TAXI at SemEval-2016 Task 13: a Taxonomy Induction Method based on Lexico-Syntactic Patterns, Substrings and Focused Crawling

Alexander Panchenko[1], Stefano Faralli[2], Eugen Ruppert[1], Steffen Remus[1], Hubert Naets[3], Cédrick Fairon[3], Simone Paolo Ponzetto[2] and Chris Biemann[1]

[1]TU Darmstadt, LT Group, Germany
[2]Mannheim University, Web and Data Science, Germany
[3]UCLouvain, CENTAL, Belgium
panchenko@lt.informatik.tu-darmstadt.de

Abstract

We present a system for taxonomy construction that reached the first place in all subtasks of the SemEval 2016 challenge on Taxonomy Extraction Evaluation. Our simple yet effective approach harvests hypernyms with substring inclusion and Hearst-style lexico-syntactic patterns from domain-specific texts obtained via language model based focused crawling. Extracted taxonomies are evaluated on English, Dutch, French and Italian for three domains each (Food, Environment and Science). Evaluations against a gold standard and by human judgment show that our method outperforms more complex and knowledge-rich approaches on most domains and languages. Furthermore, to adapt the method to a new domain or language, only a small amount of manual labour is needed.

1 Introduction

In this paper, we describe TAXI – a taxonomy induction method first presented at the SemEval 2016 challenge on Taxonomy Extraction Evaluation (Bordea et al., 2016). We consider taxonomy induction as a process that should – as much as possible – be driven solely on the basis of raw text processing. While some labeled examples might be utilized to tune the extraction and induction process, we avoid relying on structured lexical resources such as WordNet (Miller, 1995) or BabelNet (Navigli and Ponzetto, 2010). We rather envision a situation where a taxonomy shall be induced in a new domain or a new language for which such resources do not exist. In this paper, we demonstrate our methodology based on hyponym extraction from substrings and general-domain and domain-specific corpora for four languages and three domains.

2 Related Work

The extraction of taxonomic relationships from text is a long-standing challenge in ontology learning, see e.g. Biemann (2005) for a survey. The literature on hypernym extraction offers a high variability of methods, from simple lexical patterns (Hearst, 1992; Oakes, 2005), similar to those used in our method, to complex statistical techniques (Agirre et al., 2000; Ritter et al., 2009).

Snow et al. (2004) use sentences that contain two terms which are known to be hypernyms. They parse sentences and extract patterns from the parse trees. Finally, they train a hypernym classifier based on these features and applied to text corpora.

Yang and Callan (2009) presented a semi-supervised taxonomy induction framework that integrates co-occurrence, syntactic dependencies, lexical-syntactic patterns and other features to learn an ontology metric, calculated in terms of the semantic distance for each pair of terms in a taxonomy. Terms are incrementally clustered on the basis of their ontology metric scores.

Snow et al. (2006) perform incremental construction of taxonomies using a probabilistic model. They combine evidence from multiple supervised classifiers trained on large training datasets of hyponymy and co-hyponymy relations. The taxonomy learning task is defined as the problem of finding the taxonomy that maximizes the probability of individ-

ual relations extracted by the classifiers.

Kozareva and Hovy (2010) start from a set of root terms and use Hearst-like lexico-syntactic patterns to harvest hypernyms from the Web. The extracted hypernym relation graph is subsequently pruned.

Veraldi et al. (2013) proposed a graph-based algorithm to learn a taxonomy from textual definitions, extracted from a corpus and the Web. An optimal branching algorithm is used to induce a taxonomy.

Finally, Bordea et al. (2015) introduced the first shared task on Taxonomy Extraction Evaluation to provide a common ground for evaluation. Six systems participated in the competition. The top system in this challange used features based on substrings and co-occurrence statistics (Grefenstette, 2015). Lefever et al. (2015) reached the second place gathered hypernyms from patterns, substrings and WordNet. Tan et al. (2015) used word embeddings, reaching the third place.

3 Taxonomy Induction Method

Our approach is characterized by scalability and simplicity, assuming that being able to process larger input data is more important than the sophisticated extraction inference. Our approach to taxonomy induction takes as input a set of domain terms and general-domain text corpora and outputs a taxonomy. It consists of four steps. Firstly, we crawl domain-specific corpora based on terminology of the target domain (see Section 3.1). These complement general-purpose corpora, like texts of Wikipedia articles. Secondly, candidate hypernyms are extracted based on substrings and lexico-syntactic patterns (see Section 3.2). Thirdly, the candidates are pruned so that each term has only a few most salient hypernyms (see Section 3.3). The last step performs optimization of the overall taxonomy structure removing cycles and linking disconnected components to the root (see Section 3.4).

