Large-Scale Learning of Relation-Extraction Rules with Distant Supervision from the Web

Sebastian Krause, Hong Li, Hans Uszkoreit and Feiyu Xu

Language Technology Lab, DFKI
Alt-Moabit 91c, Berlin, Germany
{sebastian.krause,lihong,uszkoreit,feiyu}@dfki.de

Abstract. We present a large-scale relation extraction (RE) system which learns grammar-based RE rules from the Web by utilizing large numbers of relation instances as seed. Our goal is to obtain rule sets large enough to cover the actual range of linguistic variation, thus tackling the long-tail problem of real-world applications. A variant of distant supervision learns several relations in parallel, enabling a new method of rule filtering. The system detects both binary and n-ary relations. We target 39 relations from Freebase, for which 3M sentences extracted from 20M web pages serve as the basis for learning an average of 40K distinctive rules per relation. Employing an efficient dependency parser, the average run time for each relation is only 19 hours. We compare these rules with ones learned from local corpora of different sizes and demonstrate that the Web is indeed needed for a good coverage of linguistic variation.

Keywords: information extraction, IE, relation extraction, RE, rule based RE, web scale IE, distant supervision, Freebase

1 Introduction

Tim Berners-Lee defines the *Semantic Web* as "a web of data that can be processed directly and indirectly by machines" [4]. Today, there is still a long way to go to reach the goal of a true Semantic Web because most information available on the Web is still encoded in unstructured textual forms, e. g., news articles, encyclopedia like *Wikipedia*, online forums or scientific publications. The research area of *information extraction* (IE) aims to extract structured information from these kinds of unstructured textual data. The extracted information can be instances of concepts such as persons, locations or organizations, or relations among these concepts. *Relation extraction* (RE) deals with the automatic detection of relationships between concepts mentioned in free texts. It can be applied for automatically filling and extending knowledge databases and for semantic annotation of free texts. In recent research, *distant supervision* has become an important technique for data-driven RE (e. g. [15, 16, 22, 32]) because of the availability of large knowledge bases such as *Yago* [21] and *Freebase*[1]. Distant supervision utilizes a large number of known facts of a target domain for automatically labeling mentions of these facts in an unannotated text corpus, hence generating training data.

[1] http://www.freebase.com/

We develop a large-scale RE system that employs Freebase facts as seed knowledge for automatically learning RE rules from the Web in the spirit of distant supervision. The obtained rules can then be applied for the extraction of new instances from new texts. Freebase is a fact database containing some 360 million assertions about 22 million entities such as people, locations, organizations, films, books, etc. We extend the distant supervision approach to RE by combining it with existing means for accommodating relations with arity > 2. To the best of our knowledge, this is the first approach to RE which can learn large-scale grammar-based RE rules for n-ary relations from the Web in an efficient way. We try to learn from the Web as many such rules as possible. For these rules, we adopt the rule formalism of the DARE framework [31], because it accommodates relations of various complexity and is expressive enough to work with different linguistic formalisms, in particular, results of deeper analysis such as dependency structures. When applied to parsed sentences, the learned rules can detect relation mentions, extract the arguments and associate them with their respective roles. Therefore, the results can be directly used as input for a knowledge database. In comparison to statistical-classifier approaches like [15, 16], our approach does not only come up with a web-scale RE system but also delivers the extraction rules as an important knowledge source, which can be reused for question answering, textual entailment and other applications.

Our method is applied to 39 relations from the domains *Awards*, *Business* and *People* modeled in Freebase. About 2.8M instances of these relations were retrieved from Freebase as seed knowledge, from which about 200,000 were turned into Bing queries, resulting in almost 20M downloaded web pages. 3M sentences matched by seed facts were utilized to learn more than 1.5M RE rule candidates. Run time for learning was reduced by parallelization with three server machines (16 cores with 2.4 GHz each; 64 GB RAM). We utilize a very efficient dependency parser called MDParser [24]. In our experiments, it takes around 120 ms to parse one sentence of the average length of 25 words. For each relation, the average run time for the entire rule learning process takes only 19 hours.

Our experiments show that the large number of learned rules make useful candidates of RE rules. These rules produce a higher recall than semi-supervised bootstrapping on a domain-relevant small corpus or distant supervision on a large local corpus. However, precision is hampered by a large number of invalid candidate rules. But many of the invalid rule candidates are learned for multiple relations, even for incompatible ones. Therefore, we use the rule overlap between relations for effective filtering. This technique is a new variant of previously proposed methods, i. e., *counter training* [7, 33] and *coupled learning* [6]. It is better suited for distant supervision learning, since it works directly on the rule sets without needing a confidence feedback of extracted instances.

2 Related Work

Real-world applications often benefit from the extraction of n-ary relations, in particular, in the case of event extraction. Very often more than two arguments of an event are mentioned in a single sentence, e. g., in the following example.

Example 1. *Prince Albert has married the former Olympic swimmer Charlene Wittstock in a Catholic ceremony in Monaco.*

Here, three arguments of a wedding event are mentioned: the two persons (*Prince Albert*, *Charlene Wittstock*) and the location (*Monaco*). In general, the *binary relation only* approaches (e. g, [17, 19, 20]) do not employ the existing syntactic and semantic structures among $n > 2$ arguments and rely on a later component to merge binary relations into relations of higher complexity (e. g., [14]). As described above and explained in Section 4.2, DARE [31] provides a rule extraction strategy, which allows rules to have more than 2 arguments, when they co-occur in one sentence.

