Exposé for *Diplom* Thesis

Relation Extraction with Massive Seed and Large Corpora

Sebastian Krause[1]

Supervised by:
Dr. Feiyu Xu[2]
Prof. Dr. Hans Uszkoreit[2]
Prof. Dr. Ulf Leser[1]

May 2011

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

1 Motivation and Background

The research area of information extraction (IE) aims to extract relevant structured information from natural language texts. In addition to the named-entity recognition (NER) task, the identification and classification of relations among entities, namely, the so-called relation extraction (RE) task, is particularly important for many real-world applications. Given the sentence in Figure 1, a RE system should be able to recognize the underlined mentions of entities and their semantic relation, i. e., a *marriage*.

> Pitt met Friends actress Jennifer Aniston in 1998 and married her in a private wedding ceremony in Malibu on July 29, 2000.

Figure 1. Example sentence containing a mention of the *marriage* relation. From: `http://en.wikipedia.org/wiki/Brad_Pitt`, accessed 2011/02/08.

1.1 Problem Definition

A *named entity* is a piece of text string which refers to an entity. The entity can be, e. g., a person of the real world such as `Brad Pitt` or a temporal expression like `July 29, 2000`. The *type* of a named entity is its semantic class, e. g., person for `Brad Pitt` or date for `July 29, 2000`. Let t be a named-entity type and let $\mathcal{NE}_t$ be the set of *all* named entities of type t. Let T be a bag of named-entity types and let $n = |T|$. Then any set $\mathcal{R}$ with

$$\mathcal{R} \subseteq \underset{t \in T}{\times} \mathcal{NE}_t$$

is called an *n-ary relation*. An example for a semantic relation is the *marriage* relation: $\mathcal{R}_{\text{marriage}} \subseteq \mathcal{NE}_{\text{person}} \times \mathcal{NE}_{\text{person}} \times \mathcal{NE}_{\text{date}} \times \mathcal{NE}_{\text{location}}$, which describes on what date and at which location two persons married. Now the task of RE can be defined as to find *mentions* of given semantic relations $\mathcal{R}_1, \mathcal{R}_2, \ldots$ in natural-language texts and to extract *instances* of them.

1.2 The Field of RE

Many RE systems make use of extraction rules based on linguistic patterns. In this methodology, systems learn patterns from sentences that are known to (or assumed by) the system to contain a mention of a certain semantic relation. These patterns are then applied to new unseen sentences to extract information from them.

The underlying linguistic formalisms differ in their analytic depth of the human language. For example, Ravichandran and Hovy (2002) use *surface-text* patterns to find instances of relations like *birth date*. With learned patterns such as "x `was born in` y" or "x `(`y`-`", with x and y being placeholders, they are able to extract birth-date information from sentences like "`Mozart was born in 1756.`" and "`Gandhi (1869-1948) ...`". A similar formalism is applied, e.g., in work from Hearst (1992) or Pantel et al. (2004), who employ so called *lexico-syntactic* patterns to extract the *is-a* relation from sentences. Their patterns are regular expressions over surface-level text and part-of-speech (POS) tags.

Even though these methods are able to extract a lot of information from natural-language texts, their capabilities are limited. Surface-string and lexico-syntactic patterns are often too specific to be reused for new texts. Too many patterns are needed to cover all the possible surface variants of human-language texts. Furthermore, they can often handle only unary or binary relations. They cannot deal with complex relations where mentions of the arguments have a long distance between each other in the text. For example in the sentence shown in Figure 2, it is hard to express the semantic relation between `Bell` and `building products` in a concise way using only lexical properties. More abstract formalisms that exploit the grammatical structure of sentences are of use here. The graph in Figure 3 shows the *dependency relations* between the words in the example sentence of Figure 2. Using this representation of the sentence's structure, the relation between `Bell` and the `products` is more apparent. A system that utilizes dependency relations for RE patterns is, e. g., *DARE*, proposed by Xu et al. (2007).

```
Bell, based in Los Angeles, makes and distributes
electronic, computer and building products.
```

Figure 2. Example sentence containing long-distance semantic relations. From: (de Marneffe and Manning, 2008).