3.1 Corpora for Taxonomy Induction

To build domain-specific taxonomies we use both general and domain-specific corpora.

General Domain Corpora. We use three general purpose corpora in our approach presented in Table 1: Wikipedia, 59G and CommonCrawl[1].

[1] https://commoncrawl.org

	EN	FR	NL	IT
Wikipedia	11.0	3.2	1.4	3.0
59G	59.2	–	–	–
CommonCrawl	168000.0 ‡	–	–	–
FocusedCrawl Food	22.8	7.9	3.4	3.6
FocusedCrawl Environment	23.9	8.9	2.0	7.1
FocusedCrawl Science	8.8	5.4	6.6	5.1

Table 1: Corpora sizes used in our system in GB, where ‡is the size of the crawl archive.

The second corpus is a concatenation of the English Wikipedia, Gigaword (Parker et al., 2009), ukWaC (Ferraresi et al., 2008) and a news corpora from the Leipzig Collection (Goldhahn et al., 2012).

Domain-Specific Corpora. Lefever (2015) showed the usefulness for taxonomy extraction of domain dependent corpora crawled from the Web using *BootCat* (Baroni and Bernardini, 2004). This method takes terms as input, which are randomly combined into sequences of a pre-defined length, and sent to a Web search engine. The search results, i.e. the returned URLs, compose a domain-dependent corpus. The number of input terms, the number of queries and the amount of desired URLs impact the size of the corpus. With 1,000 web queries and 10 URLs per query, the expected size of the resulting corpus is around 300 MB. While Lefever (2015) shows that such small in-domain corpora can be already useful for taxonomy extraction, we assumed that better results can be obtained if bigger domain-specific corpora are used.

We therefore follow a different approach based on *focused crawling*, where *BootCat* is used only for initialization of seed URLs. We use the provided taxonomy terms as input for the *BootCat* method, generate 1,000 random triples, and use the retrieved URLs as a starting point for further crawling. Focused crawling is an extension to standard web crawling where URLs, expected to point to relevant web documents, are prioritized for download (Chakrabarti et al., 1999).

Remus and Biemann (2016) introduced a focused crawling approach based on language modeling. The idea is that relevant web documents refer to other relevant web documents, where the relevance of a web document is computed by considering a statistical n-gram language model of a small, initially provided, domain-defining corpus. We provide a domain-defining corpus for each category by

using Wikipedia articles, that are directly contained in the matching Wikipedia category. For example, for the the Food domain we used the Wikipedia articles of *Category:Foods* to build a language model of the Food domain. The language models for each domain were created using the 5-gram with the Kneser-Ney (1995) smoothing.

Using this technique, we are able to iteratively follow promising URLs and download web pages until a specified stopping criterion (no more pages with desired perplexity or timeout). Each domain and language was crawled for about one week on a single server machine with 24 cores and 32GB RAM, and harvested between 130 and 800 GB raw content, which results in 2 to 23 GB of unique plain-text sentences (c.f. Table 1). Note, that these sentences might contain cross-domain content.

3.2 Candidate Hypernyms via Substrings

A simple yet precise method for hypernym extraction is based on substring matching, c.f. the baseline system in Table 3 and (Lefever, 2015). For instance, "biomedical science" is a "science", "microbiology" is a "biology" and so on. We calculate the following substring-based hypernymy score $\sigma(t_i, t_j)$ between a pair of candidate terms t_i, t_j:

$$\sigma(t_i, t_j) = \begin{cases} \frac{length(t_j)}{length(t_i)} & \text{if } m(t_i, t_j) \wedge \neg m(t_j, t_i) \\ 0 & \text{otherwise} \end{cases}$$

Here $m(t_i, t_j)$ is a function that returns true in case of a match of the term t_i inside the term t_j. Such match happens if $length(t_i)$ is greater than 3. For English and Dutch, the hypernym t_i should match in the end of hyponym t_j, e.g. "natural science" is a "science". For French and Italian a match of hypernym should be in the beginning of hyponym e.g. "algèbre linéaire" is a "algèbre", not "linéaire". The same holds for English and Dutch if a hyponym contains a preposition e.g.: "toast with bacon" is a "toast", not "bacon" or "brood van gekiemd graan" is "brood", not "graan". Finally, if no match is found, we lemmatize terms t_i and t_j and retry matching. The precision-recall curve of the substring score calculated on the trial dataset is presented in Figure 1. As one can observe, precision of the substring score is constantly high, reaching 0.91 at the recall level of 0.29 with AUC of 0.61. Therefore, this score is able to retrieve a significant number of high-quality hypernyms. Yet, only hypernyms of compound words can be retrieved via substrings.

