Semantic Rule Filtering for Web-Scale Relation Extraction

Andrea Moro[1], Hong Li[2], Sebastian Krause[2],
Feiyu Xu[2], Roberto Navigli[1] and Hans Uszkoreit[2]

[1] Dipartimento di Informatica, Sapienza Università di Roma
Viale Regina Elena 295, 00161 Roma, Italy
{moro,navigli}@di.uniroma1.it,
[2] Language Technology Lab, DFKI, Alt-Moabit 91c, Berlin, Germany
{lihong,skrause,feiyu,uszkoreit}@dfki.de

Abstract. Web-scale relation extraction is a means for building and extending large repositories of formalized knowledge. This type of automated knowledge building requires a decent level of precision, which is hard to achieve with automatically acquired rule sets learned from unlabeled data by means of distant or minimal supervision. This paper shows how precision of relation extraction can be considerably improved by employing a wide-coverage, general-purpose lexical semantic network, i.e., BabelNet, for effective semantic rule filtering. We apply Word Sense Disambiguation to the content words of the automatically extracted rules. As a result a set of relation-specific relevant concepts is obtained, and each of these concepts is then used to represent the structured semantics of the corresponding relation. The resulting relation-specific subgraphs of BabelNet are used as semantic filters for estimating the adequacy of the extracted rules. For the seven semantic relations tested here, the semantic filter consistently yields a higher precision at any relative recall value in the high-recall range.

Keywords: Relation Extraction, Semantics, WSD, Rule Filtering, Web-scale, Semantic relations

1 Introduction

Information Extraction (IE) automatically finds relevant entities or relations (including facts and events) in natural language texts. The task of Relation Extraction (RE) is to recognize and extract instances of semantic relations between entities or concepts mentioned in these texts. Usually the relations are given, but they may also be induced from the data, as in Open IE [3] where tuples of potential relations are extracted without role labeling. In this paper, we address Web-scale domain-adaptive RE with semantic labeling for given relations of varying arity.

Precision and recall are two important performance measurements. In the past, recall, scalability, domain adaptability and efficiency were regarded as much greater challenges than achieving high precision, because research employed learning data limited in size, types and domains that did not give rise to the noise levels encountered when using the Web as learning corpus. Much research also concentrated on intelligence applications, where recall is much more important than precision. But the limited learning data were not sufficient to overcome the recall barriers, because of the long tail in the skewed frequency distribution of relevant linguistic patterns. Today, the availability of (i) large open-source knowledge databases such as Freebase [6], (ii) nearly unlimited textual resources on the Web and (iii) efficient NLP systems such as dependency parsers (e.g., [2, 46]) enables the creation of large-scale distantly (or minimally) supervised RE systems for many relations with acceptable efficiency [22, 25, 36, 38, 59]. These systems can achieve much better recall without the need for larger volumes of labeled data. Their drawback is their lack of precision, resulting from the large number of candidate patterns which are selected but not sufficiently constrained by the seed knowledge. Filtering by lexical features (e.g., part-of-speech information, word sequences, etc.), syntactic features such as dependency relations, or simple manually-defined heuristics [1, 4, 8, 22, 25] does not suffice. A major open challenge is the exploitation of semantic information in the text and in existing semantic resources beyond the seed data. Several recent approaches add secondary semantic features to their systems which, however, have been shown to offer only slight improvements in RE precision [19, 60].

In this paper, we propose a new method that automatically learns relation-specific lexical semantic resources from a general-purpose knowledge base without any task-specific manual annotation. The input of this unsupervised learning method is a large collection of noisy RE patterns (40K rules per relation on average) acquired by the RE system Web-DARE [22], together with their sentence mentions from 20M Web pages retrieved by searching for the named-entity tuples of the seed facts. The patterns are dependency structures extracted from the parse trees of the sentence mentions. The learning system acquires relation-relevant word senses by applying Word Sense Disambiguation [30] to the words in the patterns and then extracts the corresponding relation-specific lexical semantic subgraphs from a large-scale general purpose lexical semantic network, i.e., BabelNet [31]. These relation-specific subgraphs are utilized as semantic knowledge for filtering out bad rules. In contrast to frequency-based filters, our semantic rule filter, on the one hand, deletes those high-frequency rules which do not contain any relation-relevant words, but at the same time, on the other hand, it also preserves any low-frequency rules which are semantically relevant (owing to their low frequency such rules would previously have been, erroneously, filtered out). It thereby increases both precision and recall.

The main contributions of this paper are to:

- introduce a novel unsupervised, scalable learning method for automatically building relation-specific lexical semantic graphs representing the semantics of the considered relation. Moreover, we show the usefulness of these graphs for filtering semantically irrelevant rules and improving the precision of large-scale RE;
- report on a first comparison of WordNet and BabelNet with respect to improving RE: BabelNet achieves better recall and F-score than WordNet both in rule filtering and in RE;
- demonstrate that relation-specific lexical semantic resources can improve RE performance: For seven semantic relations tested, the semantic filter consistently yields a higher precision at any relative recall value in the high-recall range.

2 Related Work

In recent years several approaches to RE have tried to circumvent the costly, and still not satisfactory, corpus annotation needed for supervised learning. Minimally or weakly supervised methods start with limited initial knowledge and unlabeled data. By a bootstrapping process partial labeling of data and system training are performed in several iterations (e.g., [1, 7, 8, 39, 55]). However, these systems often have to cope with low recall and precision, the latter partially due to semantic drift.

A newer class of approaches, sometimes referred to as distant supervision, utilizes extensive volumes of preexisting knowledge for partially labeling large volumes of data. [25] train a *linear-regression* classifier on *Freebase* relation instances occurring in a large Wikipedia corpus. In order to achieve high precision (without much consideration for recall), lexical features, syntactic dependency information and negative examples are employed. The resulting precision is 67.6% for 10,000 sampled instances of 102 relations.