Approaches with surface-oriented rule representation (e. g., [11–13, 19]) prefer to employ shallow linguistic analyses thus circumventing less efficient full syntactic parsing for large-scale RE tasks. These formalisms are robust and efficient but only handle binary relations. They work best for relations whose arguments usually co-occur in close proximity within a sentence and whose mentions exhibit limited linguistic variation. In contrast, systems learning RE rules from syntactic structures such as dependency graphs are able to detect relation arguments spread widely across a sentence (e. g., [31, 34]). However, these approaches are usually applied only to relatively small corpora.

The minimally-supervised *bootstrapping* paradigm takes a limited number of initial examples (relation instances or patterns) and labels free texts during several iterations (e. g., [2, 20, 34]). These approaches often suffer from semantic drift and the propagation of errors across iterations [15]. Furthermore, their performance is strongly dependent on the properties of the data, i. e., on specific linguistic variation in conjunction with redundant mention of facts [23]. In contrast, distant supervision approaches [10, 16, 27, 28, 32] rely on a large amount of trustworthy facts and their performance does not hinge on corpus data properties such as redundancy, since multiple occurrences of the same instance in different sentences are not required.

Closely related to our distant supervision approach is the work described by [15], who train a *linear-regression* classifier on examples derived from mentions of *Freebase* relation instances in a large Wikipedia corpus. They focus on the 109 most populated relations of *Freebase*. The trained classifier works on shallow features such as word sequences and POS tags and on dependency relations between words. To our knowledge, neither [15], nor other existing distant supervision approaches can handle n-ary relations.

Parallel to the above approaches, a new paradigm has emerged under the name of *open IE*. A pioneering example is the *TextRunner* system [3, 35]. In contrast to *traditional* RE systems, they do not target fixed relations, thus being very useful for applications continuously faced with new relation or event types, e. g., online social media monitoring. However, the results of these systems cannot be directly taken for filling knowledge databases, because the semantics of the new relations including the roles of the entities remains unknown.

All ML systems for RE are faced with the problem of estimating the confidence of the automatically acquired information. Some approaches utilize the confidence value of the extracted instances or the seed examples as feedback for evaluation of the rules (e. g., [1, 5, 34]). Many others employ negative examples for detecting wrong rules [7, 33], so-called *counter training*. In [30], negative examples are implicitly generated by utilizing a given set of positive relation instances, which form a *closed world*. [6] introduces

coupled learning, which learns a coupled collection of classifiers for various relations by taking their logical relationships as constraints for estimating the correctness of newly extracted facts. Our current rule filtering method works directly on rules without making use of any confidence information associated with extracted instances.

3 Target Relations and Essential Type

We decide to conduct our experiments in three domains: *Award*, *Business* and *People*. All three domains contain n-ary relations with $n = 2$ and $n > 2$.

Let t be a named-entity type and let $\mathcal{NE}_t$ be the set containing *all* named entities of type t. Let T be a bag of named-entity types and let $n = |T|$. Then any of our n-ary target relations is a set $\mathcal{R}$ for some T with

$$\mathcal{R} \subseteq \prod_{t \in T} \mathcal{NE}_t \,. \tag{1}$$

Derived from the modeling in Freebase, the *marriage* relation can formally be described by:

$$\mathcal{R}_{\text{marriage}} \subseteq \mathcal{NE}_{\text{person}} \times \mathcal{NE}_{\text{person}} \times \mathcal{NE}_{\text{location}} \times \mathcal{NE}_{\text{date}} \times \mathcal{NE}_{\text{date}} \,. \tag{2}$$

Table 1. Some of the target relations of the *Award*, *Business* and *People* domains.

	ARGUMENT NAMES & Entity Types				
Relation	**Slot 1**	**Slot 2**	**Slot 3**	**Slot 4**	**Slot 5**
award nomination	AWARD award concept Ⓡ	NOMINEE organization Ⓡ person	DATE date	WORK creative work	
award honor	AWARD award concept Ⓡ	WINNER organization Ⓡ person	DATE date	WORK creative work	
hall of fame induction	HALL OF FAME award concept Ⓡ	INDUCTEE organization Ⓡ person	DATE date	-	
organization relationship	PARENT organization Ⓡ	CHILD organization Ⓡ	FROM date	TO date	-
acquisition	BUYER organization Ⓡ	ACQUIRED organization Ⓡ	DATE date	-	-
company name change	NEW organization Ⓡ	OLD organization Ⓡ	FROM date	TO date	-
spin off	PARENT organization Ⓡ	CHILD organization Ⓡ	DATE date	-	-
marriage	PERSON A person Ⓡ	PERSON B person Ⓡ	CEREMONY location	FROM date	TO date
sibling relationship	PERSON A person Ⓡ	PERSON B person Ⓡ	-	-	-
romantic relationship	PERSON A person Ⓡ	PERSON B person Ⓡ	FROM date	TO date	-
person parent	PERSON person Ⓡ	PARENT A person Ⓡ	PARENT B person	-	-

Often the first k ($k \geq 2$) arguments of an relation are *essential arguments*, since conceptually the relation holds between these entities.[2] Then we require these arguments in every text mention of an instance. For example, we require both persons in a *marriage* relation to be mentioned, whereas date and location of the wedding are considered optional, as well as a supplementary divorce date. All relations which share the NE types of their essential arguments are of the same *essential type*.