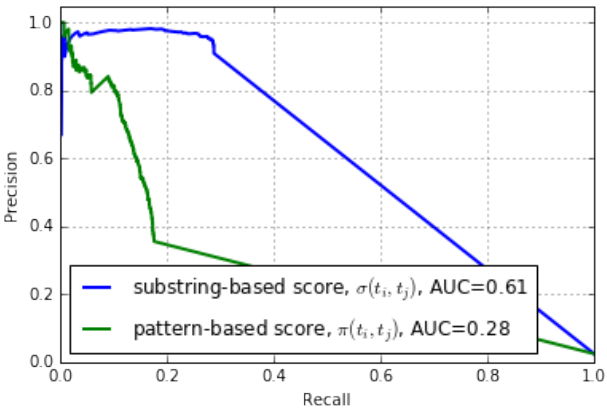

Figure 1: Precision-Recall plots of substring-based and pattern-based features of the TAXI approach measured on the trial dataset.

3.3 Candidate Hypernyms via Patterns

To extract candidate hypernym relations from texts we used three systems listed below. All of them rely on lexico-syntactic patterns in the fashion of (Hearst, 1992; Klaussner and Zhekova, 2011). We used several systems to filter noise via complementary signals. Besides, not all the systems support all the four languages of the SemEval task. Porting of Hearst patterns to a new European language is a straightforward and relatively quick procedure. Yet, due to a dense SemEval schedule, we decided to implement new rules only for two languages not supported by any available system, namely Italian and Dutch and reuse extraction rules for other languages.

PattaMaika. This system was used to process English, Italian and Dutch corpora. It implements patterns using UIMA Ruta (Kluegl et al., 2014). First, part-of-speech information is used to assign noun phrase (NP) chunk annotations to nominal phrases. Next, we use patterns to identify hypernym relations between NP chunks. We adapted the 9 English rules to the target languages, resulting in 9 patterns for Italian and 8 patterns for Dutch.

PatternSim. This system was used to process English and French corpora. It encodes patterns in the form of finite state transducers implemented

with the Unitex corpus processor.[2] PatternSim relies on 10 English patterns yielding average precision of top 5 extracted semantic relations per word of 0.69 (Panchenko et al., 2012). For French, 9 hypernym extraction patterns are used providing precision at top 5 of 0.63 (Panchenko et al., 2013).

WebISA. In addition to *PattaMaika* and *PatternSim*, we used a publicly available database of English hypernym relations extracted from the CommonCrawl corpus (Seitner et al., 2016). We used 108 million hypernym relations with frequency above one. This collection of relations was harvested using a regexp-based implementation of 59 patterns collected from the literature.

Combination of hypernyms. Result of the extraction are 18 collections of hypernym relations listed in Table 2. Even the huge *WebISA* collection extracted from tens of terabytes of text does not provide hypernyms for all rare taxonomic terms, such as "ground and whole bean coffee" and "black sesame rice cake". On the other hand, most of the collections contain many noisy relations. For instance, frequent relations for hypernyms often go in both directions, e.g. "history" is a "science", but also "science" is a "history". Therefore, we introduced an asymmetric pattern-based hypernymy score $\pi(t_i, t_j)$ between terms t_i and t_j. It combines information from different hypernym collections to filter noisy extractions. To compute the score, first we normalize extraction counts on the per word basis: $\pi^k(t_i, t_j) = \frac{freq^k(t_i,t_j)}{\max_j freq^k(t_i,t_j)}$, where $freq^k(t_i, t_j)$ is the number of relations extracted between terms t_i and t_j by the k-th extractor. These normalized scores are averaged across all extractors per language-domain pair: $\bar{\pi}(t_i, t_j) = \frac{1}{|LD|}\sum_{k \in LD} \pi^k(t_i, t_j)$, where LD is a set of hypernym collections relevant for a given language-domain pair. For instance, for the language-domain pair "English-Food", the LD contains four collections: general relations extracted by *PatternSim*, *PattaMaika* and *WebISA* plus domain-specific relations extracted by *PatternSim* (see Table 2). Finally, to get the pattern-based score, we subtract averaged scores of two terms in both directions: $\pi(t_i, t_j) = \bar{\pi}(t_i, t_j) - \bar{\pi}(t_j, t_i)$. This way, we downrank symmetrical relations like synonyms and co-hyponyms.