Open IE systems such as *TextRunner* and its successors [3, 4, 13, 56], together with subsequent developments [48, 27, 28], detect instance candidates of any unknown relation. The Open IE task, however, is faced with even higher levels of noise. Shallow linguistic analysis and numerous heuristics based on lexical features and frequencies are utilized to filter out noisy or irrelevant information for both learning and extraction.

However, all these RE methods are faced with the problem of estimating the confidence of automatically labeled information and learned rules. Some approaches use the confidence value of the extracted instances or the seed examples as feedback for estimating the confidence of rules [1, 7, 55]. In most cases, however, the confidence values rely on redundancy. Many approaches utilize negative examples for filtering [25, 51, 54]. As mentioned above, lexical features such as word sequences or part-of-speech information are often utilized for further filtering [3, 4, 25, 56]. Some approaches employ domain-relevant terms for filtering rules [35, 52]. Web-DARE [22] filters rules by their absolute frequency and their relative frequency in comparison to other related relations (overlap). In order to improve precision, NELL – a large-scale RE system designed to learn factual knowledge from the Web in a never-ending manner [8] – employs the "coupled learning" of a collection of classifiers for several relations. By exploiting this method it is possible to filter out noisy relation instances recognized by mutually exclusive classifiers. However, even if some of these approaches reach the goal of high precision, this is obtained at the cost of recall.

To obtain high precision while at the same time preserving recall, the use of semantic approaches can be highly beneficial. One of the first attempts was presented in [24] where the authors proposed a method for adding semantic features to the labeled data used for training a syntactic parser. However, even if the authors obtained promising results, the major drawback of this approach is the need for huge volumes of annotated data, which, even today, is hard to obtain. Other approaches add semantic features to feature-based RE systems that learn relation-specific extractors [20, 60]. However, none of these approaches has taken full advantage of syntactic and semantic analysis, and thus they have achieved only small improvements [19]. A recent trend in this research strand is the utilization of tree kernel-based approaches, which can efficiently represent high-dimensional feature spaces [36, 59]. However, supervision is stil required and semantic analysis is only marginally employed. In contrast, in this paper we draw only upon semantic knowledge to obtain significant improvements over non-semantic systems.

To integrate and make the most of semantics in RE systems we need a lexical representation of knowledge that can be exploited to obtain a semantic description of the relations. In contrast to many state-of-the-art resources [15, 18, 29], BabelNet [31] integrates encyclopedic (i.e., from Wikipedia) and lexicographic knowledge (i.e., from WordNet) to obtain a rich multilingual "encyclopedic dictionary".

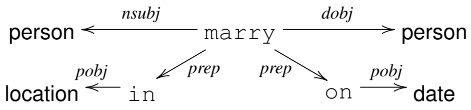
3 Web-DARE and NELL

Our goal is to leverage semantic knowledge to improve the quality of the RE rules learned by two Web-scale RE systems, i.e., Web-DARE and NELL, introduced hereafter.

3.1 Web-DARE

The Web-DARE system [22] learns RE rules for n-ary relations in a distant-supervision manner [25]. For 39 relations, 200k instances, i.e. *seeds*, were collected from the freely-available knowledge base Freebase. Utilizing these relation instances as Web-search queries, a total of 20M Web pages were retrieved and processed, extracting from them 3M sentences mentioning the arguments (entities) of a seed instance. After analyzing these sentences by additional NER and parsing, 1.5M RE rules were extracted from the dependency parses. The following example rule contains four arguments, two married persons plus the wedding location and the starting date of the marriage:

(1)



FO-Filter. The reported recall for Web-DARE is rather high. To overcome the extremely low precision, a rule filter (called FO-Filter) is introduced based on the rule frequency and mutual exclusiveness of relations with similar entity-type signatures. Whenever a particular rule has been learned for more than one relation, it will be added to one relation if its relative frequency in this relation is the highest in comparison to other relations. Rule frequency is the number of the sentence mentions from which a rule has been learned. Relative frequency of a rule in a relation is calculated on the basis of the frequency of this rule in this relation compared to the total frequency of all rules in this relation. Furthermore, a frequency threshold has been applied to exclude rules with low frequency.

3.2 NELL

NELL [8] is a system designed to learn factual knowledge from an immense corpus over a long period. NELL's background ontology contains several hundred entity types (categories) and binary relations, which are related in that certain pairs of categories or relations are marked as being sub- or supersets of each other, or as being mutually exclusive. This coupling of relations is beneficial when estimating the correctness of newly extracted facts. Earlier versions of NELL, described by [5] and [9], were based mainly on a learner of lexico-syntactic rules. The architecture was extended with an extractor working on semi-structured parts of Web pages, i.e., HTML lists and tables, by [10]. Afterwards, a classifier for categorizing noun phrases into entity types based on morphological features as well as an inference-rule learning component was added to NELL [8, 23]. NELL's rules are binary and surface-level oriented, as illustrated by the following example:

(2) person `and husband` person.

While in the NELL system these patterns are only a single piece in a bigger learning and extraction pipeline, we employ them here *on their own* for RE. The NELL rules serve mainly as an additional testing ground for our semantic filter. This implies that the RE results presented in Section 6 are not representative of the performance of NELL itself.

4 WordNet and BabelNet

In this section we give a brief overview of the knowledge bases that we use to obtain a semantic description of the considered relations. The first is WordNet [15] which is a manually-created lexical network of the English language, initiated by George A. Miller in the mid-1980s. The two main components of this resource are the synsets and the semantic relations between them. A synset is a set of synonyms representing the same concept. Each synset is connected to other synsets through lexical and semantic relations. There are roughly 20 relations, among which are hyponymy, meronymy and entailment.