Table 1 shows some of the targeted relations from the three domains *Award*, *Business* and *People*. Due to space restrictions, we only present a subset of the 39 used relations here. Required (essential) arguments are marked by ®. Relations of the same essential type are grouped by solid horizontal lines. For example, all three relations from the *Award* domain (i. e., *award nomination*, *award honor* and *hall of fame induction*) belong to the same essential type since their first two arguments are of the same NE types: award concept and person/organization. All relation definitions used in this paper were taken from Freebase.

4 Architecture

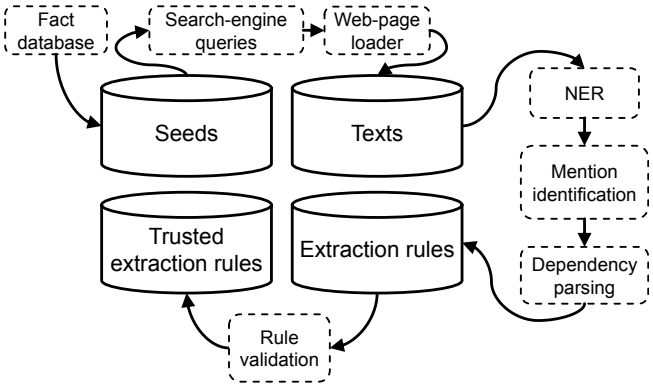

Fig. 1. Data flow of implemented system.

Figure 1 displays the general workflow of our system. First, a local database of relation instances (so-called *seeds*) is generated. The seeds are used as queries for a web search engine, which returns hits potentially containing mentions of the seeds. The web pages are downloaded and transformed into plain texts. After NER, sentences containing at least the essential seed arguments are collected, which are then processed by the dependency parser. We regard a sentence containing at least the essential arguments as a potential mention of a target relation. The parses serve as input for rule learning, which works only on individual sentences. The rule-validation component utilizes information from parallel learning of multiple relations of the same essential types to filter out low-quality rules.

[2] Assuming that relations are defined with their most important arguments preceding the others as they actually are in Freebase and most ontologies

An important design choice is the utilization of the dependency-relation formalism for our rule model. We assume that any given mention of a target-relation instance can be identified by a somehow characteristic pattern in the sentence's underlying dependency graph. This approach has limitations, e. g., it does not cover mentions requiring some kind of semantic understanding (see Section 7), or simply mentions with arguments spread across several sentences. Nevertheless, this methodology is intuitively expressive enough for many mentions. To our knowledge, there exists no systematic investigation of how quantitatively limiting a dependency-formalism based, sentence-restricted approach to RE is.

4.1 Linguistic Annotation

NER in our system is performed by a combination of two components: (a) the *Stanford Named Entity Recognizer*[3] [8] for detection of persons, organizations and locations (extended with our own date recognition), and (b) a simple string fuzzy match via a gazetteer created from the name variations of the seeds' entities as provided by Freebase. In the current system, neither complex entity linking nor coreference resolution are applied in the training phase.

After identification of sentences containing seed mentions, each sentence is processed by the dependency-relation parser *MDParser* (*Multilingual Dependency Parser*)[4] [24]. We choose this parser because it is very fast, while maintaining competitive parsing quality when used in an application, as shown by [25] for the textual entailment task. The parsing results also contain information about part-of-speech tags and word lemmas.

4.2 Rule Learning

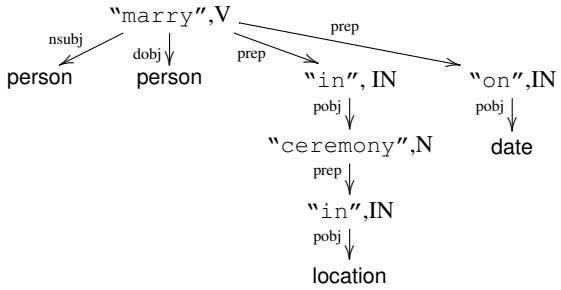

Fig. 2. Dependency parse of Example 2.

We re-use the rule-learning component of the existing DARE system [31, 29]. DARE is a minimally-supervised machine-learning system for RE on free texts, consisting of 1) rule learning (RL) and 2) relation extraction (RE). Starting from a semantic seed (a set of relation instances), RL and RE feed each other in a bootstrapping framework. In

[3] *Stanford CoreNLP* (version 1.1.0) from `http://nlp.stanford.edu/software/corenlp.shtml`

[4] See `http://mdparser.sb.dfki.de/`.

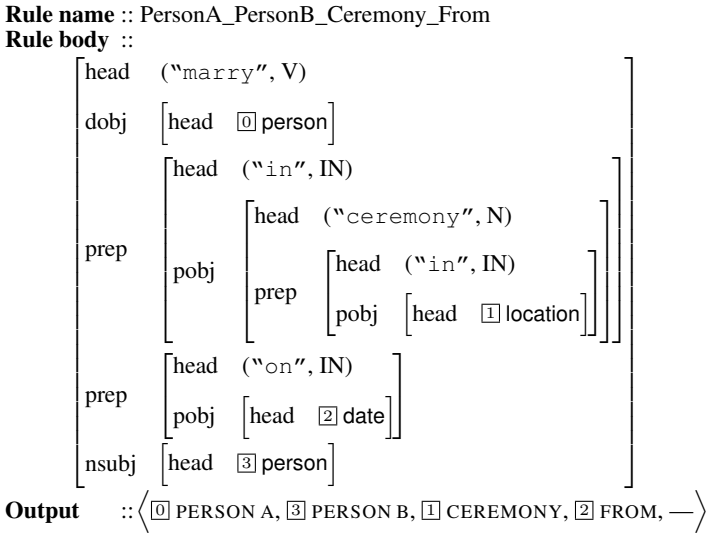

Fig. 3. Example rule for the *marriage* relation.

our system, we use the RL component for the training phase (Section 5) and the RE part in the evaluation (Section 6). DARE is able to directly handle n-ary relations through its extraction-rule formalism, which models the links between relation arguments using dependency relations.