[2] http://www-igm.univ-mlv.fr/~unitex

	EN ‡, †,§	FR‡	NL†	IT†
General	27.6‡, 4.9†, 118.9§	3.2	2.22	0.13
Food	24.1‡	3.8	0.47	0.05
Environment	26.3‡	4.5	0.32	0.95
Science	9.3‡	2.7	0.97	0.05

Table 2: Number of hypernyms in millions of relations. Systems used to extract respective hypernym collections are denoted with ‡for *PatternSim*, †for *PattaMaika* and § for *WebISA*.

The precision-recall curve of the pattern-based score on the trial data is presented in Figure 1. This plot is calculated on general corpora as we did not crawl domain specific corpora for the trial dataset domains. As one can see, precision of 0.80 is achieved at recall of 0.15 or less and drops to 0.36 at recall of 0.19. AUC of 0.28 of is less than half of the substring-based score of 0.61. Thus, patterns are a less reliable source of hypernyms than the substrings. Yet, they can capture relations between words with different spelling like "apple" and "fruit", while the substring-based score need a character overlap, like in "grapefruit" and "fruit".

3.4 Pruning of Hypernyms

Patterns and substrings together yield up to several hundreds of hypernym candidates per term. This step prunes hypernym candidates, ranking them with an unsupervised and supervised combinations of the $\sigma(t_i, t_j)$ and $\pi(t_i, t_j)$ scores.

Unsupervised Pruning. In this pruning strategy, used for French, Dutch and Italian languages, a term t_i is a hypernym of term t_j if their substring score $\sigma(t_i, t_j)$ is greater than zero or if rank of the term t_i according to the pattern-based score $\pi(t_i, t_j)$ equals to one or two. Thus a term t_i obtains all hypernyms extracted by substrings and up to two hypernyms extracted by patterns.

Supervised Pruning. This pruning strategy used for English relies on a supervised classifier trained on the trial dataset (Bordea et al., 2016). [3] This pruning approach uses 3,249 hypernymy relations from the trial taxonomies as positive training samples, e.g. hypernym relation ("biology", "science") and 128,183 automatically generated relations as negatives samples coming from two sources: 3,249 in-

[3] We did not submit to SemEval both supervised and unsupervised versions of the system for English as only one run was allowed per language-domain pair.

verted hypernyms, such as ("science", "biology") plus 124,934 co-hyponyms from the trial taxonomy, for instance ("biology", "mathematics").

The classifier used in the competition had two features that characterize a word pair (t_i, t_j), namely substring-, and pattern-based scores $\sigma(t_i, t_j)$ and $\pi(t_i, t_j)$. Note, that the same features were used in the unsupervised approach. We applied an SVM classifier with RBF kernel (Vert et al., 2004), tuning kernel meta-parameters within an internal loop cross-validation procedure.

We tested multiple alternative configurations with extra features, including term frequency, out/in degree of terms in the hypernym graph, term length, expansions of hypernyms based on term clustering and shortest paths in the graph of candidate hypernyms as well as other classifiers including Logistic Regression, Gradient Boosted Trees, and Random Forest. However, none of the above mentioned configurations yielded consistently better results on the trial data than the two feature-based SVM.

To identify hypernyms among a set of terms T, we classify using a model trained on the trial data all possible word pairs except identical ones: $\{(t_i, t_j) : i \neq j; (t_i, t_j) \in T \times T\}$. The pairs classified using the positive class are added to the taxonomy.

3.5 Taxonomy Construction

At this point of the taxonomy construction, we obtained a noisy graph, which may contain cycles and disconnected components. To remove cycles and to obtain a directed acyclic graph taxonomy we used the unsupervised graph pruning approach of Faralli et al. (2015), which searches for a cycle C using topological sorting of Tarjan (1972) and then removes a random edge of C until no cycles are detected in the graph.

Besides, to improve connectivity of the taxonomy we connect all the nodes of each disconnected component with zero out degree to the taxonomy root.