The second resource that we draw upon is BabelNet[3] [31], a large-scale multilingual semantic network which, in contrast to WordNet, was built automatically through the algorithmic integration of Wikipedia and WordNet. Its core components are the Babel synsets, which are sets of multilingual synonyms. Each Babel synset is related to other Babel synsets by semantic relations such as hypernymy, meronymy and semantic relatedness, obtained from WordNet and Wikipedia. Moreover, since BabelNet is the result of the integration of a lexical resource and an encyclopedic resource, it is perfectly in line with the multilingual linguistic Linked Open Data project [12]. This project consists of a vision of the Semantic Web in which a wealth of linguistic resources are linked to each other so as to obtain a bigger and optimal representation of knowledge [32].

One major difference between these two resources is in respect of their considerably different sizes, both in terms of number of concepts and semantic relation instances. On the one hand, WordNet provides roughly 100K synsets, 150K lexicalizations and 300K relation instances. On the other hand, BabelNet contains roughly 5.5M synsets, 15M lexicalizations and 140M relation instances. Moreover, given the multilingual nature of BabelNet (the current version 1.1.1 considers six different languages: Catalan, English, French, German, Italian and Spanish), this resource can exploit multilinguality to perform state-of-the-art knowledge-based Word Sense Disambiguation [33] (in contrast to WordNet which encodes only English lexicalizations), thereby enabling new methods for the automatic understanding of the multilingual (Semantic) Web.

[3] `http://babelnet.org`

5 Rule Filtering with Relation-Specific Semantic Graphs

Current statistical approaches to the rule filtering problem do not take into account the semantic information available within the rules. As a consequence they are not able to identify bad rules, which, from the point of view of the extracted arguments, look correct. For instance, the rule *PERSON met PERSON*, extracted for the relation *married*, is not specific to the considered semantic relation even if it extracts several good relation instances. We tackle this issue by introducing a novel approach to explicitly represent the semantics of each rule and relation. To do this, we apply Word Sense Disambiguation (WSD) to the automatically extracted rules and then build relation-specific semantic graphs which represent the semantics of the considered relation. For instance, our semantic representation of the relation *married* contains concepts that are semantically distant from the concepts usually associated with the term *met*. As a result, our approach is able to correctly filter out the aforementioned rule.

5.1 Building Semantic Graphs

Given a semantic relation ρ, we consider the set of rules R_ρ automatically extracted by the Web-DARE system, together with the set S of sentences from which these rules were extracted. Our goal is to build a semantic representation for the relation ρ. In Algorithm 1 we show the pseudocode of our semantic graph construction approach, described in the following.

WSD (lines 4–13 of Algorithm 1). In this first part of the algorithm we compute a frequency distribution over the synsets of the considered knowledge base for the semantic relation ρ. Given the set S of sentences used by the Web-DARE system and a rule $r \in R_\rho$, we define the subset $S_r \subset S$ as the set of sentences from which the rule r was extracted (see line 6). For instance, we add the sentence *It was here that the beautiful Etta Place first met Harry Longabaugh* to the set $S_{PERSON\_met\_PERSON}$. Then, for each sentence s in S_r, we perform WSD on each content word of the rule r using the remaining content words of s as context (see line 9). For instance, using the previous sentence and given the word *met*, we use as context the following words: *was, here, beautiful, Etta, Place, first, Harry, Longabaugh* obtaining the synset[4] $meet_v^1$. We use an off-the-shelf API for knowledge-based WSD[5] which exploits a knowledge base and graph connectivity measures to disambiguate words [34]. For each synset selected by the WSD API, we increment its count (see line 10 in Algorithm 1). As a result of this step, we obtain Σ_ρ, a synset frequency distribution representing the unstructured semantics of the given relation ρ (see lines 4–10 in Algorithm 1). Then, to avoid data sparsity, we discard all the synsets that occur only once (lines 11–13). For example, given the semantic relation $\rho = marriage$, the most frequent synsets returned by the WSD API are: $marry_v^1$, $wife_n^1$ and $husband_n^1$.

[4] For ease of readability, in what follows we use senses to denote the corresponding synsets. We follow [30] and denote with w_p^i the i-th sense of w with part of speech p.

[5] We did not use supervised approaches as they would have required a separated training phase for each considered domain.

Algorithm 1 Building the relation-specific semantic graph

```
1:  input:   $S$, the set of sentences from which the rules where extracted;
             $R_\rho$, the set of rules automatically extracted for the the relation $\rho$;
             $E_{kb}$, the edges, i.e. the semantic relation instances, of the knowledge base;
             $k$, our free parameter.
2:  output: $G_\rho$, the semantic graph for $\rho$.
3:  function SemanticGraph($S, R_\rho, E_{kb}, k$)
4:      $\Sigma_\rho$ := Map<Synset, Integer>
5:      for each $r \in R_\rho$ do
6:          $S_r := \{s \in S : s \text{ matches } r\}$
7:          for each $sentence \in S_r$ do
8:              for each $word \in \text{contentWords}(r)$ do
9:                  $synset := WSD(word, sentence)$
10:                 $\Sigma_\rho[synset]$++ // we increase by one the integer associated with the synset
11:     for each $synset \in keys(\Sigma_\rho)$ do
12:         if $\Sigma_\rho[synset] = 1$ then
13:             $\Sigma_\rho.remove(synset)$
14:     $\Gamma_\rho := Top(\Sigma_\rho, k)$ // we initialize the core synsets with the top-$k$ most frequent synsets
15:     for each $synset \in keys(\Sigma_\rho)$ do
16:         if $\exists synset' \in Top(\Sigma_\rho, k)$ s.t. $(synset, synset') \in E_{kb}$ then
17:             $\Gamma_\rho := \Gamma_\rho \cup \{synset\}$
18:     return $G_\rho := (\Gamma_\rho, \{(synset_1, synset_2) \in E_{kb} : synset_1, synset_2 \in \Gamma_\rho\})$
```

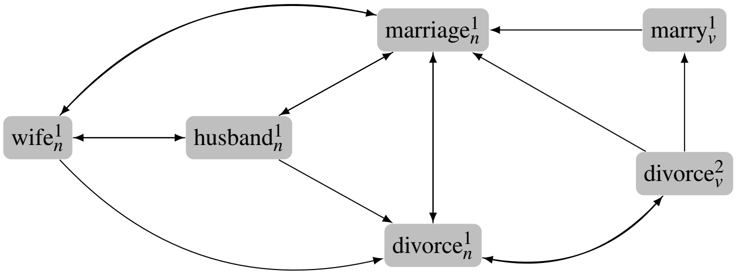


Fig. 1. An excerpt of the semantic graph associated with the relation *marriage* with $k = 2$.