Consider for example the *marriage* relation from Table 1, which has the arguments PERSONA, PERSONB, CEREMONY, FROM, and TO. Given the seed tuple ⟨*Brad Pitt, Jennifer Aniston, Malibu, 2000, 2005*⟩, the following sentence can be used for rule learning:

Example 2. Brad Pitt married Jennifer Aniston in a private wedding ceremony in Malibu on July 29, 2000.

This sentence is processed by the dependency parser, which outputs a structure like in Figure 2, where the surface strings of the named entities have already been replaced by their respective types in this tree via the NER.

From this dependency tree, DARE learns the rule in Figure 3, which contains four arguments: two married persons plus the wedding location and the starting date of the marriage. DARE additionally learns projections of this rule, namely, rules containing a subset of the arguments, e. g., only connecting the person arguments. This way, a single sentence might lead to the learning of several rules.

5 Web-Scale Rule Learning

Our rule learning consists of two phases: candidate-rule learning and rule filtering. As assumed in distant supervision, when there is a sentence containing the arguments of a relation instance, this sentence is a potential mention of the target relation. Therefore, rules learned from such sentences are also potential rules of the target relation. Because

this assumption is not true for all sentences with relation arguments, the resulting rules may be wrong. Hence, they are only *candidate* rules and need further filtering.

5.1 Learning Candidate Rules

Table 2. Number of seeds from Freebase and search hits; statistics about downloaded web pages and documents and sentences containing seed mentions; statistics for rule learning.

Relation	#Seeds	#Seeds used	#Search hits	#Documents w/ a mention	#Sentences w/ a mention	#Rules
award nomination	86,087	12,969	1,000,141	14,245	15,499	7,800
award honor	48,917	11,013	1,000,021	50,680	56,198	40,578
hall of fame induction	2,208	2,208	443,416	29,687	34,718	17,450
organization relationship	219,583	70,946	1,000,009	37,475	51,498	28,903
acquisition	1,768	1,768	308,650	40,541	71,124	50,544
company name change	1,051	1,051	124,612	8,690	10,516	6,910
spin off	222	222	32,613	3,608	5,840	4,798
marriage	16,616	6,294	1,000,174	211,186	381,043	176,949
sibling relationship	8,246	8,246	914,582	130,448	186,228	69,596
romantic relationship	544	544	280,508	82,100	172,640	74,895
person parent	23,879	3,447	1,000,023	148,598	213,869	119,238
avg. of 39 relations	72,576	6,095	635,927	60,584	73,938	41,620

In the following, we describe the experimental results of our training phase. Table 2 provides statistics for this phase. For the 39 target relations, 2.8M relation instances were extracted from Freebase (column "# Seeds"). For each relation, we tried to find 1M web documents using the search engine *Bing*[5] (column "# Search hits"), resulting in more than 20M downloaded documents in total for all relations. Note that for some relations, finding 1M web documents required only a subset of the relation instances retrieved from Freebase, while for other relations even utilizing all relation instances was not sufficient for getting 1M web documents. This explains the difference in numbers between columns "# Seeds" and "# Seeds used".

The downloaded web documents were subsequently processed by NER and sentence segmentation. Given sentences with their NE annotations, the sentences with mentions of seeds are identified. The mentions of seeds occur in a relatively small fraction of the downloaded web documents (around 10 %), as shown in column "# Documents w/ a mention". Reasons for that are 1) seed arguments being spread across sentence borders, 2) NER errors or 3) a wrong (non-English) language of the web document.

The final training corpus contains for each relation on average 74k sentences with mentions of seed instances, i. e., a total of around 3M sentences (column "# Sentences w/ a mention"). All of these mentions include at least the respective relation's essential arguments. On average, around 40k distinct rules were learned per relation (column "# Rules").

[5] `http://www.bing.com`

The total system runtime per relation was on average around 19 hours, with the processing being distributed on three server machines (16 cores with 2.4 GHz each; 64 GB RAM). The parallelization was accomplished naively by chunking the data according to the respective source seed. Of the pipeline's main processing phases, the search-engine querying and the document download with subsequent text extraction were the most time-consuming ones, with on average 6 hours 17 minutes per relation and 8 hours 40 minutes per relation, respectively. The mention-finding step (including NER) took 3 hours 11 minutes for each relation, the final dependency parsing and rule learning on average only 40 minutes per relation.

5.2 Rule Filtering

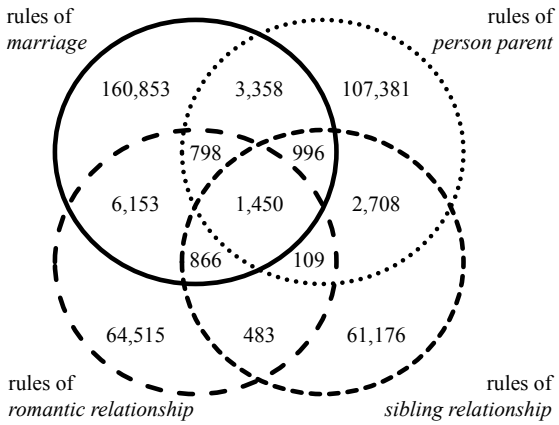

Fig. 4. Euler diagram showing numbers of rules learned for four *People* relations. Missing zones: *person parent* /*romantic relationship* (408); *marriage* /*sibling relationship* (1,808).