4 Evaluation

To assess quality of the taxonomies several complementary measures were used. The first type of measurements are structural measures, such as the number of connected components (c.c.), the number of intermediate nodes (i.i.), i.e. the number of nodes with out degree equal to zero, and the presence of cycles. Second, system outputs were compared against the corresponding domain gold standards and performances are evaluated in terms of F-score. Here precision and recall are based on the number of edges in common with the gold standard taxonomy over the number of system edges and over the number of gold standard edges respectively. To better compare against gold standard taxonomies the task included the evaluation of a cumulative measure (Velardi et al., 2013), namely Cumulative Fowlkes & Mallows Measure (F&M), where the similarity between the system and the reference taxonomies are measured as the combination of the hierarchical cluster similarities. Finally, the organizers performed manual quality assessment to estimate the precision of the hypernyms. To compute this measure, annotators labeled a sample of 100 hypernym relations as correct or wrong. The taxonomy extraction was evaluated on four languages, namely English, Dutch, French and Italian, and three different domains (Food, Science and Environment). A detailed description of the evaluation settings and metrics can be found in (Bordea et al., 2016).

5 Results

Table 3 presents a summary of evaluation of our method on the SemEval 2016 Task 13 dataset. Overall 5 systems participated in the challenge: JUNLP, TAXI, NUIG-UNLP, USAAR and QASITT. We represent the respective best scores across our four competitors in the *BestComp* column.

Gold Standard Comparison. The organizer-provided *Baseline* system implemented a string inclusion approach that covers relations between compound terms. A similar mechanism was used by the USAAR system (Tan et al., 2016), which improved over the baseline in terms of precision at the cost of recall. USAAR achieved the highest precision scores for English, as they used substring-based methods that yield high precision (c.f. Figure 1). Yet substrings cannot retrieve hypernyms of non-compound terms.

The main mechanisms we added in TAXI with regard to the substring-based methods are statistics over pattern-based extractions over large domain specific corpora and our taxonomy construction step

	Monolingual (EN)			Multilingual (NL, FR, IT)		
Measure	**Baseline**	**BestComp**	**TAXI**	**Baseline**	**BestComp**	**TAXI**
Cyclicity	*0*	**0**	**0**	*0*	**0**	**0**
Structure (F&M)	*0.005*	**0.406**	0.291	*0.009*	0.016	**0.189**
Categorisation (i.i.)	*77.67*	**377.00**	104.50	*64.28*	**178.22**	64.94
Connectivity (c.c.)	*36.83*	44.75	**1 .00**	*40.50*	34.89	**1.00**
Gold standard comparison (Fscore)	*0.330*	0.260	**0.320**	*0.009*	0.016	**0.189**
Manual Evaluation (Precision)	*n.a.*	**0.490**	0.200	*n.a.*	0.298	**0.625**

Table 3: Overall scores obtained by averaging the results over domains (Environment, Science, Food) and languages (NL, FR, IT) for the multilingual setting. The *BestComp* lists the respective best scores across four our competitors. The best scores excluding the baseline are set in boldface. Definitions of the measures are available in Section 4.

that improves structure of the resource. These united mechanisms are not used in other submissions to the challenge. The NUIG-UNLP team (Pocostales, 2016) relies on vector directionality in dense word embedding spaces. Such approximation of patterns based on distributional similarity provided good recall, but attained low precision.

The QASSIT team (Cleuziou and Moreno, 2016), who ranked second in the competition, uses patterns to extract hypernym candidates, but they rely solely on the Wikipedia. Subsequently, an optimization technique based on genetic algorithms is used to learn the parametrization of a so-called pretopological space, which leads to desired structural properties of the resulting taxonomy. While we use simpler optimization procedure based on supervised learning, TAXI outperforms QASSIT in terms of comparisons with the gold standard. Possible reasons why our method performs better are (1) QASSIT use no substring features, (2) this team relies on smaller general-purpose corpora, while we use larger domain-specific corpora.

Finally, JUNLP relies on substrings and relations extracted from BabelNet. We find the latter to be undesirable for taxonomy extraction. Indeed, a rich lexical resource, such as BabelNet can be considered as a taxonomy in itself. Interestingly, even with the BabelNet-based features the system did not always reach the top precision and recall.