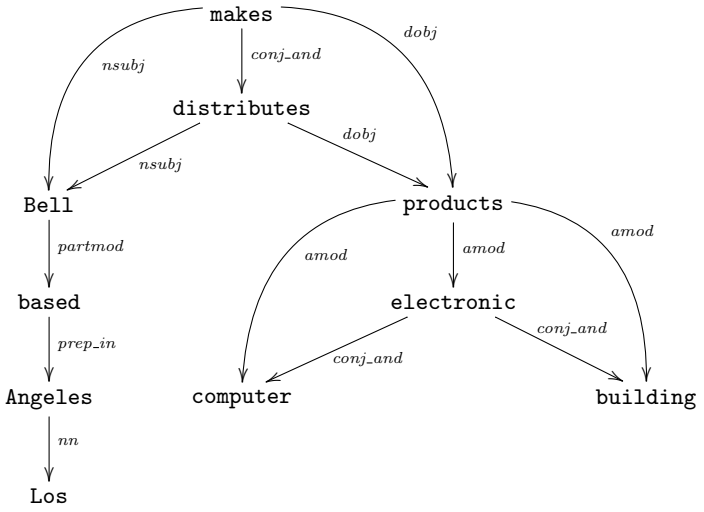

Figure 3. Dependency graph for the sentence of Figure 2. Nodes represent single words of the sentence. Edges denote dependency relations between the connected words. The edge labels describe the type of the dependency relation between two words, see (de Marneffe and Manning, 2008) for a description of possible dependencies.

Another important aspect of RE systems is the size of the supplied text corpus. Today, state-of-the-art systems aim to extract knowledge from large-scale corpora or by directly accessing the web using search engines like *Google*[3] or *Bing*[4]. E. g., Brin (1998) uses a crawl of 5 million web pages for RE and Agichtein and Gravano (2000) use a corpus which contains 320,000 news articles. Pantel et al. (2004) apply their system on a 15 GB newspaper corpus. Zhu et al. (2009) and Carlson et al. (2010a) work with similar corpus sizes, i. e., 1 million web pages and 5 billion sentences from 500 million web pages, respectively. Among the RE approaches that directly employ web search engines are the systems of Ravichandran and Hovy (2002) (using *Altavista*[5]), Etzioni et al. (2005) (using *Google*) and Kozareva and Hovy (2010b) (using *Yahoo!*[6]).

Many recent RE systems (see Section 4) combine large-sized corpora with linguistically-lean sentence analysis. Exploiting the redundancy of fact mentions on the web, i. e., the variety of ways a certain fact is expressed on many different web pages, they achieve considerable results in extracting mentions of semantic relations. But for application areas in which the available corpus is of significantly smaller size, this redundancy assumption does not hold, e. g., when a certain

[3] `http://www.google.com`
[4] `http://www.bing.com`
[5] `http://www.altavista.com`
[6] `http://search.yahoo.com/`

fact is to be found in a single text document. Because of the high specificity and limited linguistic expressiveness of surface-text and lexico-syntactic patterns, they are not well suited for this kind of task. Pattern models with a higher level of abstraction, like the dependency-relation formalism, are more favorable because significantly less rules are needed to cover a large subset of the human language. It is therefore interesting to investigate how such linguistically-rich RE systems can be built and trained using the web.

Another recent development is the growing availability of free repositories of structured data, such as *Freebase*[7], *DBpedia*[8] and *YAGO*[9]. They contain information about a huge number of entities, as well as about their relations with each other. Exploiting this readily accessible knowledge as training data for a RE system for the web is certainly a promising approach to large-scale pattern acquisition

The remainder of this exposé is organized as follows: Section 2 states the goal of the proposed work. Section 3 explains the general approach to achieving the stated goal and Section 4 describes related work in the field of IE. Finally, Section 5 lists concrete steps for realizing the proposed system.

2 Goal

The aim of this work is to examine how the web and large-scale corpora can help to learn linguistically-rich RE patterns. For this setting, a RE system based on the dependency-relation formalism will be implemented.

It will be investigated whether using a massive number of relation instances as training examples to the RE system is beneficial for the quality of the created rules. Another important aspect is the filtering and ranking of the candidate rules.

3 Approach

3.1 Basic Idea

This work proposes a system that aims to learn RE patterns from web text, thus exploiting the huge wealth of linguistic expressions found on the web for machine learning. Using a large existing database with instances of target relations, web search engines are queried with the arguments of the instances (named entities), resulting in a list of web pages mentioning the entities. The web pages' HTML tags are filtered to obtain unstructured text and are then processed by natural-language processing (NLP) tools that perform tokenization, POS tagging, named-entity recognition and dependency parsing. From these linguistically annotated texts, RE patterns are learned. After a pattern ranking and

[7] `http://www.freebase.com/`
[8] `http://dbpedia.org/`
[9] `http://www.mpi-inf.mpg.de/yago-naga/yago/`

confidence estimation is performed, a set of *trusted* patterns remains. Our assumption is that applying these patterns on a given local test corpus will extract relation instances at a higher level of precision and recall than a locally bounded RE system is able to.