Core Synsets (lines 14–18 of Algorithm 1). In the second part of Algorithm 1 we build a subset $\Gamma_\rho \subseteq \Sigma_\rho$ of *core synsets*, i.e., the most semantically representative concepts for a semantic relation ρ. We initialize Γ_ρ with the top-k most frequent synsets in Σ_ρ (line 14). For instance, with $k = 2$ and the relation $\rho = marriage$, we have $\Gamma_{marriage} := \{marry_v^1, wife_n^1\}$. We then look at each synset s in Σ_ρ and we check if there exists a semantic relation in the knowledge base that connects the synset s to any of the top-k frequent synsets. If this is the case, we augment Γ_ρ with s, i.e., we extend our initial set of core synsets with additional semantically related synsets (lines 15–17). For example, with $k = 2$ and the relation $\rho = marriage$, we add $husband_n^1$, $marriage_n^1$ and $divorce_v^2$ to $\Gamma_{marriage}$, among others (see Figure 1). Finally, the algorithm returns the subgraph G_ρ

Algorithm 2 Classifying the rules of a semantic relation

```
1: input: $G_\rho$, the semantic graph associated with the relation $\rho$;
          $R_\rho$, the set of rules associated with the relation $\rho$;
2: output: $GR$, the good rules
3: function FILTER($G_\rho, R_\rho$)
4:     $GR := \emptyset$
5:     for each $rule \in R_\rho$ do
6:         if $\exists\ word \in \text{contentWords}(rule), synset \in V(G_\rho)$ such that
             $word \in lexicalizations(synset)$ then
7:             $GR := GR \cup \{rule\}$;
8:     return $GR$;
```

of the given knowledge base induced by the set of core synsets Γ_ρ (see line 18), which will be used to filter out bad rules as described in Section 5.2. An excerpt of the kind of graphs that we obtain is shown in Figure 1.

5.2 Filtering Out Bad Rules

We now describe our semantic filter, whose pseudocode is shown in Algorithm 2, which filters out bad rules by exploiting the semantic graph G_ρ previously described. For each rule $r \in R_\rho$ associated with a semantic relation ρ, we check if any of its content words matches one lexicalization of the concepts contained in the semantic graph G_ρ (see line 6). If this is the case we mark r as a good rule, otherwise we filter out r. For instance, our filter recognizes the rule *PERSON married PERSON* as a good rule, while filtering out *PERSON met PERSON* because none of the senses of $meet_v$ matches any of the core synsets automatically associated with the relation *married*.

6 Experiments and Evaluations

6.1 Experimental Setup

Overview. We carried out two different experiments to assess the quality of our semantic filtering algorithm: an intrinsic evaluation (i.e., evaluating the quality of the filtered rules against a gold-standard rule set without taking into account the extraction performance) and an extrinsic evaluation (i.e., determining its effect on recall and precision of real RE). In both evaluations, we experiment with different values of k for Algorithm 1, ranging from 1 to 15.

Our evaluation aims at obtaining insights concerning the following aspects:

- *rule-frequency driven FO Filter vs. filtering based on lexical semantics:* We test our semantic rule filter against the previous FO-Filter to compare the performance difference.
- *impact of the selection among lexical semantic resources:* We evaluate the effects of training our filtering algorithm with two different knowledge bases: manually generated WordNet vs. BabelNet, a massive extension of WordNet automatically created from Wikipedia information (cf. Section 4).

Table 1. Statistics about (a) the input data for the rule filters, (b) the gold standard for intrinsic evaluation, (c) the baseline (pre-filtering) performance for the extrinsic evaluation. Values are shown for both Web-DARE (WD) and NELL (N) systems. “Freebase Mentions” refers to the number of correctly identified Freebase mentions in a sample of the evaluation corpus.

Relation	INPUT		INTRINSIC (SEC. 6.2)	EXTRINSIC (SEC. 6.3)						
	# Rules		# Gold-Set Rules	# Extracted Mentions		Baseline Precision		# Freebase Mentions		
	WD	N	WD (+|-)	WD	N	WD	N	WD	N	WD ∪ N
acquisition	26,986	272	52|48	17,913	296	14.20%	28.04%	93	1	93
marriage	88,350	547	47|53	92,780	2,586	11.60%	8.50%	161	9	168
person birth	22,377	995	50|50	63,819	2,607	36.50%	5.60%	77	0	77
person death	31,559	5	50|50	84,739	17	18.00%	100.00%	300	0	300
person parent	45,093	956	22|78	93,800	358	13.20%	66.20%	91	5	92
place lived	47,689	829	51|49	84,389	3,155	47.90%	92.00%	68	38	106
sibling relationship	26,250	432	12|88	59,465	211	5.60%	51.18%	48	2	49
sum	288,304	4,036	284|416	496,905	9,230	–	–	838	55	885
average	41,186	577	41|59	70,986	1,319	21.00%	50.22%	120	20	126

- *generality of the semantic filtering method:* We also apply our filtering to the NELL rule set to check whether the filtering is general enough to apply beyond DARE rules.

Table 1 lists our target relations in column “Relation” while in column “INPUT” we show the respective number of rules given by the Web-DARE and NELL[6] systems.

6.2 Intrinsic Evaluation

Dataset. For the intrinsic evaluation, we manually validate a set of 700 Web-DARE rules to create a balanced gold standard of correct rules (+) and incorrect ones (-) from all target relations. Column “INTRINSIC” of Table 1 presents the number of manually validated rules per relation.

