Unsupervised Relation Extraction of In-Domain Data from Focused Crawls

Steffen Remus

FG Language Technology
Computer Science Department, Technische Universität Darmstadt

Information Center for Education
German Institute for Educational Research (DIPF)

remus@cs.tu-darmstadt.de

Abstract

This thesis proposal approaches unsupervised relation extraction from web data, which is collected by crawling only those parts of the web that are from the same domain as a relatively small reference corpus. The first part of this proposal is concerned with the efficient discovery of web documents for a particular domain and in a particular language. We create a combined, focused web crawling system that automatically collects relevant documents and minimizes the amount of irrelevant web content. The collected web data is semantically processed in order to acquire rich in-domain knowledge. Here, we focus on fully unsupervised relation extraction by employing the extended distributional hypothesis. We use distributional similarities between two pairs of nominals based on dependency paths as context and vice versa for identifying relational structure. We apply our system for the domain of educational sciences by focusing primarily on crawling scientific educational publications in the web. We are able to produce promising initial results on relation identification and we will discuss future directions.

1 Introduction

Knowledge acquisition from written or spoken text is a field of interest not only for theoretical reasons but also for practical applications, such as semantic search, question answering and knowledge management, just to name a few.

In this work, we propose an approach for *unsupervised relation extraction* (URE) where we make use of the *Distributional Hypothesis* by Harris (1954). The underlying data set is collected from the world wide web by focusing on web documents that are from the same domain as a small initialization data set that is provided beforehand. We hereby enrich this existing, domain-defining, corpus with more data of the same kind. This is needed for practical reasons when working with the Distributional Hypothesis (Harris, 1954): A lot of data is required for plausible outcomes and an appropriate coverage. However, we want as little irrelevant data as possible. The proposal's contribution is thus twofold: *a*) focused crawling, and *b*) unsupervised relation extraction. As a particular use case, we are especially interested in scientific publications from the German educational domain. However, we would like to point out that the methodology itself is independent of language and domain and is generally applicable to any domain.

This work is structured as follows: First we will motivate our combined approach and introduce each part individually. We then present related work in Section 2. Section 3 explains the methodology of both parts, and in Section 4 we outline the evaluation procedure of each of the components individually. This is followed by some preliminary results in Section 5, and Section 6 concludes this proposal with some prospects for future work.

1.1 Motivation

The identification of relations between entities solely from text is one of many challenges in the development of language understanding system (Carlson et al., 2010; Etzioni et al., 2008); and yet it is the one step with the highest information gain. It is used e.g. for taxonomy induction (Hearst, 1992) or ontology accumulation (Mintz et al., 2009) or even for identifying facts that express general knowledge and that often recur (Chambers and Jurafsky, 2011). Davidov et al. (2007) performed unsupervised relation extraction by actively mining the web and showed major improve-

Proceedings of the Student Research Workshop at the 14th Conference of the European Chapter of the Association for Computational Linguistics, pages 11–20,
Gothenburg, Sweden, April 26-30 2014.

ments in the detection of new facts from only little initial seed. They used a major web search engine as a vital component of their system. According to Kilgarriff (2007), however, this strategy is unreliable and should be avoided. Nevertheless, the web is undeniably the largest source for any kind of data, and we feel the need for developing easy-to-use components that make it possible to create corpora from the web with only little effort (cf. e.g. Biemann et al. (2013)). When it comes to specific in-domain information, the complete world wide web is first of all too vast to be processed conveniently, and second the gain is little because of too much irrelevant information. Thus we need methods for reducing the size of data to process without losing the focus on the important information and without using web search engines. The combination of a focused crawling system with a subsequent unsupervised relation extraction system enables the acquisition of richer in-domain knowledge than just relying on little local data, but without having to process petabytes of data and still not relying of web search. And yet, by using the web as a resource, our system is generally applicable and independent of language and target domain.

1.2 Focused Crawling

The first part of this proposal is concerned with the efficient discovery of publications in the web for a particular domain. The domain definition is given as a limited number of reference documents. An extra challenge is, that non-negligible amounts of scientific publications are only available as pdf documents, which makes the necessity of new focused crawling techniques even more important. This holds especially for our target use case, the German educational domain. In Section 2.1 we will discuss this issue in more detail. We develop a focused web crawling system which collects primarily relevant documents and ignores irrelevant documents and which is particularly suited for harvesting documents from a predefined specific domain.