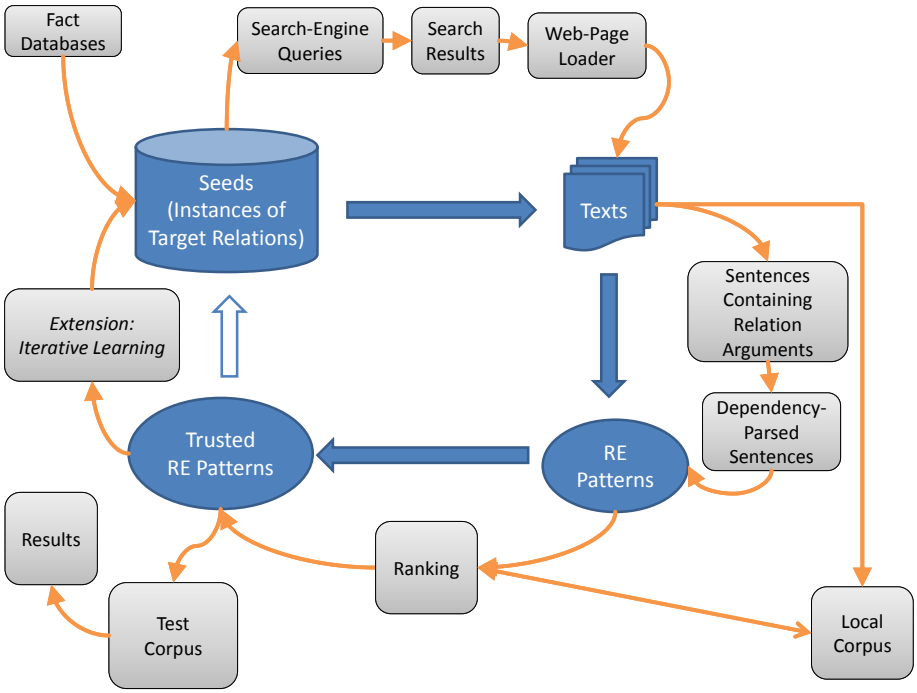

Figure 4. Data flow of proposed system.

Figure 4 depicts the data flow of the proposed system, where blue elements represent the major steps of the algorithm and the other elements describe intermediate steps. *Fact databases* are freely available ontologies and knowledge bases like *YAGO*, *Freebase* and *DBpedia*. For certain target relations, the content of these databases will give us a lot of instances, which can serve as *seeds* to initialize the learning process.

Ranking There are a couple of different approaches to the ranking of the learned patterns. A first idea is that the occurrence frequency of patterns (i. e., the number of web pages a certain pattern is learned from) might correlate to their accuracy in expressing the target relation. Checking whether there exists a natural threshold for the frequency from which on patterns tend to be correct will be the first step.

Another option is to apply some sort of *co-training* of "related" relations, i. e., exploiting mutual exclusion or inclusion between relations. For example, relations like *sibling* and *marriage* cannot share instances. Therefore, instances that are extracted by patterns from such mutually-exclusive relations can help to identify incorrect patterns. It would be interesting to examine whether automatically determining the usefulness of co-training for given target relations is possible. It is also feasible to learn patterns for an *unrelated* relation as proposed by (Mintz et al., 2009). All patterns that appear in that relation can be excluded from the set of learned patterns for actual relations as they are like to be noise.

Applying the closed-world approach from Xu et al. (2010) seems promising as well since it allows for direct incorporation of the already existing domain knowledge, i. e., the fact databases. To apply the closed-world ranking methodology, a method for transforming the initial fact database into closed worlds in an (at least semi-) automatic way has to be invented.

At last, the approach to correctness estimation used in the *SOFIE* and *PROSPERA* systems (see Section 4) could also be adapted, i. e., reasoning about the correctness of extracted instances using first-order logic.

Evaluation Ideally, a RE system is evaluated by applying it to a test corpus and comparing the results, i. e., extracted instances, to a manual annotation of the test corpus, which lists the facts that are actually mentioned in the text. Unfortunately, creating such an annotation is very time-consuming and there are only a few annotated corpora freely available. In the following, some publicly available evaluation corpora are listed, together with the relations for which they are annotated.

For rules of the *prize awarding* relation[10] the *Nobel prize* corpus[11] from Xu et al. (2007) could be employed, which consists of about 3000 newspaper documents from the years 1981–2006. For the subset of $\mathcal{R}_{\text{prize}}$ expressed in this corpus (i. e., $\mathcal{R}_{\text{Nobel prize}}$, which restricts $\mathcal{NE}_{\text{prize}}$ to {`Nobel prize`}), a gold-standard list of possible relation instances exists, i. e., the list of all Nobel prize winners[12]. Using this list, extracted instances can be validated and the quality of the rules can be determined. Apart from Xu et al. (2007), also Kim et al. (2011) used this corpus for RE experiments; both systems' performances can be used as baselines for evaluating this work's system.

