Extracting Ontology Concept Hierarchies from Text using Markov Logic

Lucas Drumond
Federal University of Maranhão
Av. dos Portugueses, s/n 65085-580
São Luís - MA, Brazil
ldrumond@gmail.com

Rosario Girardi
Federal University of Maranhão
Av. dos Portugueses, s/n 65085-580
São Luís - MA, Brazil
rosariogirardi@gmail.com

ABSTRACT

Ontologies have proven to be a powerful tool for many tasks such as natural language processing and information filtering and retrieval. However their development is an error prone and expensive task. One approach for this problem is to provide automatic or semi-automatic support for ontology construction. This work presents the Probabilistic Relational Hierarchy Extraction (PREHE) technique, an approach for extracting concept hierarchies from text that uses statistical relational learning and natural language processing for combining cues from many state-of-the-art techniques. A Markov Logic Network has been developed for this task and is described here. A preliminary evaluation of the proposed approach is also outlined.

Categories and Subject Descriptors

I.2.6 [**Learning**]: Knowledge acquisition

General Terms

Theory

Keywords

Ontology Learning, Markov Logic Networks, Knowledge Acquisition

1. INTRODUCTION

Knowledge has become an important factor for the success of organizations. As well as the advent of relational databases allowed computers to deal with information more effectively, thus giving rise to the so called Information Systems, the usage of knowledge representation formalisms for constructing knowledge bases has enabled the development of the Knowledge Systems.

Traditionally, the development of knowledge bases has been performed manually by domain experts and knowledge engineers. However, this is an expensive and error prone task. This difficulty in making explicit the knowledge implicitly contained in texts, databases or minds of domain experts is called the knowledge acquisition bottleneck and overcoming this problem is crucial to the success of knowledge based applications.

Ontologies constitute one of the formalisms for representing knowledge used by knowledge systems. They have proven to be a powerful tool for supporting natural language processing [5][14], information filtering [22], information retrieval [30], the Semantic Web [29] and many other tasks. However, these applications also suffer from the knowledge acquisition bottleneck. An approach for this problem is the automatic or semi-automatic construction of ontologies, a field of research that is usually referred to as ontology learning [10].

A concept hierarchy constitutes the backbone of an ontology and many techniques have been proposed for extracting concept hierarchies from text [5][16][21][31]. One problem with such approaches is that either they are based on statistical analysis [15][21] which does not consider the relationships between terms or on linguistic patterns [16][31], which are not able to deal with the noise present in texts that arise from natural language ambiguity. Statistical relational learning [23] approaches, such as Markov Logic Networks [27], can be used to alleviate this problem since they combine the expressive power of knowledge representation formalisms with probabilistic learning approaches.

This work proposes the Probabilistic Relational Hierarchy Extraction (PREHE) technique and investigates the suitability of statistical learning techniques for ontology learning tasks. PREHE uses Markov Logic Networks and Natural Language Processing techniques for extracting taxonomic relationships between concepts from natural language corpora. The text corpus is pre-processed for generating relational data. The concepts are extracted by the PRECE technique [11] and the taxonomic relationships are discovered by performing inference on the MLN described here. A preliminary evaluation of the proposed approach is outlined.

This paper is organized as follows. Section 2 defines ontologies and introduces basic concepts about ontology learning. Section 3 defines Markov Logic Networks and some learning tasks performed with them. Section 4 describes the PREHE technique for extracting concept hierarchies from text. Section 5 presents an experiment carried out to evaluate PREHE. Section 6 discusses related work and, finally, section 7 presents some concluding remarks.

2. ONTOLOGIES

This section introduces the ontology formal definition used in this work. For a more detailed discussion on ontologies and ontology learning, please refer to [1] and [10]. Formally, an ontology can be defined as in definition 1.

Definition 1. An ontology is a tuple $\mathcal{O} := (C, H, I, R, A)$ where:

- C is the set of entities in the domain being modeled The set C is the union of two sets, i.e., $C := C^C \cup C^I$ where C^C is the set of ontology concepts and C^I is the set of ontology instances.
- $H \subseteq C^C \times C^C$ is a set of taxonomic relationships. Such relationships define the concept hierarchy.
- $I \subseteq C^C \times C^I$ is the set instantiation relationships.
- $R := \{rel_k(c_1, c_2, ..., c_n) | \forall_i c_i \in C\}$ is the set of non-taxonomic relationships.
- $A := \{condition_x \rightarrow conclusion_y(c_1, c_2, ..., c_n) | \forall_i c_i \in C^C\}$ is a set of axioms.