Whenever two relations are of the same essential type, they may share some same relation instances, in particular, for the required arguments, for example, the same two persons might be involved in various relations such as marriage and romantic relations. This can be for good reasons, if the relations overlap or if the relevant expressions of the language are ambiguous. Most rules learned for two or more relations, however, are not appropriate for one or both relations. Rules might be learned for wrong relations because of erroneous NER & dependency parsing, false seed facts and false mentions. Especially when a rule is learned for two disjoint relations, something must be wrong. Either the rule exhibits a much higher frequency for one of the two relations, then it can be safely deleted from the other, or the rule is wrong for both relations. Figure 4 shows intersections of the sets of learned rules for four relations of the same essential type in the *People* domain: *marriage*, *romantic relationship*, *person parent*, and *sibling relationship*. Rules in the intersections either express one of the displayed relations or a non-displayed relation or no specific semantic relation at all.

We propose a general and parametrizable filtering strategy using information about the applicability of a rule w. r. t. other relations of the same essential type. If a rule

occurs significantly more often in a relation $\mathcal{R}$ than in another relation $\mathcal{R}'$, this rule most probably belongs to $\mathcal{R}$. Let $f_{r,\mathcal{R}}$ be the frequency of rule r in relation $\mathcal{R}$ (i. e., the number of sentences for $\mathcal{R}$ from which r has been learned) and let $R_{\mathcal{R}}$ be the set of learned rules for $\mathcal{R}$. Then the relative frequency of r in $\mathcal{R}$ is defined as:

$$rf_{r,\mathcal{R}} = \frac{f_{r,\mathcal{R}}}{\sum\limits_{r' \in R_{\mathcal{R}}} f_{r',\mathcal{R}}} \tag{3}$$

Next, we define the first component of our filter. Let $\mathbb{R}$ be a set of relations of the same essential type. The rule r is *valid* for the relation $\mathcal{R} \in \mathbb{R}$ if the relative frequency of r in $\mathcal{R}$ is higher than its relative frequencies for all other relations in $\mathbb{R}$:

$$valid^{\mathcal{R}}_{inter}(r) = \begin{cases} \textit{true} & \text{if } \forall \mathcal{R}' \in \mathbb{R} \backslash \{\mathcal{R}\} : rf_{r,\mathcal{R}} > rf_{r,\mathcal{R}'} \\ \textit{false} & \text{otherwise} \end{cases} \tag{4}$$

The second component is a heuristic which only filters on the frequency of a rule w. r. t. a single relation:

$$valid^{\mathcal{R}}_{freq}(r) = \begin{cases} \textit{true} & \text{if } f_{r,\mathcal{R}} \geq x, \text{where } x \geq 1 \\ \textit{false} & \text{otherwise} \end{cases} \tag{5}$$

With this filter, we ensure that in addition to the relative frequency, there is also enough evidence that r belongs to $\mathcal{R}$ from an absolute point of view. We merge the two components into our final filter, later referred to as the *combined filter*:

$$valid^{\mathcal{R}}_{c}(r) = valid^{\mathcal{R}}_{freq}(r) \wedge valid^{\mathcal{R}}_{inter}(r) \tag{6}$$

Note that all rules that do not contain any content words such as verbs, nouns or adjectives will be deleted before the actual rule filtering takes place. In addition to the frequency heuristic, we also experimented with other features, such as the arity of rules and the length of rules' source sentences. However, their general performance was not superior to the frequency of a rule.

6 Testing and Evaluation

Since we are in particular interested in the recall and coverage performance of our learned rules, we are more dependent on the gold-standard data than precision-driven evaluations as presented in [15], where they evaluate manually the top 100 or 1000 extracted instances of the most popular relations. The ACE 2005 corpus [26] is too sparse for our evaluation goal, for example, there are only 14 mentions containing the essential person arguments for the *marriage* relation. The annotation of the MUC-6 corpus [9] is document-driven and does not provide direct links between relation arguments and sentences. Therefore, we had to prepare a new gold-standard test corpus annotated with relations and their arguments sentence-wise. Because of high annotation costs, we decided to focus on one relation, namely, the *marriage* relation. On this new

gold-standard corpus, we compare our system's web rules against rules learned with the basic DARE system.

In order to know the impact of training corpus size on the coverage of the learned rules in the distant supervision approach, we also compare the recall performance of the rules learned from the Web with rules learned from local corpora of two different sizes. All learned rules are tested against the New York Times part of the English Gigaword 5 corpus [18].

6.1 Learned Marriage Relation Rules

Table 3. Distribution of *marriage* rules across arities. "Avg." – Average, "Med." – Median.

Arity	#Rules	Min. Freq.	Avg. Freq.	Med. Freq.	Max. Freq.
2	145,598	1	3.21	1	64,015
3	26,294	1	2.90	1	2,655
4	4,350	1	3.07	1	603
5	40	1	1.40	1	10

The *marriage* relation has five arguments: PERSONA, PERSONB, CEREMONY, FROM, and TO. A candidate rule must extract at least the two person arguments. The distribution of the rules with respect to their arities is depicted in Table 3. Although many rules are binary, there are more than 20 % of the total rules with arities > 2 (more than 30k). It demonstrates that it is important to learn n-ary rules for the coverage.