[10] $\mathcal{R}_{\text{prize}} \subseteq \mathcal{NE}_{\text{person}} \times \mathcal{NE}_{\text{prize}} \times \mathcal{NE}_{\text{prize area}} \times \mathcal{NE}_{\text{date}}$. Describes that a certain person won a prize in a certain prize area and year.

[11] `http://dare.dfki.de/Daten/nobel/nobel_corpus.html`

[12] `http://nobelprize.org/`

For the relations *person birth*[13], *marriage*[14], *company merger*[15] and *employment tenure*[16] a corpus from the *Automatic Content Extraction* (ACE) program[17] could be used to evaluate the respective rules, in particular, the English part of the "ACE 2005 Multilingual Training Corpus" (Walker et al., 2006). In this corpus, mentions of the relations above are annotated.

The sixth *Message Understanding Conference* (MUC)[18] presented two corpora for IE evaluation, i. e., the "MUC 6" corpus (Chinchor and Sundheim, 2003) and the "MUC 6 Additional News Text" corpus (Chinchor and Sundheim, 1996). These corpora are annotated with structured information of different kinds. Particularly interesting is the annotation from the scenario-template evaluation track, which contains information about *management succession* events suitable for a possible evaluation of rules from a corresponding *management succession* relation[19]. The follow-up conference (MUC 7 [20]) introduced another annotated corpus (Chinchor, 2001). Among other information levels, the contained documents had been analyzed for and marked up with mentions of the binary relations *employee of*[21], *product of*[22] and *location of*[23]. Creating rules for these relations with the proposed system and evaluating them on the MUC 7 corpus is another option for measuring this work' system performance.

For relations with no existing annotated corpus, evaluation must be performed differently. At first a test corpus will have to be created by crawling the web pages of newspapers. Then the trusted, i. e., high-ranked, patterns must be applied to the (linguistically preprocessed but not manually annotated) test corpus to extract relation instances. A sample of these instances is then validated by hand to estimate the extraction precision of the set of trusted patterns. Another possibility is to hold back part of the initial fact database as kind of an *almost gold standard* of the valid relation instances of the target relations. But because the fact databases do not necessarily contain *all* facts mentioned in the test corpus, validating the extracted instances against this data will result only

[13] $\mathcal{R}_{\text{birth}} \subseteq \mathcal{NE}_{\text{person}} \times \mathcal{NE}_{\text{date}} \times \mathcal{NE}_{\text{location}}$. Describes that a certain person was born in a certain location and year.

[14] see Section 1

[15] $\mathcal{R}_{\text{merger}} \subseteq \overbrace{\mathcal{NE}_{\text{organization}} \times \ldots \times \mathcal{NE}_{\text{organization}}}^{k+1 \text{ times}} \times \mathcal{NE}_{\text{date}}$. Describes that k organizations merged on a certain date to form a new organization.

[16] $\mathcal{R}_{\text{job}} \subseteq \mathcal{NE}_{\text{person}} \times \mathcal{NE}_{\text{organization}} \times \mathcal{NE}_{\text{job position}} \times \mathcal{NE}_{\text{date}} \times \mathcal{NE}_{\text{date}}$. Describes that a certain person worked for an organization during a certain period of time.

[17] `http://www.itl.nist.gov/iad/mig/tests/ace/`

[18] `http://www.cs.nyu.edu/cs/faculty/grishman/muc6.html`

[19] $\mathcal{R}_{\text{manager}} \subseteq \mathcal{NE}_{\text{person}} \times \mathcal{NE}_{\text{person}} \times \mathcal{NE}_{\text{position}} \times \mathcal{NE}_{\text{organization}}$. Describes that one person succeeds another person in a certain (job) position in an organization.

[20] `http://www.itl.nist.gov/iaui/894.02/related_projects/muc/`

[21] $\mathcal{R}_{\text{employee of}} \subseteq \mathcal{NE}_{\text{person}} \times \mathcal{NE}_{\text{organization}}$. Describes that a person works for a certain organization.

[22] $\mathcal{R}_{\text{product of}} \subseteq \mathcal{NE}_{\text{product}} \times \mathcal{NE}_{\text{organization}}$. Describes that a company sells a product.

