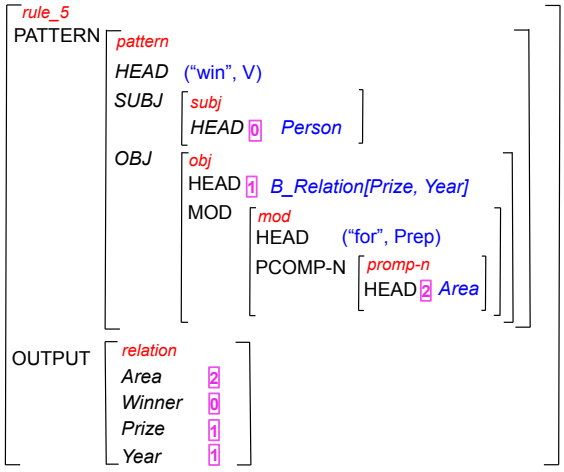

Figure 4: Learned relation extraction rule example

5 Seeds Selector for Seed Construction

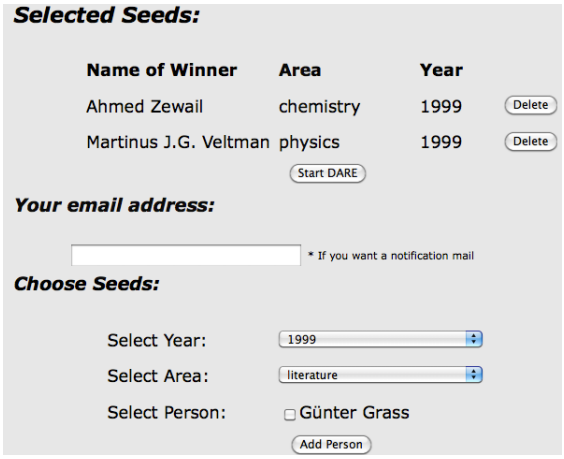

Figure 5: Seed selector

META-DARE offers users a web interface for seed construction[2]. Figure 5 illustrates a seed construction example. Users can choose their seed examples according to the following parameters:

- **Size**: users can select as many winning events as available.
- **Year**: users can choose winners belonging to a certain year.
- **Area**: users can add their preferred area.
- **Person name**: users are allowed to select their preferred person name.

Given a valid email address from the user, the system is able to dispatch a notification automatically when the experiment ends.

6 Visualization for Monitoring

META-DARE allows users to access and monitor the following elements of the bootstrapping process:

- Learning graph: Users have access to the whole learning graph and can also zoom in the graph and interact with each node and view its content.
- Learned rule: Each learned rule is presented as a feature structure and is linked to its seeds and sentences from which it is extracted.
- Evaluation results: The distribution of the extracted instances and their precision is presented in tabular form.

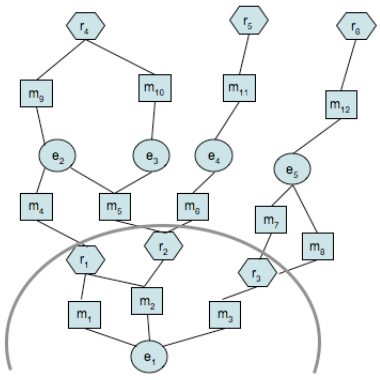

Figure 6: Learning graph starting from semantic seed. e_i: *relation instances*; r_i: *extraction rules*; m_j: *textual snippets*

6.1 Learning Graph

A learning graph in DARE is a graph whose vertices are relation instances, extraction rules and text units as depicted in Figure 6. The learning process starts with instances (e.g., e_1) as seeds and finds textual snippets (e.g., m_1, m_2, m_3) which

[2] http://dare.dfki.de/start_demo.jsp

	4 arity	3-arity			2 arity	sum
		(W. P. A.)	(W. P. Y.)	sum		
correct	142	61	20	81	74	297
sum	155	88	21	109	107	371
precision	91.61%	69.32%	95.24%	74.31%	69.16%	80.05%

Table 1: Distribution of extracted instances and their precision

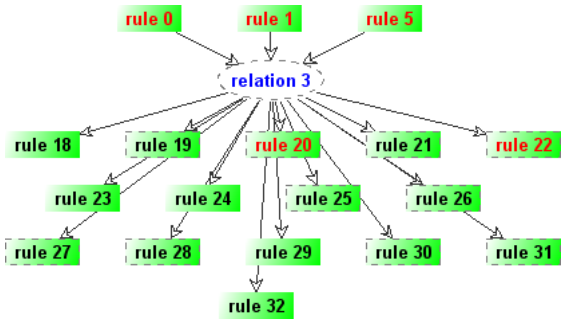

Figure 7: Interaction of rule application and rule learning

match the seeds and then extract pattern rules (e.g, r_1, r_2, r_3). Figure 6 represents the extraction and learning process as a growing graph (Uszkoreit et al., 2009).