Results. In this section we describe the intrinsic evaluation of our filtering algorithm. To evaluate the filtered rules, we compute their precision, recall and F-score against the manually built gold standard rule set. We do this without considering the relation extraction performance of the filtered rules, i.e., how many good relation instances are effectively extracted by these rules, as this is the focus of the extrinsic evaluation.

Figure 2 displays precision, recall and F-score values for the total set of seven semantic relations using WordNet and BabelNet as knowledge bases and varying the parameter *k* from 1 to 15 (see Algorithm 1 in Section 5). As Figure 2 shows, we obtain a considerable increase in recall by using BabelNet instead of WordNet (with a maximum

[6] NELL rules were taken from iteration 680, `http://rtw.ml.cmu.edu/resources/results/08m/`

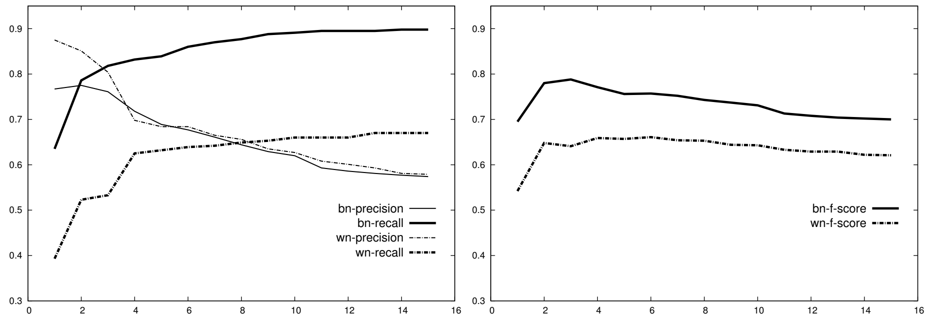


Fig. 2. Precision, Recall and F-score considering all the 7 semantic relations, using WordNet (dotted) and BabelNet (solid), varying the value *k* from 1 to 15.

value of roughly 90% for BabelNet and 70% for WordNet). Despite the gain in recall for the BabelNet filter, precision stays roughly the same as for the WordNet filter (for each value of *k*), which yields an F-score boost of roughly 10%.

The main reason for the observed improvement can be found in the rich set of semantic relation instances of BabelNet, i.e. when using BabelNet as our knowledge base, the filtering method is able to discover semantic connections between concepts that are not provided by WordNet. For instance, WordNet does not contain a semantic connection between the concepts of *marriage* and *divorce*, whereas BabelNet does.

6.3 Extrinsic Evaluation

Dataset. In the extrinsic evaluation, we use the Los Angeles Times/Washington Post (henceforth LTW) portion of the English Gigaword v5 corpus [37] for RE. LTW is comprised of 400K newswire documents from the period 1994–2009. We match all Web-DARE and NELL rules against the LTW corpus, resulting in more than 500K detected relation mentions, shown in column "EXTRINSIC" of Table 1. To estimate the precision of RE, we manually check a random sample of 1K extracted mentions per relation and system, giving us the pre-filtering performance depicted in column "Baseline Precision". To estimate the RE coverage of the rules, we investigate how many mentions of Freebase facts the systems find on LTW. The values are listed in the last three columns of Table 1, labeled "Freebase Mentions". Only actual mentions are taken into account, i.e., sentences containing the entities of a Freebase fact and actually referring to the corresponding target relation. Relative recall values stated in this section are to be understood as recall with respect to the set of Freebase-fact mentions found by at least one of the two rule sets (Web-DARE/NELL), i.e., relative to the very last column of Table 1.

Semantic Filter for Web-DARE Rules. Figure 3 presents the precision vs. relative recall results of RE when performed with the baseline Web-DARE rules, the statistical approach (FO-Filter) and our semantic filtering algorithm (S-Filter) using BabelNet and WordNet. The FO-filter is able to increase the precision from the baseline of 20% up to close to 100%, since by varying the frequency threshold, any value between 40% and

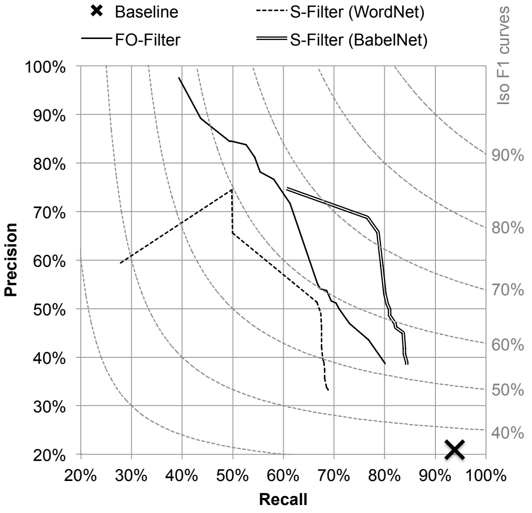


Fig. 3. RE performance of Web-DARE rules with different applied filters. Dashed curves in gray depict points with equal F1 score. For the semantic filter ("S-Filter"), the curves resulted from varying *k* from 1 to 15. FO-Filter is described in [22]. Results are averaged over seven relations.

Table 2. Impact of WordNet (WN) vs. BabelNet (BN) utilization on Web-DARE rule filtering. Results are averaged over seven relations, all values are in %.

k (Alg. 1)	Precision		Recall		F1	
	WN	**BN**	**WN**	**BN**	**WN**	**BN**
(Basel.)	21.00		93.83		34.32	
15	33.24	38.50	68.87	84.37	44.84	52.87
10	38.89	46.16	68.01	82.20	49.48	59.12
5	49.07	52.99	67.40	80.04	56.79	63.76
3	65.57	65.76	49.93	78.69	56.69	71.64
2	74.43	68.79	49.84	76.61	59.70	72.49
1	59.43	74.66	27.84	60.73	37.92	66.98

98% can be reached. But this filtering sacrifices a large portion of the initial recall. In contrast, the semantic filter trained with BabelNet does not permit precision levels above 75% for the average of the relations targeted in this paper, but it has at the same time a more reasonable precision-recall trade-off, e.g., by retaining about 15 percentage points recall above the FO-Filter at a precision level of around 70%. In the recall range covered by the BabelNet filter, its precision is consistently higher.