[23] $\mathcal{R}_{\text{location of}} \subseteq \mathcal{NE}_{\text{location}} \times \mathcal{NE}_{\text{organization}}$. Describes that a company has an office in a certain location.

in a *lower* boundary for the actual extraction precision. To determine whether applying the proposed system for learning RE patterns has indeed a positive impact on recall on this corpus, *DARE* by Xu et al. (2007) could also be applied to it. With the (estimated) precision values and the absolute number of extracted instances of the proposed system as well as the baseline *DARE* system, calculating a *relative* recall relating the two approaches is possible.

Other issues An important aspect of the proposed system is the time it needs to process certain parts of the web, i. e., querying search-engines, linguistically processing web pages and learning rules from them. In order to reduce this time, the work flow of the system will be parallelized, in particular, the components performing linguistic analyses. Furthermore, dependency parsing of a single sentence has to be as fast as possible. Hence, choosing the right tools and parsing only promising sentences is vital. E. g., the comparison of Cer et al. (2010) of different parsers for creating dependency relations must be considered.

3.2 Extension: Iterative Learning

It might be beneficial to apply the *trusted* patterns to extract new relation instances from the web. This represents a *bootstrapping* approach to RE, as it is done, e. g., by *DARE*. The question here is how search engines can be queried with dependency-based RE patterns, i. e., how such patterns can be transformed to information retrieval queries. If no applicable solution is found, this problem can be circumvented by extending the pattern learning with an acquisition of lexico-syntactic patterns or pure surface-level string patterns. As detailed in Section 4, they were already successfully used to apply RE to the web. Exploiting these patterns in our system would bridge the gap between learned dependency-based patterns and new seed instances and hence enable the system to learn more patterns from the web.

4 Related Work

This section explains recent approaches to IE on the web.

YAGO and Extensions *YAGO*, presented by Suchanek et al. (2007, 2008), is a large ontology about entities and their relations from *Wikipedia*[24]. The ontology was created by processing the (semi-)structured parts of Wikipedia, i. e., infoboxes and categories. *SOFIE* and *PROSPERA* are RE systems for unstructured text, which are proposed by Suchanek et al. (2009) and Nakashole et al. (2010), respectively. The pattern model is based on lexico-syntactic expressions. *YAGO* is utilized in two ways in these systems. On the one hand it is used to create a limited amount of training examples for the learning process and on the

[24] `http://en.wikipedia.org/`

other hand as trusted base knowledge for a reasoning component. This reasoning model utilizes hand-crafted consistency rules to construct *Horn clauses* from extracted facts, which enables them to view confidence-estimation of extracted facts as a (weighted) *MaxSat* problem. Recently, *YAGO*, the extraction systems and the reasoning components were extended to deal with the time- and space-dependent validity of facts, which led to the system *TOB* (Zhang et al., 2008) and the ontologies *T-YAGO* (Wang et al., 2010) and *YAGO2* (Hoffart et al., 2011).

In contrast to these approaches, this work focuses on learning patterns rather than facts. A shared feature is the exploitation of previously existing knowledge, which we use as initial knowledge for the learning process and for ranking learned rules, as described in Section 3.

Self-supervised Learning and Open IE *KnowItAll* by Etzioni et al. (2004a,b, 2005) is a system mainly intended for harvesting entities and their types from web pages. Using hand-crafted lexico-syntactic patterns as queries to search engines, they are capable of extracting thousands of facts from the web. Their subsequent *TextRunner* system (Banko et al., 2007; Yates et al., 2007) introduces the notion of *open IE*. In contrast to *traditional* RE systems, the relations for which mentions are to be found are not previously named. In the system's training phase, a dependency-parsed corpus is filtered for sentences containing at least two noun phrases. Lexical sentence features and POS tags from the words connecting the noun phrases on the sentence's dependency graph are used as training examples for a classifier. This classifier is then applied to a large POS-tagged corpus to determine for each sentence, whether it contains a mention of a relation. As there are no manually constructed examples, this approach is entitled by the authors as *self-supervised*. In this approach, the name of a relation is derived from the words connecting two entities (noun phrases in this case) in a sentence. Drawback of this approach is that the semantics of a relation between entities is not clear and has to be determined in a separate step (Yates and Etzioni, 2007, *RESOLVER* system).

Similar self-supervised approaches are presented in the *Kylin* system (Wu and Weld, 2007; Wu et al., 2008; Weld et al., 2008), the *WOE* system (Wu and Weld, 2010) and the *Luchs* system (Hoffmann et al., 2010). They parse Wikipedia infoboxes to generate relation instances, which they use as training examples for learning extractors basing on conditional random fields over shallow sentence features. *WOE* also learns relation-independent patterns based on dependency parses of sentences.