The learning graph visualized in META-DARE mainly focuses on the interaction between the learned rules and their seed instances[3]. Figure 7 shows that all three learned rules $rule$ 0, $rule$ 1 and $rule$ 5 detect the same relation instance $relation$ 3 as follows:

(3) ⟨*Robert Mundell, Nobel, Economics, 1999*⟩

which further helps to learn many new rules including $rule$ 18 and $rule$ 19 etc. The nodes not framed by dashed lines, such as $rule$ 23 and $rule$ 24 are rules that cannot discover any new relation instances. The foreground colors of the nodes indicate the evaluation information (see Section 6.2).

If users click one of these rules, they can view the rule presentation as depicted in Figure 4.

The sentences mentioning extraction rules or instances are also presented on the web page. The following example shows two sentences from which $relation$ 3 is extracted.

(4) 1. *Canadian economist Robert Mundell won the Nobel in economics for introducing foreign trade, capital movements, and currency swings into Keynesian economics in the early 1960s. (nyt, 1999-10-13)*

2. *The Canadian-born professor Robert Mundell has won the 1999 Nobel Prize for Economics. (bbc, 1999-10-14)*

[3] http://dare.dfki.de/graph.jsp?f_id=example

6.2 Visualization of Evaluation Results

With the help of the gold standard database about the Nobel prize winners, we are able to automatically evaluate the extracted instances. In our evaluation, we take following aspects into account:

- overall performance of the relation extraction: precision and recall
- detailed analysis of the extracted instances: distribution of relation instances with various arities and their precision.
- highlighting of the wrong instances and indications of error sources

Table 1 lists the extraction results and their evaluations after one experiment run with only one example as seed. This seed is mentioned in example (1). We classify the extracted relation instances into different groups depending on their argument combinations. The overall precision of this experiment is 80.05% with 297 correct instances. The precision of instances with all four arguments given is pretty high, namely, 91.61%. They cover almost half of extracted instances. Among the instances with three arguments, there are two argument combinations where *W* stands for winners, *P* for prize names, *Y* for years and *A* for areas. The combination *(W.P.Y)* has achieved a very good precision but contains only few instances. In our experiment, we consider only instances at least containing a person name as instance candidates. This experiment confirms our observation that instances which cover more arguments of the target relation have in general better precision values.

In Table 2 and Table 3, we summarize four different experiments depending on different seed configurations. Table 2 lists the configuration of

id	instance number		prize area	year
1	1		chemistry	1999
2	1		chemistry	1998
3	2		peace	1998
4	12	3	medicine	1998
		2	chemistry	
		2	peace	
		1	literature	
		3	physics	
		1	economics	

Table 2: Different seed constructions

id	bootstrapping steps	extracted instances		learned rules
		sum	4-arity	
1	7	372	156	1151
2	10	374	156	1146
3	6	373	159	1147
4	5	374	163	1117

Table 3: Performance comparison of different seed constructions mentioned in Table 2

seed construction in the four experiments. The first two experiments apply only one seed example and both seed examples are in the same area *Chemistry*, but in a different year. The seed in the third experiment contains two examples in the area *Peace*, while the fourth contains all twelve winners in the year *1998*. If we compare the number of the learned rules and the learned instances in Table 3, all four experiments do not differ too much from each other. However, with more examples in the fourth run, the system needs only five iterations. As reported in (Uszkoreit et al., 2009), the Nobel corpus owns a data property close to a small world. With one single example, the system can achieve very good performance. Therefore, all four experiments share similar performance in our evaluations.

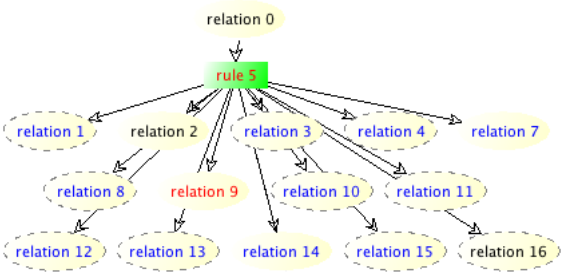

Figure 8: Highlighting of the wrong instances and indications of error sources

As illustrated in Figure 7 and 8, META-DARE also highlights the dangerous or bad rules and wrong relation instance. As described in Xu et al. (2010), the acquired rules are divided into four groups according to the extraction results:

- useless, if the rule does not extract any instances.
- good, if the rule extracts only correct instances.
- dangerous, if the rule extract both correct and wrong instances.
- bad, if the rule extract only bad instances.