As illustrated by the chart, training the S-Filter with WordNet instead of BabelNet leads to inferior performance. Table 2 shows the Web-DARE RE performance for different parameter values of Algorithm 1. The use of BabelNet consistently leads to a higher F-score compared to WordNet. For example at $k = 2$, the F-score is roughly thirteen percentage points higher.

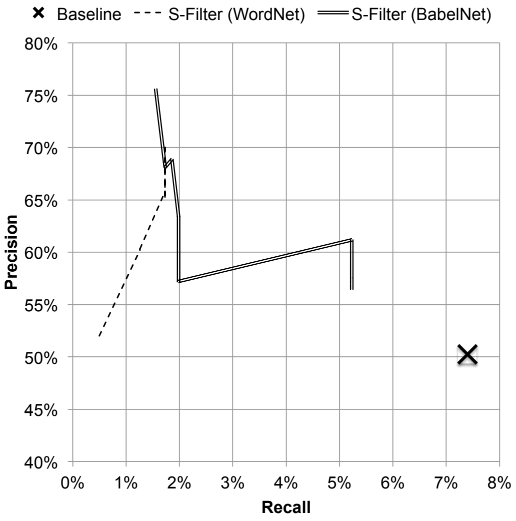


Fig. 4. RE performance of NELL rules, both with and without semantic filter ("S-Filter"). *k* varies from 1 to 15. Results are averaged over seven relations.

Semantic Filter for NELL Rules. Figure 4 shows the precision versus relative recall results of the baseline and our semantic filtering algorithms when applied to NELL's patterns. Again, the RE precision increases. The relative recall values on our test data do not permit any conclusions to be drawn for the NELL system. Due to the low number of mentions found in the NELL recall baseline (see Table 1), the filter application has a high impact on the depicted recall values and thus the curves show a non-monotonic growth. Nevertheless, as the chart indicates, the proposed filter can also be applied to pattern sets of different RE rule formalisms. Similarly to Figure 3, Figure 4 demonstrates that training the filter on BabelNet leads to superior RE performance compared to the filter variant trained on WordNet.

6.4 Result Analysis and Insights

Generality. Both Figures 3 & 4, as well as Table 2, show significant performance improvements after the application of the semantic filter, regardless of the underlying pattern formalism, i.e., dependency-analysis-based or surface-level-based. This means that our algorithm could be applied in a large variety of application scenarios, as long as the patterns or rules contain content words to which the semantic filter can be applied.

BabelNet vs. WordNet. The semantically-enhanced RE performance values of Web-DARE and NELL as given in Sections 6.2 & 6.3 fully support our initial expectation that BabelNet, with its richer inventory of lexical semantic relations, is better suited for effective rule filtering.

Consider the following example from the marriage relation:

(3) $$\textsf{person} \xrightarrow{appos} \texttt{widow} \xrightarrow{prep} \texttt{of} \xrightarrow{pobj} \textsf{person}$$

This rule draws on the concept of deceased spouses, i.e. *widow*, for detecting the target relation. Since the semantic graph created with BabelNet contains this concept, the rule is identified as being useful for RE and hence it is not filtered out, in contrast to the filter from WordNet, which erroneously excludes it.

Individual Relations. The performance of the filter varies across relations. Due to space limitations we cannot show detailed per-relation results here. The filter works particularly well for relations like *acquisition* and *person birth/death*, whereas the results are rather discouraging for *place lived*. Investigating the sampled mentions of the latter relation, we found that this can be attributed to the larger lexical diversity of this relation. Often the semantic information is carried by constructions such as "Belfast writer J. Adams", where the lexical anchor "writer" is semantically insignificant to the relation. To get high coverage on such mentions extraction rules would have to match a certain set of semantically diverse nouns here, without matching all nouns ("Belfast visitor Cameron"). The relation seems to require much background knowledge, which may have to include entailment and other inferences. For example, a mention of a person being a senator for some (US) state could, depending on legal requirements, indeed be a mention for *place lived*.

Semantic Filter vs. FO-Filter. Finally, we investigated the causes of the superior performance of our new semantic filter compared to the pre-existing FO-Filter. In addition to the problem of always finding mutually exclusive relations with compatible entity signatures, the FO-Filter also has the disadvantage of not excluding erroneous rules which belong neither to the particular target relation nor to any of the compatible relations. In contrast, the new semantic filter works independently for each relation.

The following low-precision Web-DARE rules illustrate this point, all learned for the *marriage* relation:

(4) $$\textsf{person} \xleftarrow{nsubj} \texttt{lose} \xrightarrow{prep} \texttt{to} \xrightarrow{pobj} \textsf{person}$$

(5) $$\textsf{person} \xleftarrow{nsubj} \texttt{date} \xrightarrow{dobj} \textsf{person}$$

(6) $$\textsf{person} \xleftarrow{nsubj} \texttt{meet} \xrightarrow{dobj} \textsf{person}$$

These rules, as they express typical relations for married couples, get strong statistical support for the *marriage* relation against the other relations. Therefore, the FO-Filter is not able to correctly identify them as wrong. In contrast, the semantic filter correctly disposes of them.