When dealing with large amounts of extracted knowledge, effective methods for reasoning and inference are helpful. *Markov Logic Networks*, as proposed by Richardson and Domingos (2006), represent such a method. An application is given in the systems *Holmes* (Schoenmackers et al., 2008) and *Sherlock* (Schoenmackers et al., 2010), which build on the data extracted by TextRunner to learn and apply inference rules between the relations.

Our approach differs to the ones above mainly in that we make extensive use of already structured prior knowledge as training examples and in that we focus on learning only selected relations.

Learning Hyponyms and Binary Relations from the Web Pantel et al. (2004) aim to learn instances of the *is-a* relation from large corpora. They compare the results when applying a clustering-based approach using features from dependency parses of sentences and when utilizing a lexico-syntactic pattern approach. Finding that a higher accuracy of the dependency-parsing approach comes with significantly higher runtime, their conclusion is that for extraction on large corpora, a linguistically-lightweight (shallow) approach is the best. Consequently, related publications, e. g., by Kozareva et al. (2008); Hovy et al. (2009); Kozareva and Hovy (2010b), repeatedly use lexico-syntactic patterns for learning hyponyms from the web and construction of taxonomies. Kozareva and Hovy (2010a) extend the learning of hyponyms to the learning of selectional restrictions for open IE patterns, i. e., determining the valid entity types of the relation arguments .

Ravichandran and Hovy (2002) present an algorithm that aims to extract binary relations from the web using surface-level text patterns. This algorithm is embedded in the *Espresso* system by Pantel and Pennacchiotti (2006), which extends it by a ranking component that utilizes search-engine queries to estimate the correctness of patterns.

Unlike these approaches, we focus explicitly on learning patterns for dependency-parsed sentences. Furthermore, we aim to extract instances for n-ary (i. e., not only binary) relations, which is not part of the work described above.

Large-scale IE on the Web *NELL* 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 Betteridge et al. (2009) and Carlson et al. (2009), relied mainly on a learner of lexico-syntactic patterns. The architecture is extended with an extractor working on semi-structured parts of web pages, i. e., HTML lists and tables, by Carlson et al. (2010b). Recently, a classifier for categorizing noun phrases into entity types based on morphological features as well as an inference-rule learning component has been added to *NELL* by Carlson et al. (2010a).

A major similarity to our approach is the aim to learn knowledge from a corpus of huge size. However, the focus of our work is to not to extract as much facts as possible, but instead to learn patterns which can be applied for fact extraction later. Moreover, we exploit dependency-relation analyses for pattern creation and do not use only lexico-syntactic patterns.

Distant Supervision A particularly interesting approach to RE is proposed by Mintz et al. (2009). Their idea is to train a linear-regression classifier on examples derived from mentions of *Freebase* relation instances in a large Wikipedia corpus. They focus on approximately 100 relations which are among the most frequent ones in *Freebase*. The learned classifier works on shallow features like word sequences and POS tags and on dependency relations between words. With using both a large corpus and a vast amount of relation instances as training examples, their work is closely related to ours. However, our goal is not to learn a classifier, but instead explicit patterns.

5 Work plan

This section briefly states a list of steps towards the goal of Section 2.

May 2011 NER and parsing tools will be tested for speed and accuracy. Fact databases will be processed to obtain initial seeds. Closed-world subsets of fact databases will be created. Proper target relations will be selected with respect to a possible co-training of several relations.

June–July 2011 Proposed system will be implemented. Evaluation will be performed, both for the system on its own and in comparison to the *DARE* system.

August 2011 Basic system will be extended with iterative learning of patterns. Extended system will be evaluated against the basic system and *DARE*.

References

Agichtein, E. and Gravano, L. (2000). Snowball: extracting relations from large plain-text collections. In *DL '00: Proceedings of the fifth ACM conference on Digital libraries*, pages 85–94, New York, NY, USA. ACM.

Banko, M., Cafarella, M. J., Soderland, S., Broadhead, M., and Etzioni, O. (2007). Open information extraction from the web. In *Proceedings of the 20th International Joint Conference on Artificial Intelligence*, pages 2670–2676.

Betteridge, J., Carlson, A., Hong, S. A., Hruschka Jr., E. R., Law, E. L. M., Mitchell, T. M., and Wang, S. H. (2009). Toward never ending language learning. In *Proceedings of the 2009 AAAI Spring Symposium on Learning by Reading and Learning to Read*.

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

Carlson, A., Betteridge, J., Hruschka Jr., E. R., and Mitchell, T. M. (2009). Coupling semi-supervised learning of categories and relations. In *Proceedings of the NAACL HLT 2009 Workskop on Semi-supervised Learning for Natural Language Processing*.