Manual Evaluation. Our system was ranked first in terms of manual judgments for the Dutch, Italan and French reaching the average precision across languages and domains of 0.625. Precision for different language-domain pairs ranged from 0.90 for the Italian-Science pair to 0.23 for the French-Environment pair. For English, our system was ranked second with the average score of 0.20, while the substring-based USAAR system obtained the score of 0.49 and the third-ranked system obtained the score of 0.09. We attribute lower precision in the English run to absence in the supervised ranking scheme of a limit on the number of extracted hypernyms per word.

Further detailed comparisons with other systems with breakdowns with regard to different languages, domains and evaluation schemes are presented by Bordea et al. (2016) and on SemEval website.[4]

Discussion. One shortcoming of our method is its coverage: for instance 774 terms of 1,555 of the English food domain are still not attached to any node, which is a typical issue with pattern-based approaches since not all taxonomic relationships are spelled out explicitly in corpora. To tackle this shortcoming, we plan to use hypernym expansion based on distributional semantics (Biemann and Riedl, 2013).

6 Conclusion

We presented a technique for taxonomy induction from a domain vocabulary. It extracts hypernyms from substrings and large domain-specific corpora bootstrapped from the input vocabulary. Multiple evaluations based on the SemEval taxonomy extraction datasets of four languages and three domains show state-of-the-art performance of our approach. An implementation of our method featuring all language resources, is available for download.[5]

Acknowledgments

We acknowledge the support of the Deutsche Forschungsgemeinschaft under the JOIN-T project.

[4] http://alt.qcri.org/semeval2016/task13

[5] http://tudarmstadt-lt.github.io/taxi

References

Eneko Agirre, Olatz Ansa, Eduard H. Hovy, and David Martínez. 2000. Enriching very large ontologies using the WWW. In *ECAI Workshop on Ontology Learning*, pages 28–33, Berlin, Germany.

Marco Baroni and Silvia Bernardini. 2004. BootCaT: Bootstrapping corpora and terms from the web. In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC)*, pages 1313–1316, Lisbon, Portugal.

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

Chris Biemann. 2005. Ontology Learning from Text: A Survey of Methods. *LDV Forum*, 20(2):75–93.

Georgeta Bordea, Paul Buitelaar, Stefano Faralli, and Roberto Navigli. 2015. Semeval-2015 task 17: Taxonomy extraction evaluation (TExEval). *SemEval-2015*, 452(465):902–910.

Georgeta Bordea, Els Lefever, and Paul Buitelaar. 2016. Semeval-2016 task 13: Taxonomy extraction evaluation (TExEval-2). In *Proceedings of the 10th International Workshop on Semantic Evaluation*, San Diego, CA, 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.

Guillaume Cleuziou and Jose G. Moreno. 2016. QASSIT at SemEval-2016 Task 13: On the integration of Semantic Vectors in Pretopological Spaces for Lexical Taxonomy Acquisition. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval 2016)*. Association for Computational Linguistics, June.

Stefano Faralli, Giovanni Stilo, and Paola Velardi. 2015. Large scale homophily analysis in twitter using a twixonomy. In *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence, IJCAI 2015*, pages 2334–2340, Buenos Aires, Argentina.

Adriano Ferraresi, Eros Zanchetta, Marco Baroni, and Silvia Bernardini. 2008. Introducing and evaluating ukWaC, a very large web-derived corpus of English. In *Proceedings of the 4th Web as Corpus Workshop (WAC-4) Can we beat Google*, pages 47–54, Marakech, Morocco.

Dirk Goldhahn, Thomas Eckart, and Uwe Quasthoff. 2012. Building Large Monolingual Dictionaries at the Leipzig Corpora Collection: From 100 to 200 Languages. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, pages 759–765, Istanbul, Turkey.

Gregory Grefenstette. 2015. INRIASAC: Simple hypernym extraction methods. *SemEval-2015*, pages 911–914.

Marti A Hearst. 1992. Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the 14th conference on Computational linguistics-Volume 2*, pages 539–545.

Carmen Klaussner and Desislava Zhekova. 2011. Lexico-syntactic patterns for automatic ontology building. In *Student Research Workshop at RANLP 2011*, pages 109–114.

Peter Kluegl, Martin Toepfer, Philip-Daniel Beck, Georg Fette, and Frank Puppe. 2014. UIMA Ruta: Rapid development of rule-based information extraction applications. *Natural Language Engineering*, 22(1):1–40.