Each ontology concept and relationship has a unique identifier. Besides that, they are associated to one or more natural language terms. Because of that, some ontologies also have a *lexicon* associated with them. A *lexicon* is a structure that maps natural language terms to concepts and relations of an ontology. Definition 2 defines a lexicon.

Definition 2. A lexicon is a tuple $L_{\mathcal{O}} := (L^C, L^R, F, G)$ where:

- $L_{\mathcal{O}}$ is a lexicon L associated with an ontology $\mathcal{O}$;
- L^C and L^R are the sets of lexical entries for concepts and relations, respectively;
- $F \subseteq L^C \times C$ a set of relationships that associates a lexical entry to a certain concept in the ontology $\mathcal{O}$;
- $G \subseteq L^R \times R$ a set of relationships that associates a lexical entry to a certain relation in the ontology $\mathcal{O}$.

3. MARKOV LOGIC NETWORKS

Statistical relational learning is an approach for machine learning that combines the expressive power of knowledge representation formalisms with probabilistic learning, thus enabling one to represent syntactic dependencies between words and capturing statistical information of words in text. Many statistical relational learning approaches have been proposed in the literature [23][26][27]. Markov Logic Networks (MLNs) [27] constitute an approach for statistical relational learning that combines first order logic with Markov networks.

A MLN is a first order logic knowledge base with weights, that can be either positive or negative, associated to each formula. While a traditional first order logic knowledge base is a set of hard constraints on the set of possible worlds, i.e. each world that violates a formula is impossible, an MLN is a set of softened constraints. The higher the weight of a formula, the less probable is a world that violates it. Worlds that violate formulas with negative weights are more probable instead.

Two common inference tasks in Markov Logic are the MAP (maximum a posteriori) and probabilistic inference. MAP inference aims at finding the most probable state of the world given some evidence. In Markov Logic this task is the same as finding the truth assignment that maximizes the sum of the weights of satisfied formulas. This can be done by any weighted satisfiability solver. For instance, variants of the WalkSat algorithm [17] like the MaxWalkSat have been used for this task [27].

Probabilistic inference aims at determining the probability of a formula given a set of constants and, maybe, other formulas as evidence. The probability of a formula is the sum of the probabilities of the worlds where it holds. Computing such probabilities can be expensive, thus approximate methods such as MCMC (Markov chain Monte Carlo) inference [13] constitute a reasonable alternative and are generally used in combination with the MC-SAT algorithm [25].

There are two approaches for learning the weights of a given set of formulas: generative and discriminative learning. Generative learning aims at maximizing the joint likelihood of all predicates while discriminative, at maximizing the conditional likelihood of the query predicates given the evidence ones. For a detailed discussion of inference and learning in MLNs, the reader is referred to [27].

4. PREHE: PROBABILISTIC RELATIONAL HIERARCHY EXTRACTION

This section describes PREHE (Probabilistic RElational Hierarchy Extraction), a technique for extracting concept hierarchies from natural language corpora that uses probabilistic relational learning. PREHE is part of a process for ontology learning and uses as input a set of concepts extracted by the PRECE technique [11]. A brief description of the PRECE technique is provided here. Since MLNs work with relational data, natural language corpora must be pre-processed in order to extract relational data. The pre-processing phase is described in subsection 4.1. Once the corpus is pre-processed, it can be used as input for hierarchy extraction. Subsection 4.2 provides a brief description of PRECE technique. Finally, subsection 4.3 describes the PREHE technique.

4.1 Corpus Pre-Processing

The first step of the pre-processing phase is to tokenize the corpus. After that, the tokens are annotated with part of speech (POS) tags and their lemmas. After the lemmatization, the chunking step takes place. The goal of this phase is to discover sets of words that, together, form a syntactic unit.

After that, the syntactic analysis takes place in order to extract the syntactic dependencies between words. The syntactic dependencies are relationships that words hold within a sentence. They indicate, for instance, who are the subject and the object of a given verb or which noun is modified by a given adjective. The syntactic dependencies considered in this work are represented according to the Stanford dependencies [8]. Then, the tokens containing terms from a stop list are removed. The remaining terms are weighted using TF-IDF scores [28] only terms with weights above a certain value are selected. At last, the WordNet [12] is used for extracting hypernym relations between the selected terms.