1.3 Unsupervised Relation Extraction

The second part of this proposal is the semantic structuring of texts — in our particular use case scientific publications from the educational domain — by using data-driven techniques of computational semantics. The resulting structure enables forms of post-processing like inference or reasoning. In the semantic structuring part, the overall goal is to discover knowledge which can then be used in further steps. Specifically, we will focus on unsupervised relation extraction.

2 Related Work

2.1 Focused Crawling

The development of high-quality data-driven semantic models relies on corpora of large sizes (Banko and Brill, 2001; Halevy et al., 2009), and the world wide web is by far the biggest available source of textual data. Nowadays, a large number of research projects rely on corpora that comes from data in the world wide web. The Web-as-Corpus Kool Yinitiative[1] (WaCKy) (Baroni et al., 2009) for example produced one of the largest corpora used in linguistic research which comes from web documents. Another research initiative which produces a variety of corpora by crawling the web is the COW[2] (corpora from the web) project (Schäfer and Bildhauer, 2012). Currently one of the largest N-gram corpora coming from web data is the Google V1 and Google V2 (Lin et al., 2010), which are used e.g. for improving noun phrase parsing (Pitler et al., 2010). Also the predecessor Google Web1T (Brants and Franz, 2006), which is computed from 1 Trillion words from the web, is heavily used in the community.

All these corpora are generated from general texts which either come from crawling specific *top-level-domains* (tlds) or preprocessing and filtering very large amounts of texts for a specified language. Additionally, we are not aware of any corpus that is created by collecting pdf documents. This is especially an issue when aiming at a corpus of scientific publications, such as e.g. the ACL anthology[3] (Bird et al., 2008). As of today, electronic publications are primarily distributed as pdf documents. Usually these are omitted by the particular crawler because of a number of practical issues, e.g. difficulties in extracting clean plaintext.

Further, we are not interested in sheer collection size, but also in domain specificity. Crawling is a time-consuming process and it comes with logistic challenges for processing the resulting data. While standard breadth-first or depth-first crawling strategies can be adjusted to include pdf files, we want to avoid to harvest the huge bulk of data

[1] http://wacky.sslmit.unibo.it/
[2] http://hpsg.fu-berlin.de/cow/
[3] http://acl-arc.comp.nus.edu.sg/

that we are not interested in, namely those documents that are of a different topical domain as our initial domain definition.

In focused crawling, which is sometimes also referred to as topical crawling, web crawlers are designed to harvest those parts of the web first that are more interesting for a particular topic (Chakrabarti et al., 1999). By doing so, task-specific corpora can be generated fast and efficient. Typical focused crawlers use machine learning techniques or heuristics to prioritize newly discovered URIs (unified resource identifier) for further crawling (Blum and Mitchell, 1998; Chakrabarti et al., 1999; Menczer et al., 2004). In our scenario however, we do not rely on positively and negatively labeled data. The source documents that serve as the domain definition are assumed to be given in plain text. The development of tools that are able to generate in-domain web-corpora from focused crawls is the premise for further generating rich semantic models tailored to a target domain.

2.2 Unsupervised Relation Extraction

The unsupervised relation extraction (URE) part of this proposal is specifically focused on extracting relations between *nominals*. Typically the choice of the entity type depends merely on the final task at hand. Kinds of entities which are usually considered in relation extraction are named entities like persons or organizations. However, we will focus on nominals which are much more general and also include named entities since they are basically nouns or noun phrases (Nastase et al., 2013). Nominals are discussed in more detail in Section 3.2. Unsupervised methods for relation extraction is a particularly interesting area of research because of its applicability across languages without relying on labeled data. In contrast to *open information extraction*, in unsupervised relation extraction the collected relations are aggregated in order to identify the most promising relations for expressing interesting facts. Here, the grouping is made explicit for further processing.

One possible application of relation extraction is the establishment of so-called *knowledge graphs* (Sowa, 2000), which encode facts that manifest solely from text. The knowledge graph can then be used e.g. for reasoning, that is finding new facts from existing facts.

Many approaches exist for acquiring knowledge from text. Hearst (1992) first discovered that relations between entities occur in a handful of well developed text patterns. For example '*X is a Y*' or '*X and other Ys*' manifest themselves as hyponymic relations. However, not every kind of relation is as easy to identify as those '*is-a*' relations. Often semantic relations cannot be expressed by any pattern. A variety of methods were developed that automatically find new patterns and entities with or without supervision. These methods reach from *bootstrapping methods* (Hearst, 1992) over *distant supervision* (Mintz et al., 2009) and *latent relational analysis* (LRA) (Turney, 2005) to *extreme unsupervised relation extraction* (Davidov and Rappoport, 2008a), just to name a few. The importance of unsupervised methods for relation extraction is obvious: The manual creation of knowledge resources is time consuming and expensive in terms of manpower. Though manual resources are typically very precise they are almost always lacking of lexical and relational coverage.