6.2 Evaluation on Gold-Standard Corpus

Our gold-standard corpus, dubbed *Celebrity-Gold*, contains crawled news articles from the *People* magazine.[6] This corpus consists of 25,806 sentences with 259 annotated mentions of *marriage*. Out of curiosity, we compare the web-based learning to the bootstrapping approach using the same system components and the same seed (6,294 relation instances). The learning corpus for bootstrapping, dubbed *Celebrity-Training*, is of the same kind and size as Celebrity-Gold. Compared to the 176,949 candidate rules from the Web, the bootstrapping system learned only 3,013 candidate rules.

The learned rules are then applied to Celebrity-Gold for evaluation. It tuns out that our web-based system achieves much higher recall than the bootstrapping system: 49.42 % vs. 30.5 %. As we know, the learned web rules are in fact only candidates for RE rules. Therefore, the baseline precision is relatively low, namely, 3.05 %. Further processing is needed to filter out the wrong rules. Nevertheless, investigating the recall at this stage is very important because even an excellent rule filtering might produce below-average results if there were not enough correct rules to separate from wrong ones during the filtering phase.

[6] `http://www.people.com/`

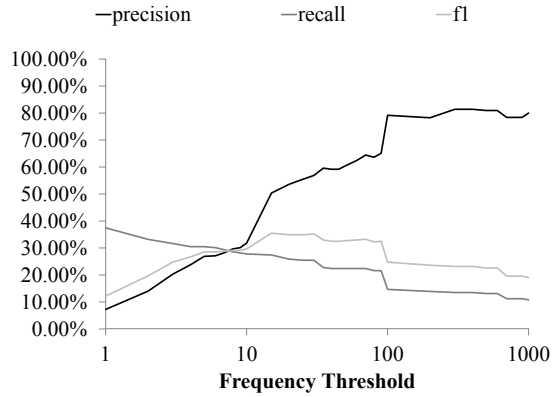

Fig. 5. Performance of web rules after filtering. X-axis: frequency thresholds

Figure 5 depicts the extraction performance after the combined filter $valid_c^{\mathcal{R}}(r)$ is applied to the learned *marriage* rules. The precision improves considerably, in particular, grows with high frequency. The best f-measure can be obtained by setting the frequency to 15 with a precision of around 50% and a recall of around 28%.

6.3 Evaluation with Different Corpus Sizes

After the encouraging results on the small-sized Celebrity-Gold corpus, we evaluated our rules by applying them to a larger corpus, the NYT subset of the English Gigaword 5 corpus (abbr. NYT). Because there is no gold-standard annotation of the *marriage* relation available for this corpus, we use two alternative validation methods: (a) manual checking of all mentions detected by our rules in a random partition of NYT (100,000 sentences) and (b) automatic matching of extracted instances against the Freebase facts about *marriage*s. Note that before RE was performed, we removed all web training sentences from NYT, to avoid an overlap of training and test data.

The performance of the web rules is compared to rules learned on two local corpora in a distant-supervision fashion. The first corpus is the Los Angeles Times/Washington Post part of the Gigaword corpus (abbr. LTW). The second local corpus for rule learning is the corpus used for bootstrapping in Section 6.2: Celebrity-Training. Here, only the rules learned in the first bootstrapping iteration were employed for relation extraction to allow for better comparison. For both local training corpora, the same seed set as for the

Table 4. Statistics about corpus sizes and rule learning.

Corpus	# Docs	# Sentences	# Seeds w/ match	# Generated training sentences	# Rules learned
Web (train.)	873,468	81,507,603	5,993	342,895	176,949
LTW (train.)	411,032	13,812,110	1,382	2,826	1,508
Celebrity-Training (train.)	150	17,100	76	204	302
NYT (test)	1,962,178	77,589,138	–	–	–

web learning was used (i. e., 6,294 instances). Table 4 shows statistics about the corpora and provides information about the learned rules.

Table 5 shows the extraction results of the different rule sets on NYT. The web candidate rules without rule filtering find the highest number of positive marriage mentions of Freebase instances in the corpus, namely, 1,003. This experiment confirms the hypothesis that the extraction coverage of the learned rules increases with the size of the training corpus. After the rule filtering, the web system has improved the precision effectively without hurting recall too much. Note that different kinds of rule filtering may be applied also to the rules learned from Celebrity-Training and LTW. Because the focus of this research is web learning, we only show the results for the web system here.

Table 5. Extraction results on NYT corpus for rules from distant-supervision learning on different corpus sizes. "# Freebase" is short for "# Extracted instances confirmed as correct by Freebase".

			Mentions in sample		
Source of rules	**Filter applied**	**# Freebase**	**# correct**	**# wrong**	**Precision**
Web	–	1,003	76	1,747	4.17 %
LTW	–	721	47	414	10.20 %
Celebrity-Training	–	186	7	65	9.72 %
Web	$valid^{\mathcal{R}}_{inter}(r)$	884	69	869	7,36 %
Web	$valid^{\mathcal{R}}_{c}(r)$, with $x = 15$	627	52	65	44.44 %
Web	$valid^{\mathcal{R}}_{c}(r)$, with $x = 30$	599	51	18	73.91 %

7 Error Analysis

Section 6.2 states that the learned rules covered 49.42 % of the gold-standard mentions in Celebrity-Gold. In this section, we analyze why the system missed the other half of mentions. Table 6 shows the results of a manual investigation of the false negatives of our system on Celebrity-Gold.