4.2 The PRECE technique

Concept Identification is performed by inference in Markov Logic. PRECE [11] extracts concepts by grouping terms by their co-occurrences and syntactic dependencies, i.e., terms that co-occur frequently or are usually related with the same terms through the same syntactic dependencies are grouped into the same concept, i.e. it learns the sets C^C and F.

The goal of the inference process is to infer the truth values of the possible groundings of the $F(concept, term)$ predicate, which means that a term is a lexical realization of a concept, i.e. $(w_j, c_k) \in F$, based on the evidence. The evidence is composed of a set of groundings of the $HasTerm(document, term)$ predicate, which means that a term is present in a document and a set of groundings of the $Depends(term, term, dependency)$ predicate, which means that a term governs another term through the specified syntactic dependency.

4.3 The PREHE technique

The goal of the PREHE technique is to organize hierarchically the concepts extracted with the PRECE technique. In other words, its goal is to learn the set H from definition 1. This task can be defined as a link prediction task [26]. Link prediction aims at discovering whether a given relationship holds for two objects. More formally, given a relationship r, the link prediction task is to find out the pairs of objects x and y for which $r(x, y)$ is true. In the hierarchy extraction context, the relationship r is the taxonomic relationship $Kind\_Of$ and the objects are the classes in C^C.

The PREHE technique aims at finding out the most likely truth values for the groundings of the query predicate given an MLN, and some evidence. This is clearly a MAP inference task in Markov logic.

In the PREHE technique, the $Kind\_Of(concept, concept)$ predicate is the query predicate and the evidence is composed by the groundings of the $Hypernym(term, term)$ and $Depends(term, term, type)$ predicates extracted from the corpus during the pre-processing phase as well as the groundings of the F predicate, the output of PRECE technique.

The structure of the MLN used by PREHE is given in Table 1. Together with its weight, Formula 1 in Table 1 captures the probability that the taxonomic relationship holds between two randomly chosen concepts. The higher the weight of this formula, the lower the number of taxonomic relationships extracted. It is worth stating that the weight is estimated automatically from training data.

Table 1: Structure of the MLN used by the PREHE Technique

Identifier	Formula
1	$\neg Kind\_Of(c_1, c_2)$
2	$Kind\_Of(c_1, c_2) \Rightarrow \neg Kind\_Of(c_2, c_1)$
3	$F(c_1, t_1) \wedge F(c_2, t_2) \wedge Depends(t_3, t_1, +dep) \wedge Depends(t_3, t_2, +dep) \Rightarrow Kind\_Of(c_1, c_2)$
4	$F(c_1, t_1) \wedge F(c_2, t_2) \wedge Hypernym(t_1, t_2) \Rightarrow Kind\_Of(c_1, c_2)$
5	$F(c_1, t_1) \wedge F(c_2, t_1) \wedge F(c_1, t_2) \wedge \neg F(c_2, t_2) \Rightarrow Kind\_Of(c_1, c_2)$

Formula 2 in Table 1 states that the $Kind\_Of$ relationship is not symmetric, i.e., if $Kind\_Of(c_1, c_2)$ is true then $Kind\_Of(c_2, c_1)$ is false.

The existence of certain syntactic dependencies between terms may be an evidence of the existence of a taxonomic relationship between them. This fact is captured by the formula 3 in Table 1. The notation $+dep$ means that a different weight will be computed for each type of syntactic dependency.

Intuitively, one can realize that the existence of a hypernym relationship between two terms belonging to the lexical realizations of two different concepts is an evidence that a taxonomic relationship between them might exist. For instance, the concept designated by the terms "city" and "urban center" is likely to be a generalization of the concept designated by the term "capital" since the term "city" is an hypernym of the term "capital". This is expressed by formula 4 in Table 1.

Approaches for hierarchy extraction based on hierarchical clustering [3][6] assume that the set of lexical realizations of a concept is a subset of the lexical realizations of its superconcept. This means that if the intersection between the lexical realizations of two concepts is not empty, they are likely to be related by the taxonomic relationship. But which one of them is the superconcept? Since the superconcept is more general, it may be referred by more natural language terms, having terms in its lexical realizations set that are not present in the subconcept. This is expressed by formula 5 in Table 1.