Reinhard Kneser and Hermann Ney. 1995. Improved backing-off for m-gram language modeling. In *Acoustics, Speech, and Signal Processing, 1995. ICASSP-95., 1995 International Conference on*, volume 1, pages 181–184. IEEE.

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

Els Lefever. 2015. LT3: a multi-modular approach to automatic taxonomy construction. *SemEval-2015*, pages 944–948.

George A Miller. 1995. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

Roberto Navigli and Simone Paolo Ponzetto. 2010. BabelNet: building a very large multilingual semantic network. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 216–225, Uppsala, Sweden.

Michael P. Oakes. 2005. Using Hearst's rules for the automatic acquisition of hyponyms for mining a pharmaceutical corpus. In *RANLP Text Mining Workshop'05*, pages 63–67, Borovets, Bulgaria.

Alexander Panchenko, Olga Morozova, and Hubert Naets. 2012. A semantic similarity measure based on lexico-syntactic patterns. In Jeremy Jancsary, editor, *Proceedings of KONVENS 2012*, pages 174–178, Vienna, Austria. ÖGAI. Main track: poster presentations.

Alexander Panchenko, Hubert Naets, Laetitia Brouwers, Pavel Romanov, and Cédrick Fairon. 2013. Recherche et visualisation de mots sémantiquement liés. In *Proceedings of the TALN-RÉCITAL 2013*, pages 747–754, Les Sables d'Olonne, France.

Robert Parker, David Graff, Junbo Kong, Ke Chen, and Kazuaki Maeda. 2009. *English gigaword fourth edition*. Linguistic Data Consortium, Philadelphia, USA.

Joel Pocostales. 2016. NUIG-UNLP at SemEval-2016 Task 13: A Simple Word Embedding-based Approach for Taxonomy Extraction. In *Proceedings of the 10th International Workshop on Semantic Evaluation*. Association for Computational Linguistics.

Steffen Remus and Chris Biemann. 2016. Domain-specific corpus expansion with focused webcrawling. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC)*, Portorož, Slovenia.

Alan Ritter, Stephen Soderland, and Oren Etzioni. 2009. What is this, anyway: Automatic hypernym discovery. In *Proceedings of the 2009 AAAI Spring Symposium on Learning by Reading and Learning to Read*, pages 88–93, Palo Alto, California.

Julian Seitner, Christian Bizer, Kai Eckert, Stefano Faralli, Robert Meusel, Heiko Paulheim, and Simone Ponzetto. 2016. A Large DataBase of Hypernymy Relations Extracted from the Web. In *Proceedings of the 10th edition of the Language Resources and Evaluation Conference*, Portorož, Slovenia.

Rion Snow, Daniel Jurafsky, and Andrew Y. Ng. 2004. Learning syntactic patterns for automatic hypernym discovery. In Lawrence K. Saul, Yair Weiss, and Léon Bottou, editors, *Proc. of NIPS 2004*, pages 1297–1304, Cambridge, Mass. MIT Press.

Rion Snow, Dan Jurafsky, and Andrew Ng. 2006. Semantic taxonomy induction from heterogeneous evidence. In *Proceedings of COLING-ACL 2006*, pages 801–808, Sydney, Australia.

Liling Tan, Rohit Gupta, and Josef van Genabith. 2015. USAAR-WLV: hypernym generation with deep neural nets. *SemEval-2015*, pages 932–937.

Liling Tan, Francis Bond, and Josef van Genabith. 2016. USAAR at SemEval-2016 Task 13: Hyponym Endocentricity. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval 2016)*. Association for Computational Linguistics.

Robert Tarjan. 1972. Depth-first search and linear graph algorithms. *SIAM Journal on Computing*, 1:146–160.

Paola Velardi, Stefano Faralli, and Roberto Navigli. 2013. Ontolearn reloaded: A graph-based algorithm for taxonomy induction. *Computational Linguistics*, 39(3):665–707.

Jean-Philippe Vert, Koji Tsuda, and Bernhard Schölkopf. 2004. A primer on kernel methods. In Bernhard Schlkopf, Koji Tsuda, and Jean-Philippe Vert, editors, *Kernel Methods in Computational Biology*, pages 35–70. Cambridge, MA: MIT Press.

Hui Yang and Jamie Callan. 2009. A metric-based framework for automatic taxonomy induction. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the Association for Computational Linguistics (ACL) and the 4th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing (IJCNLP)*, pages 271–279, Suntec, Singapore.