Another shortcoming of the FO-Filter is the recurring exclusion of high-quality patterns for which there is only limited support in the training data. When taking only the frequency of a pattern into account, these patterns cannot be distinguished from erroneously learned ones. Our use of an additional lexical-semantic resource, such as WordNet/BabelNet, provides a filtering mechanism that correctly identifies the appropriate meaning of the target relation. Consider the following example rule, which, as it has a low frequency, gets filtered out by the FO-Filter, whereas, as it expresses a relevant word sense for the considered relation, gets classified as correct by our semantic filter:

(7) $$\textsf{person} \xleftarrow{\textit{poss}} \texttt{widower} \xrightarrow{\textit{appos}} \textsf{person}$$

7 Conclusion and Outlook

After the successful utilization of parsing for large-scale RE, the time seems ripe for injecting more semantics into this most challenging task within IE. This paper demonstrates that exploiting advanced comprehensive semantic knowledge resources can significantly improve extraction performance.

This is just the beginning, opening the way for new lines of research. The semantic classifier should now be extended for rule classification with respect to relations, building a bridge between traditional IE and open IE. The synonyms provided by semantic resources could also be applied to extend the rule set for increased coverage, in addition to filtering it. As a side result of the comparison between the FO-Filter and the new semantic filter, we observed that the two methods exhibit different shortcomings, giving rise to the hope that a combination may further improve RE performance.

Acknowledgements. This research was partially supported by the European Research Council through the “MultiJEDI” Starting Grant No. 259234, by the German Federal Ministry of Education and Research (BMBF) through the projects Deependance (contract 01IW11003) and Software Campus (contract 01IS12050, sub-project Intellektix) and by Google through a Faculty Research Award granted in July 2012.

References

1. Agichtein, E.: Confidence estimation methods for partially supervised information extraction. In: Proc. of the Sixth SIAM International Conference on Data Mining. (2006)
2. Ballesteros, M., Nivre, J.: Maltoptimizer: An optimization tool for maltparser. In: Proc. of EACL. (2012) 58–62
3. Banko, M., Etzioni, O.: The Tradeoffs Between Open and Traditional Relation Extraction. In: Proc. of ACL/HLT. (2008) 28–36
4. Banko, M., Cafarella, M.J., Soderland, S., Broadhead, M., Etzioni, O.: Open information extraction from the Web. In: Proc. of the 20th IJCAI. (2007) 2670–2676
5. Betteridge, J., Carlson, A., Hong, S.A., Hruschka Jr., E.R., Law, E.L.M., Mitchell, T.M., Wang, S.H.: Toward never ending language learning. In: Proc. of the 2009 AAAI Spring Symposium on Learning by Reading and Learning to Read. (2009)
6. Bollacker, K.D., Evans, C., Paritosh, P., Sturge, T., Taylor, J.: Freebase: a collaboratively created graph database for structuring human knowledge. In: Proc. of SIGMOD. (2008) 1247–1250
7. Brin, S.: Extracting patterns and relations from the World Wide Web. In: Proc. of WebDB. (1998) 172–183

8. Carlson, A., Betteridge, J., Kisiel, B., Settles, B., Jr., E.H., Mitchell, T.: Toward an Architecture for Never-Ending Language Learning. In: Proc. of AAAI. (2010) 1306–1313
9. Carlson, A., Betteridge, J., Hruschka Jr., E.R., Mitchell, T.M.: Coupling semi-supervised learning of categories and relations. In: Proc. of the NAACL HLT 2009 Workskop on Semi-supervised Learning for Natural Language Processing. (2009)
10. Carlson, A., Betteridge, J., Wang, R.C., Hruschka Jr., E.R., Mitchell, T.M.: Coupled semi-supervised learning for information extraction. In: Proc. of WSDM. (2010)
11. Chan, Y.S., Roth, D.: Exploiting Syntactico-Semantic Structures for Relation Extraction. In: Proc. of ACL. (2011) 551–560
12. Chiarcos, C., Nordhoff, S., Hellmann, S.: Linked Data in Linguistics. Representing and Connecting Language Data and Language Metadata. Springer, Heidelberg (2012)
13. Etzioni, O., Fader, A., Christensen, J., Soderland, S., Mausam: Open Information Extraction: The Second Generation. In: Proc. of IJCAI. (2011) 3–10
14. Fader, A., Soderland, S., Etzioni, O.: Identifying Relations for Open Information Extraction. In: Proc. of EMNLP. (2011) 1535–1545
15. Fellbaum, C.: WordNet: an electronic lexical database, Cambridge, MA, USA (1998)
16. Finkel, J.R., Grenager, T., Manning, C.: Incorporating non-local information into information extraction systems by gibbs sampling. In: Proc. of ACL. (2005) 363–370
17. Grishman, R., Sundheim, B.: Message understanding conference - 6: A brief history. In: Proc. of the 16th International Conference on Computational Linguistics, Copenhagen (June 1996)
18. Hoffart, J., Suchanek, F.M., Berberich, K., Weikum, G.: YAGO2: A spatially and temporally enhanced knowledge base from Wikipedia. Artificial Intelligence **194** (2013) 28–61
19. Jiang, J., Zhai, C.: A Systematic Exploration of the Feature Space for Relation Extraction. In: Proc. of NAACL. (2007) 113–120
20. Kambhatla, N.: Combining lexical, syntactic, and semantic features with maximum entropy models for information extraction. In: Proc. of ACL (Demonstration). (2004) 178–181
21. Kozareva, Z., Hovy, E.H.: A semi-supervised method to learn and construct taxonomies using the Web. In: Proc. of EMNLP 2010. (2010) 1110–1118
22. Krause, S., Li, H., Uszkoreit, H., Xu, F.: Large-scale learning of relation-extraction rules with distant supervision from the web. In: Proc. of 11th ISWC, Part I. (2012) 263–278
23. Lao, N., Mitchell, T., Cohen, W.W.: Random walk inference and learning in a large scale knowledge base. In: Proc. of EMNLP. (2011) 529–539
24. Miller, S., Fox, H., Ramshaw, L., Weischedel, R.: A Novel Use of Statistical Parsing to Extract Information from Text. In: Proc. of NAACL. (2000) 226–233
25. Mintz, M., Bills, S., Snow, R., Jurafsky, D.: Distant supervision for relation extraction without labeled data. In: Proc. of ACL/AFNLP. (2009) 1003–1011
26. Mohamed, T., Hruschka, E., Mitchell, T.: Discovering relations between noun categories. In: Proc. of EMNLP. (2011) 1447–1455
27. Moro, A., Navigli, R.: WiSeNet: building a wikipedia-based semantic network with ontologized relations. In: Proc. of CIKM. (2012) 1672–1676
28. Moro, A., Navigli, R.: Integrating Syntactic and Semantic Analysis into the Open Information Extraction Paradigm. In: Proc. of IJCAI. (2013)
29. Nastase, V., Strube, M.: Transforming Wikipedia into a large scale multilingual concept network. Artificial Intelligence **194** (2013) 62–85
30. Navigli, R.: Word Sense Disambiguation: A survey. ACM Comput. Surv. **41**(2) (2009) 1–69
31. Navigli, R., Ponzetto, S.P.: BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. Artificial Intelligence **193** (2012) 217–250
32. Navigli, R.: BabelNet goes to the (Multilingual) Semantic Web. In: Proc. of MSW. (2012)
33. Navigli, R., Ponzetto, S.P.: Joining forces pays off: Multilingual Joint Word Sense Disambiguation. In: Proc. of EMNLP-CoNLL. (2012) 1399–1410