The extraction of relations between entities is a crucial process which is performed by every modern language understanding system like NELL[4] (Carlson et al., 2010) or machine reading[5], which evolved among others from TextRunner[6] (Etzioni et al., 2008). The identification of relations in natural language texts is at the heart of such systems.

3 Methodology

3.1 Focused Crawling

Language models (LMs) are a rather old but well understood and generally accepted concept in Computational Linguistics and Information Retrieval. Our focused crawling strategy builds upon the idea of utilizing a language model to discriminate between relevant and irrelevant web documents. The key idea of this methodology is that web pages which come from a certain domain — which implies the use of a particular vocabulary (Biber, 1995) — link to other documents of the same domain. The assumption is that the crawler will most likely stay in the same topical domain as the initial language model was generated from. Thus the crawling process can be terminated when enough data has been collected.

[4] Never Ending Language Learner: http://rtw.ml.cmu.edu/

[5] http://ai.cs.washington.edu/projects/open-information-extraction

[6] http://openie.cs.washington.edu/

A language model is a statistical model over short sequences of consecutive tokens called N-grams. The order of a language model is defined by the length of such sequences, i.e. the 'N' in N-gram. The probability of a sequence of m words, that could be for example a sentence, is computed as:

$$p(w_1, ..., w_m) \approx \prod_{i=1}^{m} p(w_i | w_{i-N+1:i-1}) \ , \quad (1)$$

where N is the order of the language model and $p(w_i|w_{i-n+1:i-1})$ is the probability of the particular N-gram. In the simplest case the probability of an N-gram is computed as:

$$p(w_i | w_{i-n+1:i-1}) = \frac{count(w_{i-N+1:i})}{count(w_{i-N+1:i-1})} \ , \quad (2)$$

where $count$(N-gram) is a function that takes as argument an N-gram of length N or an N-gram of length $N-1$ and returns the frequency of observations in the source corpus. This model has some obvious limitations when it comes to *out-of-vocabulary* (OOV) terms because of probabilities being zero. Due to this limitation, a number of LMs were proposed which handle OOV terms well.

One of the most advanced language models is the Kneser-Ney language model (Kneser and Ney, 1995), which applies an advanced interpolation technique for OOV issues. According to Halevy et al. (2009), simpler models that are trained on large amounts of data often outperform complex models with training procedures that are feasible only for small data. Anyway, we have only little data in the initial phase, thus we use Kneser and Ney's model.

Perplexity is used to measure the amount of compatibility with another model X:

$$Perplexity(X) = 2^{H(X)} \ , \quad (3)$$

where $H(X) = -\frac{1}{|X|} \sum_{x \in X} \log_2 p(x)$ is the cross entropy of a model X. Using perplexity we are able to tell how well the language model fits the data and vice versa.

The key idea is that documents which come from a certain register or domain — which implies the use of a particular vocabulary (Biber, 1995) — link to other documents of the same register. Using perplexity, we are able to rank outgoing links by their deviation from our initial language model. Hence weblinks that are extracted from a highly deviating webpage are less prioritized for harvesting. The open source crawler software Heritrix[7] (Mohr et al., 2004) forms the basis of our focused crawling strategy, since it provides a well-established framework which is easily extensible through its modularity.

3.2 Identification of Nominals

Nominals are defined to be expressions which syntactically act like nouns or noun phrases (Quirk et al., 1985, p.335). Another definition according to Nastase et al. (2013) is that nominals are defined to be in one of the following classes: *a*) common nouns, *b*) proper nouns, *c*) multi-word proper nouns, *d*) deverbal nouns, *e*) deadjectival nouns, or *f*) non-compositional (adjective) noun phrases. In this work we will follow the definition given by Nastase et al. (2013). We will further address only relations that are at least realized by verbal or prepositional phrases and ignore relations that are implicitly present in compounds, which is a task of its own, cf. (Holz and Biemann, 2008). Note however we do not ignore relations between compounds, but within compounds.

The identification of nominals can be seen as the task of identifying reliable *multi-word-expressions* (MWEs), which is a research question of its own right. As a first simplified approach we only consider nouns and heads of noun compounds to be representatives for nominals. E.g. a compound is used as an entity, but only the head is taken into further consideration as a representative since it encapsulates the main meaning for that phrase.