Carlson, A., Betteridge, J., Kisiel, B., Settles, B., Hruschka Jr., E. R., and Mitchell, T. M. (2010a). Toward an architecture for never-ending language learning. In *Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2010, Atlanta, Georgia, USA, July 11-15, 2010*.

Carlson, A., Betteridge, J., Wang, R. C., Hruschka Jr., E. R., and Mitchell, T. M. (2010b). Coupled semi-supervised learning for information extraction. In *Proceedings of the Third International Conference on Web Search and Web Data Mining, WSDM 2010, New York, NY, USA, February 4-6, 2010*.

Cer, D., de Marneffe, M.-C., Jurafsky, D., and Manning, C. D. (2010). Parsing to stanford dependencies: Trade-offs between speed and accuracy. In *Proceedings of the International Conference on Language Resources and Evaluation, LREC 2010, 17-23 May 2010, Valletta, Malta*. European Language Resources Association.

Chinchor, N. (2001). Message understanding conference (MUC) 7. Linguistic Data Consortium, Philadelphia.

Chinchor, N. and Sundheim, B. (1996). Message understanding conference (MUC) 6 additional news text. Linguistic Data Consortium, Philadelphia.

Chinchor, N. and Sundheim, B. (2003). Message understanding conference (MUC) 6. Linguistic Data Consortium, Philadelphia.

de Marneffe, M.-C. and Manning, C. D. (2008). *Stanford Dependencies manual*. The Stanford Natural Language Processing Group.

Etzioni, O., Cafarella, M., Downey, D., Kok, S., Popescu, A., Shaked, T., Soderland, S., Weld, D., and Yates, A. (2004a). Web-scale information extraction in KnowItAll (preliminary results). In *Proceedings of the 13th international conference on World Wide Web*, pages 100–110. ACM.

Etzioni, O., Cafarella, M., Downey, D., Popescu, A.-M., Shaked, T., Soderland, S., Weld, D. S., and Yates, A. (2004b). Methods for domain-independent information extraction from the web: an experimental comparison. In *Proceedings of the 19th National Conference on Artificial Intelligence (AAAI 2004)*.

Etzioni, O., Cafarella, M., Downey, D., Popescu, A.-M., Shaked, T., Soderland, S., Weld, D. S., and Yates, A. (2005). Unsupervised named-entity extraction from the web: An experimental study. *Artificial Intelligence*, 165(1):91 – 134.

Hearst, M. A. (1992). Automatic acquisition of hyponyms from large text corpora. In *COLING 1992 Volume 2: The 15th International Conference on Computational Linguistics*, pages 539–545, Nantes, France.

Hoffart, J., Suchanek, F. M., Berberich, K., Kelham, E. L., de Melo, G., and Weikum, G. (2011). Yago2: Exploring and querying world knowledge in time, space, context, and many languages. In *Proceedings of the 20th International World Wide Web Conference (WWW 2011)*, Hyderabad, India.

Hoffmann, R., Zhang, C., and Weld, D. S. (2010). Learning 5000 relational extractors. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, Uppsala, Sweden. Association for Computational Linguistics.

Hovy, E. H., Kozareva, Z., and Riloff, E. (2009). Toward completeness in concept extraction and classification. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, EMNLP 2009, 6-7 August 2009, Singapore, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 948–957.

Kim, S., Jeong, M., and Lee, G. G. (2011). A local tree alignment approach to relation extraction of multiple arguments. *Information Processing & Management*, In Press, Corrected Proof.

Kozareva, Z. and Hovy, E. (2010a). Learning arguments and supertypes of semantic relations using recursive patterns. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1482–1491. Association for Computational Linguistics.

Kozareva, Z. and Hovy, E. H. (2010b). A semi-supervised method to learn and construct taxonomies using the web. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, EMNLP 2010, 9-11 October 2010, MIT Stata Center, Massachusetts, USA, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1110–1118.

Kozareva, Z., Riloff, E., and Hovy, E. H. (2008). Semantic class learning from the web with hyponym pattern linkage graphs. In *ACL 2008, Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics, June 15-20, 2008, Columbus, Ohio, USA*, pages 1048–1056.

Mintz, M., Bills, S., Snow, R., and Jurafsky, D. (2009). Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, volume 2, pages 1003–1011. Association for Computational Linguistics.

Nakashole, N., Theobald, M., and Weikum, G. (2010). Find your advisor: Robust knowledge gathering from the web. In Dong, X. L. and Naumann, F., editors, *Proceedings of the 13th International Workshop on the Web and Databases 2010, WebDB 2010, Indianapolis, Indiana, USA, June 6, 2010*.