34. Navigli, R., Ponzetto, S.P.: Multilingual WSD with Just a Few Lines of Code: the BabelNet API. In: Proc. of ACL (System Demonstrations). (2012) 67–72
35. Nguyen, Q., Tikk, D., Leser, U.: Simple tricks for improving pattern-based information extraction from the biomedical literature. Journal of Biomedical Semantics **1**(1) (2010)
36. Nguyen, T.V.T., Moschitti, A.: Joint distant and direct supervision for relation extraction. In: Proc. of 5th IJCNLP. (2011) 732–740
37. Parker, R.: English Gigaword fifth edition (2011) Linguistic Data Consortium, Philadelphia.
38. Pasca, M., Lin, D., Bigham, J., Lifchits, A., Jain, A.: Names and Similarities on the Web: Fact Extraction in the Fast Lane. In: Proc. of ACL/COLING. (2006)
39. Ravichandran, D., Hovy, E.H.: Learning surface text patterns for a Question Answering System. In: Proc. of ACL. (2002) 41–47
40. Shinyama, Y., Sekine, S.: Preemptive Information Extraction using Unrestricted Relation Discovery. In: Proc. of HLT-NAACL. (2006)
41. Soderland, S., Roof, B., Qin, B., Xu, S., Mausam, Etzioni, O.: Adapting Open Information Extraction to Domain-Specific Relations. AI Magazine **31**(3) (2010) 93–102
42. Suchanek, F.M., Kasneci, G., Weikum, G.: YAGO: A large ontology from Wikipedia and WordNet. J. Web. Semant. **6** (2008) 203–217
43. Surdeanu, M., Ciaramita, M.: Robust information extraction with perceptrons. In: Proc. of the NIST 2007 Automatic Content Extraction Workshop (ACE07). (March 2007)
44. 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: Proc. of TAC. (2011)
45. Uszkoreit, H.: Learning relation extraction grammars with minimal human intervention: Strategy, results, insights and plans. In: Proc. of CICLing, Part II. (2011) 106–126
46. Volokh, A., Neumann, G.: Comparing the benefit of different dependency parsers for textual entailment using syntactic constraints only. In: Proc. of SemEval. (2010) 308–312
47. Weld, D.S., Hoffmann, R., Wu, F.: Using Wikipedia to bootstrap open information extraction. SIGMOD Record **37** (2008) 62–68
48. Wu, F., Weld, D.S.: Open Information Extraction Using Wikipedia. In: Proc. of ACL. (2010)
49. Wu, F., Hoffmann, R., Weld, D.S.: Information extraction from Wikipedia: moving down the long tail. In: Proc. of KDD. (2008) 731–739
50. Xu, F.: Bootstrapping Relation Extraction from Semantic Seeds. PhD thesis, Saarland University (2007)
51. Xu, F., Uszkoreit, H., Krause, S., Li, H.: Boosting relation extraction with limited closed-world knowledge. In: Proc. of COLING (Posters). (2010) 1354–1362
52. Xu, F., Uszkoreit, H., Li, H.: A seed-driven bottom-up machine learning framework for extracting relations of various complexity. In: Proc. of ACL. (2007)
53. Xu, W., Grishman, R., Zhao, L.: Passage retrieval for information extraction using distant supervision. In: Proc. of IJCNLP. (2011) 1046–1054
54. Yangarber, R.: Counter-training in discovery of semantic patterns. In: Proc. of ACL. (2003)
55. Yangarber, R., Grishman, R., Tapanainen, P.: Automatic acquisition of domain knowledge for information extraction. In: Proc. of COLING. (2000) 940–946
56. Yates, A., Cafarella, M., Banko, M., Etzioni, O., Broadhead, M., Soderland, S.: TextRunner: open information extraction on the Web. In: Proc. of HLT-NAACL (Demo). (2007) 25–26
57. Yates, A., Etzioni, O.: Unsupervised Resolution of Objects and Relations on the Web. In: Proc. of HLT-NAACL. (2007) 121–130
58. Zelenko, D., Aone, C., Richardella, A.: Kernel methods for relation extraction. The Journal of Machine Learning Research **3** (2003) 1083–1106
59. Zhou, G., Qian, L., Fan, J.: Tree kernel-based semantic relation extraction with rich syntactic and semantic information. Inf. Sci. **180**(8) (2010) 1313–1325
60. Zhou, G., Zhang, M.: Extracting relation information from text documents by exploring various types of knowledge. Inf. Process. Manage. **43**(4) (2007) 969–982