3.3 Unsupervised Relation Extraction

Our system is founded in the idea of distributional semantics on the level of dependency parses. The *Distributional Hypothesis* by Harris (1954) (cf. also (Miller and Charles, 1991)) states that words which tend to occur in similar contexts tend to have similar meanings. This implies that one can estimate the meaning of an unknown word by considering the context in that it occurs. Lin and Pantel (2001) extended this hypothesis to cover shortest paths in the dependency graph — so-called dependency paths — and introduced the *Extended Distributional Hypothesis*. This extended hypothesis states that dependency paths which tend to occur in similar contexts, i.e. they connect the simi-

[7] `http://crawler.archive.org`

lar sets of words, also tend to have similar meanings.

Sun and Grishman (2010) used an agglomerative hierarchical clustering based approach in order to group the patterns found by Lin and Pantel's method. The clusters are used in a semi-supervised way to extract relation instances that are used in a bootstrapping fashion to find new relations. While Sun and Grishman (2010) performed a hard clustering, meaning every relation is assigned exactly to one cluster, we argue that relations are accompanied by a certain degree of ambiguity. Think for example about the expression '*X comes from Y*' which could be both, a causal relation or a locational relation depending on the meaning of *X* and *Y*.

That being said, we use the Extended Distributional Hypothesis in order to extract meaningful relations from text. We follow Lin and Pantel (2001) and use the dependency path between two entities to identify both, similar entity pairs and similar dependency paths. Specifically we use the Stanford Parser[8] (Klein and Manning, 2003) to get a collapsed dependency graph representation of a sentence, and apply the JoBimText[9] (Biemann and Riedl, 2013) software for computing the distributional similarities.

By using the JoBimText framework, we accept their theory, which states that dimensionality-reduced vector space models are not expressive enough to capture the full semantics of words, phrases, sentences, documents or relations. Turney and Pantel (2010) surveyed that vector space models are commonly used in computational semantics and that they are able to capture the meaning of words. However, by doing various kinds of vector space transformations, e.g. dimensionality reduction with SVD[10] important information from the long tail, i.e. items that do not occur often, is lost. Instead, Biemann and Riedl (2013) introduced the scalable JoBimText framework, which makes use of the Distributional Hypothesis. We take this as a starting point to steer away from the use of vector space models.

For each entity pair 'X::Y', where 'X' and 'Y' are nominals, we collect all dependency paths that co-occur with it in the complete dataset. A particular path for a particular relation instance has form '@1-PATH-@2', where '-PATH-' is the instantiation of the directed shortest path in the collapsed dependency path starting from a particular 'X' and ending in a particular 'Y'. The @1, resp. @2, symbolizes the place where 'X' and 'Y' were found in the path. Here we restrict the path to be shorter than five edges and additionally we ignore paths that have only *nn* relations, i.e. compound dependency relations. See Figure 1 for an illustration of this strategy on two small example sentences. Note that this procedure strongly coheres with the methodologies proposed by Lewis and Steedman (2013) or Akbik et al. (2013).

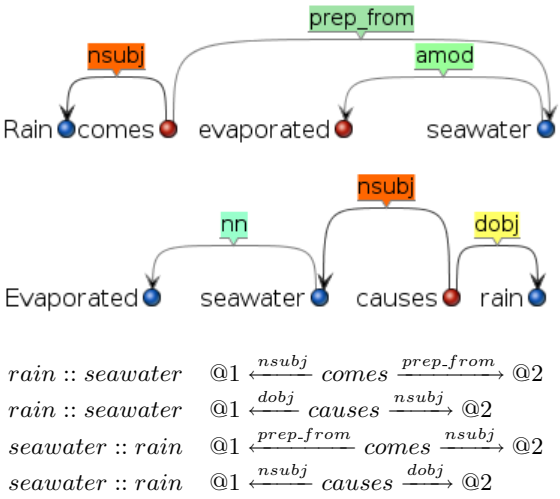

Figure 1: Upper[12]: collapsed dependency parses of the example sentences '*Rain comes from evaporated seawater.*' and *'Evaporated seawater causes rain'*. Lower: extracted entity pairs plus shortest dependency paths per entity pair from both sentences.

We then compute the distributional similarities for both directions: *a*) similarities of entity pairs by paths, and *b*) similarities of paths by entity pairs. This gives us two different views on the data.

4 Evaluation

The two major directions of this paper, i.e. the focused crawling part and the unsupervised relation extraction part are evaluated individually and independent of each other. First we will present an

[8] http://nlp.stanford.edu/downloads/lex-parser.shtml

[9] http://sf.net/p/jobimtext

[10] Singular Value Decomposition, used for example in latent semantic analysis, latent relational analysis, principal component analysis and many more.