Pantel, P. and Pennacchiotti, M. (2006). Espresso: Leveraging generic patterns for automatically harvesting semantic relations. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, Sydney, Australia. The Association for Computer Linguistics.

Pantel, P., Ravichandran, D., and Hovy, E. (2004). Towards terascale semantic acquisition. In *Proceedings of Coling 2004*, pages 771–777, Geneva, Switzerland. COLING.

Ravichandran, D. and Hovy, E. H. (2002). Learning surface text patterns for a question answering system. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), Philadelphia, July 2002*, pages 41–47.

Richardson, M. and Domingos, P. (2006). Markov logic networks. *Machine Learning*, 62:107–136.

Schoenmackers, S., Davis, J., Etzioni, O., and Weld, D. (2010). Learning first-order horn clauses from web text. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 1088–1098, Cambridge, MA. Association for Computational Linguistics.

Schoenmackers, S., Etzioni, O., and Weld, D. S. (2008). Scaling textual inference to the web. In *2008 Conference on Empirical Methods in Natural Language Processing, EMNLP 2008, Proceedings of the Conference, 25-27 October 2008, Honolulu, Hawaii, USA, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 79–88.

Suchanek, F. M., Kasneci, G., and Weikum, G. (2007). Yago: A core of semantic knowledge. In *16th international World Wide Web conference (WWW 2007)*, New York, NY, USA. ACM Press.

Suchanek, F. M., Kasneci, G., and Weikum, G. (2008). Yago: A large ontology from wikipedia and wordnet. *Web Semantics: Science, Services and Agents on the World Wide Web*, 6(3):203–217. World Wide Web Conference 2007 Semantic Web Track.

Suchanek, F. M., Sozio, M., and Weikum, G. (2009). Sofie: a self-organizing framework for information extraction. In Quemada, J., León, G., Maarek, Y. S., and Nejdl, W., editors, *Proceedings of the 18th International Conference on World Wide Web, WWW 2009, Madrid, Spain, April 20-24, 2009*, pages 631–640. ACM.

Walker, C., Strassel, S., Medero, J., and Maeda, K. (2006). Ace 2005 multilingual training corpus. Linguistic Data Consortium, Philadelphia.

Wang, Y., Zhu, M., Qu, L., Spaniol, M., and Weikum, G. (2010). Timely yago: harvesting, querying, and visualizing temporal knowledge from wikipedia. In *EDBT 2010, 13th International Conference on Extending Database Technology, Lausanne, Switzerland, March 22-26, 2010, Proceedings*, pages 697–700.

Weld, D. S., Hoffmann, R., and Wu, F. (2008). Using wikipedia to bootstrap open information extraction. *SIGMOD Record*, 37(4):62–68.

Wu, F., Hoffmann, R., and Weld, D. S. (2008). Information extraction from wikipedia: moving down the long tail. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Las Vegas, Nevada, USA, August 24-27, 2008*, pages 731–739.

Wu, F. and Weld, D. S. (2007). Autonomously semantifying wikipedia. In *Proceedings of the Sixteenth ACM Conference on Information and Knowledge*

Management, CIKM 2007, Lisbon, Portugal, November 6-10, 2007, pages 41–50.

Wu, F. and Weld, D. S. (2010). Open information extraction using wikipedia. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 118–127, Uppsala, Sweden. Association for Computational Linguistics.

Xu, F., Uszkoreit, H., Krause, S., and Li, H. (2010). Boosting relation extraction with limited closed-world knowledge. In *Coling 2010: Posters*, pages 1354–1362, Beijing, China. Coling 2010 Organizing Committee.

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

Yates, A., Cafarella, M., Banko, M., Etzioni, O., Broadhead, M., and Soderland, S. (2007). Textrunner: open information extraction on the web. In *Proceedings of Human Language Technologies: The Annual Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, NAACL-Demonstrations '07, pages 25–26, Morristown, NJ, USA. Association for Computational Linguistics.

Yates, A. and Etzioni, O. (2007). Unsupervised resolution of objects and relations on the web. In *Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings, April 22-27, 2007, Rochester, New York, USA*, pages 121–130.

Zhang, Q., Suchanek, F. M., Yue, L., and Weikum, G. (2008). Tob: Timely ontologies for business relations. In *11th International Workshop on the Web and Databases, WebDB 2008, Vancouver, BC, Canada, June 13, 2008*.

Zhu, J., Nie, Z., Liu, X., Zhang, B., and Wen, J.-R. (2009). Statsnowball: a statistical approach to extracting entity relationships. In *Proceedings of the 18th International Conference on World Wide Web, WWW 2009, Madrid, Spain, April 20-24, 2009*, pages 101–110.