The weights for the MLN in Table 1 are automatically learned using discriminative weight learning. The MAP inference, i.e. finding the most probable truth assignment for the groundings of the $Kind\_Of$ predicate, is performed by the MaxWalkSAT algorithm [27]. The results of the inference (i.e. the set of true groundings of the $Kind\_Of$ predicate) are then written in an OWL ontology.

5. EVALUATION

The PREHE technique was evaluated by comparing its output with a gold-standard. The training of the MLNs used by the PREHE technique was performed using the GENIA[1] corpus. This corpus is semantically annotated according to the GENIA ontology, developed by the Tsuji laboratory [24].

For comparing the evaluated techniques, the *LonelyPlanet* corpus [18], consisting of 1801 documents, and an ontology in the tourism domain developed in the context of the GETESS project[2], from now on called $\mathcal{O}_{Tourism}$, were used. $\mathcal{O}_{Tourism}$ is composed of 969 concepts and 961 taxonomic relationships.

In this work, three hierarchy extraction techniques were used in order to extract the concepts from the *LonelyPlanet* corpus. The extracted sets of concepts, as well as the respective techniques used for extracting them are:

- H_{PREHE} - hierarchy extracted using the PREHE technique. Learning and inference in MLNs were performed using the Alchemy software package [20]. The tasks related to the corpus pre-processing were performed using GATE [7];
- H_{FCA} - hierarchy extracted using Formal Concept Analysis (FCA) as described in [5];

[1]http://www-tsujii.is.s.u-tokyo.ac.jp/~genia/topics/Corpus/
[2]http://www.aifb.uni-karlsruhe.de/WBS/pci/TourismGoldStandard.isa

- H_{Hearst} - hierarchy extracted using Hearst patterns [16].

These sets were compared with the $\mathcal{O}_{Tourism}$ ontology, using the taxonomic recall (TR), taxonomic precision (TP) and f-measure (TF), as defined in [9]. Taxonomic recall and precision are two global measures used for assessing the quality of a hierarchy in comparison to a gold standard. The f-measure is an harmonic mean of both.

The results of the experiments are shown in Figure 1. This figure shows that the PREHE technique obtained the best combined output between taxonomic precision and recall (F-measure), due to its higher recall. However it had the worst results regarding taxonomic precision. In Table 2 it is possible to see that H_{Hearst} is composed by a low number of taxonomic relationships. The high precision obtained by the Hearst patterns suggests that such patterns are indeed a strong evidence of taxonomic relationships, but they occur with an extremely low frequency on the corpus which explains its low taxonomic recall and the low number of taxonomic relationships extracted.

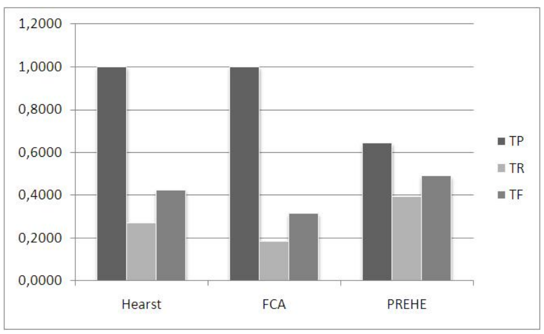

Figure 1: Results of the experiments on the *LonelyPlanet* corpus

The high taxonomic precision achieved by FCA and the higher number of taxonomic relationships extracted, indicate a good effectiveness of this approach which, however, presents a low taxonomic recall. By considering more types of syntactic dependencies than FCA (FCA considers only dependencies between nouns and verbs) and also the hypernym relationship, the PREHE technique managed to achieve a higher taxonomic recall but at the cost of a lower precision. Nevertheless PREHE obtained the best combined results for taxonomic recall and precision.

Table 2: Number of taxonomic relationships in each hierarchy

Hierarchy	Number of extracted taxonomic relationships
H_{PREHE}	170
H_{FCA}	364
H_{Hearst}	17
$H_{Tourism}$	961

6. RELATED WORK

Work has been done on learning concept hierarchies from text and the PREHE technique combines ideas from some of the state-of-the-art techniques. For instance, in [5] Formal Concept Analysis is used in order extract concepts by grouping terms related to the same sets of verbs. The hierarchy is extracted by applying the partial ordering operator to the extracted concepts (represented as FCA formal contexts).