[12] Images generated with GrammarScope: http://grammarscope.sf.net .

evaluation methodology to assess the quality of the crawler and second we will outline the evaluation of relations. While we can only show anecdotical evidence of the viability of this approach, since the work is in progress, we are able to present encouraging preliminary results in Section 5.

4.1 Focused Crawling

The quality of a focused crawl is measured in terms of perplexity (cf. Section 3.1) by creating a language model from the harvested data during a particular crawl. Perplexity is then calculated with respect to a held out test set. The following three phases describe the evaluation procedure more precisely:

1. The source corpus is split i.i.d.[13] into a training and test set.

2. We create a language model U of the training data, which is applied according to Section 3.1 for automatically focusing the crawl. In order to compare the data of different crawls, the repeated crawls are initialized with the same global parameter settings, e.g. politeness settings, seed, etc. are the same, and are terminated after reaching a certain number of documents.

3. From the harvested data, another language model V is produced which is used for the evaluation of the test data. Here we argue that a crawl which collects data that is used for evaluating V and V results in a lower perplexity score, is preferred as it better models the target domain.

Figure 2 shows a schematic overview of the three phases of evaluation.

4.2 Unsupervised Relation Extraction

The evaluation of relation extraction is a non-trivial task, as unsupervised categories do usually not exactly match the distinctions taken in annotation studies. For the evaluation of our method we consider the following three approaches:

1. We test our relations directly on datasets that were provided as relation classification challenge datasets (Girju et al., 2007; Hendrickx et al., 2010). Whereas the first dataset is provided as a binary classification task, the second is a multi-way classification task. However, both datasets can be transformed to address the one or the other task. This is possible because the challenge is already finished.

2. We apply our extracted relations for assisting classification algorithms for the task of *textual entailment* (Dagan et al., 2006).

3. Following Davidov and Rappoport (2008b) we would further like to apply our system to the task of question answering.

While the first approach is an intrinsic evaluation, the other three approaches are extrinsic, i.e. the extracted relations are used in a particular task which is then evaluated against some gold standard.

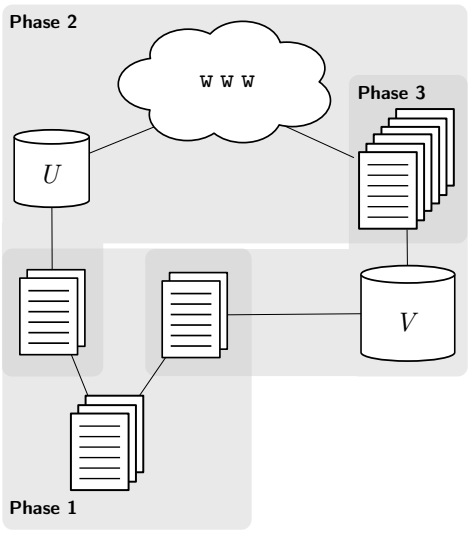

Figure 2: Schematic overview of the evaluation procedure for a particular crawl.

5 Preliminary Results

5.1 Focused crawling

Table 1 shows some quantitative characteristics of a non-focused crawl. Here the crawl was performed as a *scoped crawl*, which means that it was bounded to the German top-level-domain *'.de'* and additionally by a maximum number of 20 hops from the start seed[14]. The crawl was terminated after about two weeks. Although these numbers

[13] independent and identically distributed

[14] The start seed for the first crawl consists of five web page urls which are strongly connected to German educational research.

	pdf	html
size in GBytes	17	400
number of documents	43K	9M
runtime	$\approx$ 2 weeks	

Table 1: Numbers are given as approximate numbers.

do not seem surprising, they do support the main argument of this proposal. Focused crawling is necessary in order to reduce the massive load of irrelevant data.

Initial encouraging results on the comparison of a focused vs. a non-focused crawl are shown in Figure 3. The crawls were performed under the same conditions and we recorded the perplexity value during the process. We plot the history for the first 300,000 documents. Although these results are preliminary, a trend is clearly observable. The focused crawl harvests more relevant documents as it proceeds, whereas the non-focused crawl deviates more as longer the crawl proceeds, as indicated by higher perplexity values for later documents — an effect that is likely to increase as the crawl proceeds. The focused crawl, on the other hand, stays within low perplexity limits. We plan to evaluate settings and the interplay between crawling parameters and language modeling more thoroughly in future evaluations.