Because our system operates on top of NER and parsing results, it heavily depends on correct output of these preprocessing tools. On 41.22 % of false negatives, flawed NER rendered annotated mentions undetectable for extraction rules, even if we had learned *all* possible rules in the training phase. Example errors include unrecognized person entities and broken coreference resolution. Even worse, the parser returned for 59.54 % of the false negatives dependency graphs with errors on the paths between mention arguments, again stopping extraction rules from finding the mentions.

To approximate the system's recall in a setting with perfect linguistic preprocessing, we removed the mistakenly annotated mentions and fixed the errors in NER and parsing. We then reassessed whether a matching extraction rule had been learned in the training phase. Surprisingly, for about half of the remaining false negatives an extraction rule had actually been learned, i. e., the system's main problem is the unreliability of linguistic preprocessing, not a lack of coverage in its rules. In other words, the recall value stated in Section 6.2 would have been about 25 percentage points higher, if NER and parsing had worked perfectly.

Table 6. Analysis of false negatives (abbr.: "fn.") on Celebrity-Gold.

	% of fn.	# of fn.
Total	100.00	131
Annotation error	4.58	6
Linguistic preprocessing error[7]	84.73	111
• NER error	41.22	54
• Parsing error	59.54	78
Total	100.00	125
Matching rule actually learned	50.40	63
No matching rule learned	27.20	34
Semantic understanding required	22.40	28

[7]Note: A fn. might suffer from errors in both NER and parsing.

An error class that cannot be attributed to accuracy deficits of linguistic processing contains sentences that require semantic understanding. These sentences mention an instance of the *marriage* relation, but in an ambiguous way or in a form were the relation is understood by a human, although it is not directly represented in the sentence's structure. The following sentence from the gold-standard corpus is a typical example for this class since the syntactic dependencies do not link "husband" directly to "Ruiz."

Example 3. "... that sounded good to a tired mom like me," says Ruiz, 34, who has two children, James, 8, and Justine, 6, with husband Donald, 42, ...

For a human reader, it is obvious that the phrase "*with husband Donald*" belongs to the person *Ruiz*, because of her mentioning of mother role in the family context. However, attaching the phrase to *Justine* might very well be a reasonable decision for a parser. This becomes clearer when the sentence is slightly changed:

Example 4. "... that sounded good to a tired mom like me," says Ruiz, 34, who awaits her guests, James, 33, and Justine, 35, with husband Donald, 42.

Here, even a human reader cannot decide whose husband *Donald* is. Another example is the following sentence:

Example 5. Like countless mothers of brides, Ellen Mariani smiled until her cheeks ached as she posed for wedding pictures with her daughter Gina, 25, and newly minted son-in-law Christopher Bronley, 22, on Saturday, Sept. 15.

Here it is not clear from the structure that *Christopher Bronley* and *Gina* are spouses. Inference is needed to entail that *Gina* is married to *Christopher Bronley* because she is the daughter of *Ellen Mariani*, who in turn is the mother-in-law of *Christopher Bronley*.

8 Conclusion and Future Work

Our system for the extraction of n-ary relations exploits the Web for training. After achieving an improvement of recall, precision was raised by a rule-filtering scheme that exploits negative evidence obtained from the applicability of a rule to other relations of

the same essential type. The parallel learning of several relations hence proved to be beneficial. We demonstrate that web-scale distant-supervision based rule learning can achieve better recall and coverage than working with local large corpora or bootstrapping on small local corpora. Furthermore, rules with arities > 2 are useful resources for RE.

The error analysis clearly indicates that recall could be much higher if named entity recognition (NER) and parsing worked more accurately. As a consequence of this insight, we will concentrate on the improvement of NER using the rapidly growing resources on the Web and on the adaptation of parsers to the needs of RE, by experimenting with specialized training and parse re-ranking. Another direction of future research will be dedicated to the incorporation of more sophisticated methods for rule filtering. A first step is to exploit additional information on the relationships among the target relations for estimating the validity of rules, another strategy is to re-estimate the confidence of rules during the application phase utilizing constraints derived from the domain model.

Acknowledgments. This research was partially supported by the German Federal Ministry of Education and Research (BMBF) through the project Deependance (contract 01IW11003), by the German Research Foundation (DFG) through the Cluster of Excellence on Multimodal Computing and Interaction (M2CI), and by Google Inc through a Faculty Research Award granted in July 2012.