Another popular approach for ontology learning is to make use of hierarchical clustering [3][6] techniques. Such techniques learn concepts by grouping terms according to some similarity measure. Such measures may be based on statistical analysis such as the Harris hypothesis [15] and co-occurrence [21]. A comparative study of clustering techniques for ontology learning is found in [4].

There is also some work on extracting taxonomic relations from text using linguistic patterns, as shown in [16] and [31].

Markov logic networks have also been successfully used for learning concept hierarchies. For instance, in [32] MLNs and Support Vector Machines [2] are used in order to extract subsumption hierarchies from Wikipedia's infoboxes. As well as the PREHE technique, the approach in [32] also uses WordNet as a source of background knowledge. However, it also exploits the semi-structured nature of the infoboxes, while PREHE works with completely unstructured data sources in combination with the PRECE technique [11]. The PRECE technique uses MLNs to extract ontology concepts from textual sources, but the concepts are not organized into a hierarchy (which is extracted by the PREHE technique).

7. CONCLUDING REMARKS

This work introduced PREHE, a technique for learning concept hierarchies from text. The approach proposed here makes use of ideas from different state-of-the-art methods for hierarchy learning together with new statistical relational learning techniques.

Traditional statistical techniques assume that data is independent and identically distributed (i.i.d.). Since syntactic analysis reveals relationships between the terms represented by syntactic dependencies, the i.i.d. assumption does not hold for ontology learning from text. The experiments conducted in this work showed that by considering relationships between terms such as hypernym and syntactic dependencies, the PREHE technique achieved reasonable results compared to the state-of-the-art techniques thus, giving evidence that statistical relational learning is a suitable approach for ontology learning.

However, the gold standard based evaluation used in this work measures how good the automatically extracted ontologies correlated to the judgement of the domain experts that built the gold standard ontology. On the top of that, the gold standard is a hand crafted ontology, developed by an error prone process. If the gold standard ontology presents modeling problems, the evaluation method rewards ontologies with similar problems and penalizes ontologies with concepts or relationships not appearing in the gold standard. Therefore, the evaluation of the proposed technique in a running application is also an ongoing work. For such evaluation, an ontology based information retrieval system [30] is being developed, and its effectiveness using ontologies built with different approaches will be measured.

PRECE and PREHE technique are able together to extract a concept hierarchy from text. The next step is to extend this approach for learning non-taxonomic relationships and axioms. For instance, statistical predicate invention [19] may be used for discovering non-taxonomic relationships and MLN structure learning for axioms.

8. REFERENCES

[1] P. Buitelaar, P. Cimiano, and B. Magnini. Ontology learning from text: An overview. *Frontiers in Artificial Intelligence and Applications Series*, 123, 2005.

[2] C. Burges. A tutorial on support vector machines for pattern recognition. *Data mining and knowledge discovery*, 2(2):121–167, 1998.

[3] S. Caraballo. Automatic construction of a hypernym-labeled noun hierarchy from text. *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, pages 120–126, 1999.

[4] P. Cimiano, A. Hotho, and S. Staab. Comparing conceptual, divisive and agglomerative clustering for learning taxonomies from text. In *Proceedings of the European Conference on Artificial Intelligence (ECAI)*, pages 435–443, 2004.

[5] P. Cimiano, A. Hotho, and S. Staab. Learning concept hierarchies from text corpora using formal concept analysis. Technical report, Institute AIFB, Universität Karlsruhe, 2004.

[6] P. Cimiano and S. Staab. Learning concept hierarchies from text with a guided hierarchical clustering algorithm. In *ICML workshop on Learning and Extending Lexical Ontologies with Machine Learning Methods*, 2005.

[7] H. Cunningham. GATE, a general architecture for text engineering. *Computers and the Humanities*, 36(2):223–254, 2002.

[8] M. de Marneffe, B. MacCartney, and C. Manning. Generating typed dependency parses from phrase structure parses. In *LREC 2006*, 2006.

[9] K. Dellschaft and S. Staab. On how to perform a gold standard based evaluation of ontology learning. In *Proceedings of ISWC-2006 International Semantic Web Conference*, 2006.