5.2 Unsupervised Relation Extraction

The unsupervised extraction of relations was performed on a small subset of one Million sentences of the news corpus from the Leipzig Corpora Collection (Richter et al., 2006).

Preliminary example results are shown in Table 2 and in Table 3. Table 2 shows selected results for similar entity pairs, and Table 3 shows selected results for similar dependency paths.

In Table 2, three example entity pairs are shown together with their most similar counterparts. It is interesting to see that the relation of *gold* to *ounce* is the same as *stock* to *share* or *oil* to *barrel* and we can easily agree here, since the one is the measuring unit for the other.

Table 3 shows for three example prepositional paths the similar paths. We have chosen prepositional phrases here because of their intuitive interpretability. The example output shows that the similar phrases which were identified by the system are also interpretable for humans.

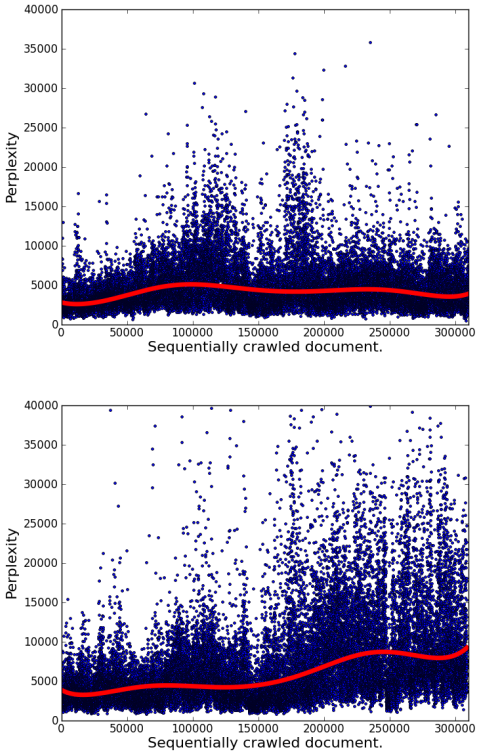

Figure 3: Two crawl runs under same conditions and with same settings. Upper: a focused crawl run. Lower: a non-focused crawl run.

6 Conclusion and Future Work

This research thesis proposal addressed the two major objectives:

1. crawling with a focus on in-domain data by using a language model of an initial corpus, which is small compared to the expected result of the crawls, in order to discriminate relevant web documents from irrelevant web documents, and

2. unsupervised relation extraction by following the principles of the Distributional Hypothesis by Harris (1954) resp. the Extended Distributional Hypothesis by Lin and Pantel (2001).

The promising preliminary results encourage us to examine this approach for further directions. Specifically the yet unaddressed parts of the evaluation will be investigated. Further, the unsupervised relation extraction techniques will be applied on the complete set of in-domain data, thus finalizing the workflow of enriching a small amount of domain defining data with web data

gold/NN :: ounce/NN
$crude/NN :: barrel/NN$
$oil/NN :: barrel/NN$
$futures/NNS :: barrel/NN$
$stock/NN :: share/NN$
graduate/NN :: University/NNP
$graduate/NN :: School/NNP$
$graduate/NN :: College/NNP$
goals/NNS :: season/NN
$points/NNS :: season/NN$
$points/NNS :: game/NN$
$touchdowns/NNS :: season/NN$

Table 2: Example results for selected entity pairs. Similar entity pairs with respect to the boldface pair are shown.

@1 <= prep_above = @2
@1 <= $prep\_below$ = @2
@1 <= $nsubj = rose/VBD = dobj$ => @2
@1 <= $nsubj = dropped/VBD = dobj$ => @2
@1 <= $nsubj = fell/VBD = dobj$ => @2
@1 <= prep_regarding = @2
@1 <= $prep\_about$ = @2
@1 <= $prep\_on$ = @2
@1 <= prep_like = @2
@1 <= $prep\_such\_as$ = @2
@1 <= $prep\_including$ = @2
@1 <= $nsubj = are/VBP = prep\_among$ => @2

Table 3: Example results for selected dependency paths. Similar paths with respect to the boldface path are shown.

from focused crawls in order to extract rich in-domain knowledge, particularly from the german educational domain as our application domain. While we made clear that crawling the web is a crucial process in order to get the amounts of in-domain data needed by the unsupervised relation extraction methods, we did not yet point out that we will also examine the reverse direction, i.e. the possibility to use the extracted relations for further improving the focused crawler. A focused crawler that is powered by semantic relations between entities would raise a new level of semantically focused crawls. Additionally, we will investigate possibilities for further narrowing the relations found by our system. Here it is possible to further categorize or cluster the relations by using either the similarity graph or the features itself, as done by Pantel and Lin (2002).