References

1. Agichtein, E.: Confidence estimation methods for partially supervised information extraction. In: Ghosh, J., Lambert, D., Skillicorn, D.B., Srivastava, J. (eds.) SDM 2006. SIAM (2006)
2. Agichtein, E., Gravano, L.: Snowball: extracting relations from large plain-text collections. In: Fifth ACM Conference on Digital Libraries. pp. 85–94. ACM (2000)
3. Banko, M., Cafarella, M.J., Soderland, S., Broadhead, M., Etzioni, O.: Open information extraction from the Web. In: Veloso, M.M. (ed.) IJCAI 2007. pp. 2670–2676 (2007)
4. Berners-Lee, T.: Weaving the Web: The Original Design and Ultimate Destiny of the World Wide Web by Its Inventor. HarperCollins, New York (1999)
5. Brin, S.: Extracting patterns and relations from the World Wide Web. In: Atzeni, P., Mendelzon, A.O., Mecca, G. (eds.) WebDB'98. LNCS, vol. 1590, pp. 172–183. Springer, Heidelberg (1998)
6. Carlson, A., Betteridge, J., Hruschka Jr., E.R., Mitchell, T.M.: Coupling semi-supervised learning of categories and relations. In: NAACL HLT 2009 Workshop on Semi-supervised Learning for Natural Language Processing. pp. 1–9 (2009)
7. Etzioni, O., Cafarella, M., Downey, D., Popescu, A.M., Shaked, T., Soderland, S., Weld, D.S., Yates, A.: Unsupervised named-entity extraction from the Web: An experimental study. Artif. Intell. 165, 91–134 (2005)
8. Finkel, J.R., Grenager, T., Manning, C.: Incorporating non-local information into information extraction systems by Gibbs sampling. In: ACL 2005 (2005)
9. Grishman, R., Sundheim, B.: Message understanding conference - 6: A brief history. In: COLING 1996. pp. 466–471 (1996)
10. Hoffmann, R., Zhang, C., Weld, D.S.: Learning 5000 relational extractors. In: ACL 2010. pp. 286–295 (2010)
11. Hovy, E.H., Kozareva, Z., Riloff, E.: Toward completeness in concept extraction and classification. In: EMNLP 2009. pp. 948–957 (2009)

12. Kozareva, Z., Hovy, E.H.: A semi-supervised method to learn and construct taxonomies using the Web. In: EMNLP 2010. pp. 1110–1118 (2010)
13. Kozareva, Z., Riloff, E., Hovy, E.H.: Semantic class learning from the Web with hyponym pattern linkage graphs. In: ACL 2008. pp. 1048–1056 (2008)
14. McDonald, R., Pereira, F., Kulick, S., Winters, S., Jin, Y., White, P.: Simple algorithms for complex relation extraction with applications to biomedical IE. In: ACL 2005 (2005)
15. Mintz, M., Bills, S., Snow, R., Jurafsky, D.: Distant supervision for relation extraction without labeled data. In: Su, K.Y., Su, J., Wiebe, J. (eds.) ACL/IJCNLP 2009. pp. 1003–1011 (2009)
16. Nguyen, T.V.T., Moschitti, A.: End-to-end relation extraction using distant supervision from external semantic repositories. In: ACL 2011, Short Papers. pp. 277–282 (2011)
17. Pantel, P., Ravichandran, D., Hovy, E.: Towards terascale semantic acquisition. In: COLING 2004 (2004)
18. Parker, R.: English gigaword fifth edition. Linguistic Data Consortium, Philadelphia (2011)
19. Pasca, M., Lin, D., Bigham, J., Lifchits, A., Jain, A.: Names and similarities on the web: Fact extraction in the fast lane. In: ACL/COLING 2006 (2006)
20. Ravichandran, D., Hovy, E.H.: Learning surface text patterns for a question answering system. In: ACL 2002. pp. 41–47 (2002)
21. Suchanek, F.M., Kasneci, G., Weikum, G.: YAGO: A large ontology from Wikipedia and WordNet. J. Web. Semant. 6, 203–217 (2008)
22. Surdeanu, M., Gupta, S., Bauer, J., McClosky, D., Chang, A.X., Spitkovsky, V.I., Manning, C.D.: Stanford's distantly-supervised slot-filling system. In: Proceedings of the Fourth Text Analysis Conference (2011)
23. Uszkoreit, H.: Learning relation extraction grammars with minimal human intervention: Strategy, results, insights and plans. In: Gelbukh, A.F. (ed.) CICLing 2011, Part II, LNCS, vol. 6609, pp. 106–126. Springer (2011)
24. Volokh, A.: MDParser. Tech. rep., DFKI GmbH (2010)
25. Volokh, A., Neumann, G.: Comparing the benefit of different dependency parsers for textual entailment using syntactic constraints only. In: SemEval-2 Evaluation Exercises on Semantic Evaluation PETE (2010)
26. Walker, C., Strassel, S., Medero, J., Maeda, K.: ACE 2005 multilingual training corpus. Linguistic Data Consortium, Philadelphia (2006)
27. Weld, D.S., Hoffmann, R., Wu, F.: Using Wikipedia to bootstrap open information extraction. SIGMOD Record 37, 62–68 (2008)
28. Wu, F., Hoffmann, R., Weld, D.S.: Information extraction from Wikipedia: moving down the long tail. In: KDD 2009. pp. 731–739 (2008)
29. Xu, F.: Bootstrapping Relation Extraction from Semantic Seeds. Ph.D. thesis, Saarland University (2007)
30. Xu, F., Uszkoreit, H., Krause, S., Li, H.: Boosting relation extraction with limited closed-world knowledge. In: COLING 2010, Posters. pp. 1354–1362 (2010)
31. Xu, F., Uszkoreit, H., Li, H.: A seed-driven bottom-up machine learning framework for extracting relations of various complexity. In: ACL 2007 (2007)
32. Xu, W., Grishman, R., Zhao, L.: Passage retrieval for information extraction using distant supervision. In: IJCNLP 2011. pp. 1046–1054 (2011)
33. Yangarber, R.: Counter-training in discovery of semantic patterns. In: ACL 2003. pp. 343–350 (2003)
34. Yangarber, R., Grishman, R., Tapanainen, P.: Automatic acquisition of domain knowledge for information extraction. In: COLING 2000. pp. 940–946 (2000)
35. Yates, A., Cafarella, M., Banko, M., Etzioni, O., Broadhead, M., Soderland, S.: TextRunner: open information extraction on the Web. In: HLT-NAACL 2007, Demonstrations. pp. 25–26 (2007)