[10] L. Drumond and R. Girardi. A survey of ontology learning procedures. In F. L. G. de Freitas, H. Stuckenschmidt, H. S. Pinto, A. Malucelli, and Óscar Corcho, editors, *WONTO*, volume 427 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2008.

[11] L. Drumond and R. Girardi. An experiment using markov logic networks to extract ontology concepts from text. In *Proceedings of the II Workshop on Web and Text Intelligence (STIL-WTI 2009)*, São Carlos, Brazil, 2009.

[12] C. Fellbaum. *Wordnet: An Electronic Lexical Database*. MIT Press, 1998.

[13] W. Gilks, S. Richardson, and D. Spiegelhalter. *Markov chain Monte Carlo in practice*. Chapman & Hall/CRC, 1996.

[14] R. Girardi and B. Ibrahim. Using english to retrieve software. *Journal of Systems Software, Special Issue on Software Reusability*, 30(3):249–270, 1995.

[15] Z. Harris. *Mathematical Structures of Language*. Wiley, 1968.

[16] M. Hearst. Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the 14th International Conference on Computational Linguistics*, pages 539–545, 1992.

[17] H. Kautz, B. Selman, and Y. Jiang. A general stochastic approach to solving problems with hard and soft constraints. *The Satisfiability Problem: Theory and Applications*, pages 573–586, 1997.

[18] M. Kavalec and V. Svatek. A study on automated relation labelling in ontology learning. *Ontology Learning from Text: Methods, Evaluation and Applications*, pages 44–58, 2005.

[19] S. Kok and P. Domingos. Statistical predicate invention. In *Proceedings of the 24th international conference on Machine learning*, pages 433–440. ACM New York, NY, USA, 2007.

[20] S. Kok, M. Richardson, P. Singla, H. Poon, D. Lowd, J. Wang, and P. Domingos. The alchemy system for statistical relational ai. Technical report, Department of Computer Science and Engineering, University of Washington, Seattle, WA, 2009.

[21] A. Maedche and S. Staab. *Ontology learning*, pages 173–190. Springer, 2004.

[22] S. Middleton, N. Shadbolt, and D. D. Roure. Ontological user profiling in recommender systems. *ACM Transactions on Information Systems*, 22:54–88, 2004.

[23] J. Neville, M. Rattigan, and D. Jensen. Statistical relational learning: Four claims and a survey. In *Workshop SRL, Int. Joint. Conf. on AI*, 2003.

[24] T. Ohta, Y. Tateisi, and J. Kim. The GENIA corpus: An annotated research abstract corpus in molecular biology domain. In *Proceedings of the second international conference on Human Language Technology Research*, pages 82–86. Morgan Kaufmann Publishers Inc. San Francisco, CA, USA, 2002.

[25] H. Poon and P. Domingos. Sound and efficient inference with probabilistic and deterministic dependencies. In *PROCEEDINGS OF THE NATIONAL CONFERENCE ON ARTIFICIAL INTELLIGENCE*, volume 21, page 458. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999, 2006.

[26] A. Popescul and L. Ungar. Statistical relational learning for link prediction. In *IJCAI workshop on learning statistical models from relational data*, 2003.

[27] M. Richardson and P. Domingos. Markov logic networks. *Machine Learning*, 62(1):107–136, 2006.

[28] G. Salton and C. Buckley. Term Weighting Approaches in Automatic Text Retrieval. 1987.

[29] N. Shadbolt, W. Hall, and T. Berners-Lee. The semantic web revisited. *Intelligent Syst*, 21(3):96–101, 2006.

[30] F. Silva, R. Girardi, and L. Drumond. A knowledge-based retrieval model. In *Proceedings of the 21st International Conference on Software Engineering & Knowledge Engineering (SEKE'2009)*, pages 558–563, Boston, USA, 2009. Knowledge Systems Institute Graduate School.

[31] R. Snow, D. Jurafsky, and A. Ng. Learning syntactic patterns for automatic hypernym discovery. *Advances in Neural Information Processing Systems*, 17:1297–1304, 2005.

[32] F. Wu and D. S. Weld. Automatically refining the wikipedia infobox ontology. In *WWW '08: Proceeding of the 17th international conference on World Wide Web*, pages 635–644, New York, NY, USA, 2008. ACM.