Acknowledgments

This work has been supported by the German Institute for Educational Research (DIPF) under the KDSL program.

References

Alan Akbik, Larysa Visengeriyeva, Johannes Kirschnick, and Alexander Löser. 2013. Effective selectional restrictions for unsupervised relation extraction. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing (IJCNLP)*, pages 1312–1320, Nagoya, Japan.

Michele Banko and Eric Brill. 2001. Scaling to very very large corpora for natural language disambiguation. In *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics (ACL)*, pages 26–33, Toulouse, France.

Marco Baroni, Silvia Bernardini, Adriano Ferraresi, and Eros Zanchetta. 2009. The wacky wide web: a collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation*, 43(3):209–226.

Douglas Biber. 1995. *Dimensions of Register Variation: A Cross-Linguistic Comparison*. Cambridge University Press, Cambridge.

Chris Biemann and Martin Riedl. 2013. Text: Now in 2d! a framework for lexical expansion with contextual similarity. *Journal of Language Modelling (JLM)*, 1(1):55–95.

Chris Biemann, Felix Bildhauer, Stefan Evert, Dirk Goldhahn, Uwe Quasthoff, Roland Schäfer, Johannes Simon, Swiezinski Swiezinski, and Torsten Zesch. 2013. Scalable construction of high-quality web corpora. *Journal for Language Technology and Computational Linguistics (JLCL)*, 27(2).

Steven Bird, Robert Dale, Bonnie Dorr, Bryan Gibson, Mark Joseph, Min-Yen Kan, Dongwon Lee, Brett Powley, Dragomir Radev, and Yee Fan Tan. 2008. The acl anthology reference corpus: A reference dataset for bibliographic research in computational linguistics. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC)*, Marrakech, Morocco.

Avrim Blum and Tom Mitchell. 1998. Combining labeled and unlabeled data with co-training. In *Proceedings of the Eleventh Annual Conference on Computational Learning Theory (COLT)*, pages 92–100, Madison, Wisconsin, USA.

Thorsten Brants and Alex Franz. 2006. *Web 1T 5-gram Version 1*. Linguistic Data Consortium, Philadelphia.

Andrew Carlson, Justin Betteridge, Bryan Kisiel, Burr Settles, Estevam R. Hruschka Jr., and Tom M. Mitchell. 2010. Toward an architecture for never-ending language learning. In *Proceedings of the 24th Conference on Artificial Intelligence (AAAI)*, Atlanta, GA, USA.

Soumen Chakrabarti, Martin van den Berg, and Byron Dom. 1999. Focused crawling: a new approach to topic-specific web resource discovery. *Computer Networks*, 31(11–16):1623–1640.

Nathanael Chambers and Dan Jurafsky. 2011. Template-based information extraction without the templates. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 976–986, Portland, Oregon, USA.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2006. The pascal recognising textual entailment challenge. In *Machine Learning Challenges. Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Tectual Entailment*, volume 3944 of *Lecture Notes in Computer Science*, pages 177–190. Springer Berlin Heidelberg.

Dmitry Davidov and Ari Rappoport. 2008a. Classification of semantic relationships between nominals using pattern clusters. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-HLT)*, pages 227–235, Columbus, Ohio.

Dmitry Davidov and Ari Rappoport. 2008b. Unsupervised discovery of generic relationships using pattern clusters and its evaluation by automatically generated SAT analogy questions. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-HLT)*, pages 692–700, Columbus, Ohio.

Dmitry Davidov, Ari Rappoport, and Moshe Koppel. 2007. Fully unsupervised discovery of concept-specific relationships by web mining. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics (ACL)*, pages 232–239, Prague, Czech Republic.

Oren Etzioni, Michele Banko, Stephen Soderland, and Daniel S. Weld. 2008. Open information extraction from the web. *Communications of the ACM*, 51(12):68–74.

Roxana Girju, Preslav Nakov, Vivi Nastase, Stan Szpakowicz, Peter Turney, and Deniz Yuret. 2007. Semeval-2007 task 04: Classification of semantic relations between nominals. In *Proceedings of the Fourth International Workshop on Semantic Evaluation (SemEval)*, pages 13–18, Prague, Czech Republic.

Alon Halevy, Peter Norvig, and Fernando Pereira. 2009. The unreasonable effectiveness of data. *IEEE Intelligent Systems*, 24(2):8–12.

Zellig S. Harris. 1954. Distributional structure. *Word*, 10(23):146–162.