In the learning graph, the rules from different group are colored in the following way:

- useless rules: not framed by dashed lines
- good rules: black foreground
- dangerous or bad rules: red foreground

In a similar way, the extracted instances are colored as follows:

- correct instance: blue foreground
- wrong instance: red foreground
- not evaluable: black foreground, such as instance about other prize-winning events but not noble-prize-winning
- useless seed: not framed by dashed lines. With these instances no rules are learned.

For example, in Figure 7 $rule$ 23 and $rule$ 24 are the useless rules, while $rule$ 20 and $rule$ 22 have extracted the wrong instances. $Rule$ 0, $rule$ 1 and $rule$ 5 are the dangerous rules. In Figure 8 $Relation$ 9 is a wrong instance but it does not contribute more errors. $rule$ 5 is a dangerous rule. The users can study the rule and the corresponding sentences from which this rule is learned.

7 Conclusion and Future Work

We demonstrate the META-DARE system which implements the minimally supervised machine learning approach DARE for learning rules and extracting relation instances. META-DARE provides a user-friendly web interface to allow researchers to conduct their own experiments and to

obtain insights in the bootstrapping process such as the learning graphs, the learned rules and the iteration behaviors. Furthermore, the evaluation results and the highlighting of the errors are very useful to investigate the learning algorithms and to develop improvement solutions.

META-DARE is an initial approach to an online monitoring system of seed-based minimally supervised machine learning approaches. We plan to integrate more domains and target relations as described in (Xu, 2007; Li et al., 2011). Since DARE is domain adaptive, the META-DARE can be easily customized if users might provide additional corpora and definitions of new relations for a new domain. It might be also useful if META-DARE can display the ranking information computed by the confidence estimation component (Xu et al., 2010) for the instances and the rules. Furthermore, in addition to seed construction, we would like to allow more interactions with the DARE system in the near future, such as adding or selecting negative examples for learning negative rules (Uszkoreit et al., 2009), evaluating the instances or rules during the bootstrapping or correcting the linguistic annotation of NLP tools. An even ambitious plan is to integrate other similar rule learning systems and compare their performance with each other.

Acknowledgements

This research was conducted in the context of the DFG Cluster of Excellence on Multimodal Computing and Interaction (M2CI), projects Theseus Alexandria and Alexandria for Media (funded by the German Federal Ministry of Economy and Technology, contract 01MQ07016), and project TAKE (funded by the German Federal Ministry of Education and Research, contract 01IW08003).

References

Eugene Agichtein and Luis Gravano. 2000. Snowball: Extracting relations from large plain-text collections. In *Proceedings of the 5th ACM International Conference on Digital Libraries (DL'00)*, San Antonio, TX, June.

S. Blohm 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 (PKDD)*, September.

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

Witold Drozdzynski, Hans-Ulrich Krieger, Jakub Piskorski, Ulrich SchŁfer, and Feiyu Xu. 2004. Shallow processing with unification and typed feature structures — foundations and applications. *Knstliche Intelligenz*, (1):17–23.

Mark A. Greenwood and Mark Stevenson. 2006. Improving semi-supervised acquisition of relation extraction patterns. In *Proceedings of the Workshop on Information Extraction Beyond The Document*, pages 29–35, Sydney, Australia, July. Association for Computational Linguistics.

M.A. Hearst. 1992. Automatic Acquisition of Hyponyms om Large Text Corpora. *Proceedings of the Fourteenth International Conference on Computational Linguistics*.

Hong Li, Feiyu Xu, and Hans Uszkoreit. 2011. Minimally supervised rule learning for the extraction of biographic information from various social domains. In *Proceedings of RANLP 2011*.

D. Lin. 1998. Dependency-based evaluation of MINIPAR. *Workshop on the Evaluation of Parsing Systems*, pages 317–330.

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

K. Sudo, S. Sekine, and R. Grishman. 2003. An improved extraction pattern representation model for automatic IE pattern acquisition. *Proceedings of ACL 2003*, pages 224–231.

Hans Uszkoreit, Feiyu Xu, and Hong Li. 2009. Analysis and improvement of minimally supervised machine learning for relation extraction. In *14th International Conference on Applications of Natural Language to Information Systems*. Springer.

Feiyu Xu, Hans Uszkoreit, and Hong Li. 2007. A seed-driven bottom-up machine learning framework for extracting relations of various complexity. In *Proceedings of ACL 2007, 45th Annual Meeting of the Association for Computational Linguistics*, Prague, Czech Republic, 6.

Feiyu Xu, Hans Uszkoreit, Sebastian Krause, and Hong Li. 2010. Boosting relation extraction with limited closed-world knowledge. In *Proceedings of the 23rd International Conference on Computational Linguistics*. Association for Computational Linguistics, o.A.

Feiyu Xu. 2007. *Bootstrapping Relation Extraction from Semantic Seeds*. Phd-thesis, Saarland University.