Marti A. Hearst. 1992. Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the 15th Conference on Computational Linguistics (Coling)*, pages 539–545, Nantes, France.

Iris Hendrickx, Su Nam Kim, Zornitsa Kozareva, Preslav Nakov, Diarmuid Ó Séaghdha, Sebastian Padó, Marco Pennacchiotti, Lorenza Romano, and Stan Szpakowicz. 2010. Semeval-2010 task 8: Multi-way classification of semantic relations between pairs of nominals. In *Proceedings of the Fifth International Workshop on Semantic Evaluation (SemEval)*, pages 33–38, Los Angeles, California.

Florian Holz and Chris Biemann. 2008. Unsupervised and knowledge-free learning of compound splits and periphrases. In *CICLing 2008: Proceedings of the Conference on Intelligent Text Processing and Computational Linguistics*, pages 117–127, Haifa, Israel.

Adam Kilgarriff. 2007. Googleology is bad science. *Computational Linguististics (CL)*, 33(1):147–151.

Dan Klein and Christopher D. Manning. 2003. Accurate unlexicalized parsing. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics (ACL)*, pages 423–430, Sapporo, Japan.

Reinhard Kneser and Hermann Ney. 1995. Improved backing-off for m-gram language modeling. In *In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 181–184, Detroit, Michigan.

Mike Lewis and Mark Steedman. 2013. Unsupervised induction of cross-lingual semantic relations. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 681–692, Seattle, WA, USA.

Dekang Lin and Patrick Pantel. 2001. Dirt - discovery of inference rules from text. In *Proceedings of the Seventh International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 323–328, San Francisco, California.

Dekang Lin, Kenneth Church, Heng Ji, Satoshi Sekine, David Yarowsky, Shane Bergsma, Kailash Patil, Emily Pitler, Rachel Lathbury, Vikram Rao, Kapil Dalwani, and Sushant Narsale. 2010. New tools for web-scale n-grams. In *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC)*, pages 2221–2227, Valletta, Malta.

Filippo Menczer, Gautam Pant, and Padmini Srinivasan. 2004. Topical web crawlers: Evaluating adaptive algorithms. *ACM Transactions Internet Technology (TOIT)*, 4(4):378–419.

George A Miller and Walter G Charles. 1991. Contextual correlates of semantic similarity. *Language and Cognitive Processes (LCP)*, 6(1):1–28.

Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. 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*, pages 1003–1011, Suntec, Singapore.

Gordon Mohr, Michele Kimpton, Micheal Stack, and Igor Ranitovic. 2004. Introduction to heritrix, an archival quality web crawler. In *Proceedings of the 4th International Web Archiving Workshop IWAW'04*, Bath, UK.

Vivi Nastase, Preslav Nakov, Diarmuid Ó Séaghdha, and Stan Szpakowicz. 2013. Semantic relations between nominals. In *Synthesis Lectures on Human Language Technologies*, volume 6. Morgan & Caypool Publishers.

Patrick Pantel and Dekang Lin. 2002. Document clustering with committees. In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 199–206.

Emily Pitler, Shane Bergsma, Dekang Lin, and Kenneth Church. 2010. Using web-scale n-grams to improve base np parsing performance. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling)*, pages 886–894, Beijing, China.

Randolph Quirk, Sidney Greenbaum, Geoffrey Leech, and Jan Svartvik. 1985. *A Comprehensive Grammar of the English Language*. Longman, London.

Matthias Richter, Uwe Quasthoff, Erla Hallsteinsdóttir, and Chris Biemann. 2006. Exploiting the leipzig corpora collection. In *Proceesings of the IS-LTC*, Ljubljana, Slovenia.

Roland Schäfer and Felix Bildhauer. 2012. Building large corpora from the web using a new efficient tool chain. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC)*, pages 486–493, Istanbul, Turkey.

John Sowa. 2000. *Knowledge Representation: Logical, Philosophical and Computational Foundations*. Brooks Cole Publishing Co., Pacific Grove, CA.

Ang Sun and Ralph Grishman. 2010. Semi-supervised semantic pattern discovery with guidance from unsupervised pattern clusters. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling)*, pages 1194–1202, Beijing, China.

Peter D. Turney and Patrick Pantel. 2010. From frequency to meaning: Vector space models of semantics. *Journal for Artificial Intelligence Research (JAIR)*, 37:141–188.

Peter D. Turney. 2005. Measuring semantic similarity by latent relational analysis. In *Proceedings of the 19th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 1136–1141, Edinburgh, Scotland, UK.