An Experiment Using Markov Logic Networks to Extract Ontology Concepts from Text

Lucas Drumond and Rosario Girardi

Federal University of Maranhão, Computer Science Department,
Av. dos Portugueses, s/n 65085-580 São Luís - MA, Brazil
ldrumond@gmail.com,rgirardi@deinf.ufma.br
http://gesec.deinf.ufma.br

Abstract. Since ontology development is currently an error prone, time consuming and expensive task, ontology based systems suffer from the so called knowledge acquisition bottleneck. One approach for this problem is to provide automatic or semi-automatic support for ontology construction. This field of research is known as ontology learning. Many techniques for learning ontologies from text have been proposed, most of them based on statistical learning and natural language processing methods. This work presents an approach for extracting ontology concepts from text that combines both ideas through statistical relational learning. A Markov Logic Network as been developed for this task and is described here.

Key words: Ontology Learning; Knowledge Acquisition; Statistical Relational Learning

1 Introduction

Though ontologies hold a great importance in modern knowledge based systems, their development is currently an error prone, time consuming and expensive task. 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 [1][2].

According to Buitelaar [1] one of the sub-phases of ontology learning is concept extraction. Many techniques for concept extraction rely either on statistical analysis [3][4][5] or on linguistic patterns [6][7]. Statistical methods make use of the bag-of-words approach, which assumes that the terms are not correlated, i.e. they do not consider relation between words, represented by their syntactic dependencies. Such relations can be used by relational learning techniques by representing them through knowledge representation formalisms such as first order logic. However, relational learning techniques are not able to deal with the noise that arises from polysemy and ambiguity present in natural language texts.

This work presents the Probabilistic Relational Concept Extraction (PRECE) technique and investigates the suitability of statistical learning techniques for ontology learning tasks. Statistical relational learning [8] combines the expressive

power of relational learning with probabilistic learning approaches, thus providing a suitable framework for representing word relationships and capture statistical information of words in text. PRECE makes use of Markov Logic Networks [9] for extracting ontology concepts from a natural language corpus.

The paper is organized as follows. Section 2 introduces the definition for the term "ontology" considered in this work. Section 3 briefly introduces Markov Logic Networks. Section 4 introduces the PRECE technique. Section 5 presents the results of the evaluation of PRECE. Section 6 discusses related work and Section 7, some concluding remarks.

2 Ontologies

This section presents the ontology definition used in this work. For a more detailed discussion on ontologies and ontology learning, please refer to [1] and [2]. Formally, an ontology can be defined, according to [10], as in definition 1.

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

- C *is the set of ontology concepts.*
- $H \subseteq C \times C$ *is a set of taxonomic relationships. Such relationships define the concept hierarchy.*
- R *is the set of non-taxonomic relationships.*
- rel *is a function* $rel : R \rightarrow C \times C$ *that assigns identifiers to the relations in the set* R*.*
- $A^{\mathcal{O}}$ *is a set of axioms, usually formalized into some logic language.*

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 according to Maedche [10].

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 combines the expressive power of knowledge representation formalisms with probabilistic learning approaches, 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. Markov Logic Networks (MLNs) [9] constitute an approach for statistical relational learning that combines first order logic with Markov random fields.

An 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 a world the violates it is. Worlds that violate formulas with negative weights are more probable instead.

Two common inference tasks in Markov Logic are the maximum a posteriori (MAP) 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 [11] like the MaxWalkSat have been used for this task [9].

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 [12] constitute a reasonable alternative and are generally used in combination with the MC-SAT algorithm [13].

For a detailed discussion of inference in MLNs, the reader is referred to [9].

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. In both cases, the existence of some atoms with unknown truth values (open world assumption) can be handled with a form of the EM (Expectation Maximization) algorithm. For a detailed discussion of inference and learning in MLNs, the reader is referred to [9].

4 PRECE: Probabilistic RElational Concept Extraction

This section describes PRECE (Probabilistic RElational Concept Extraction), a technique for extracting ontology concepts from natural language corpora that uses probabilistic relational learning. 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 concept extraction. Subsection 4.2 describes how PRECE extracts concepts from the pre-processed corpus.

4.1 Corpus Pre-Processing

In the pre-processing phase, the corpus is tokenized. 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.

At last, 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 [14]. Before starting the Concept Extraction phase, the tokens containing terms from a stop list are removed.

4.2 Concept Extraction

The problem here is to learn the set of ontology concepts (the set C from definition 1) from a given text corpus D. The approach proposed here learns concepts through their linguistic realizations, i.e. each concept is learned as a set of natural language terms. This implies in learning the sets L^C and F from definition 2 as well as the set C from definition 1. This task is performed by three steps: Term Extraction, Concept Identification and Concept Labeling as shown in Figure 1.

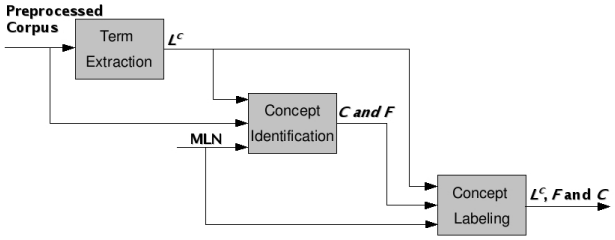

Fig. 1. Steps of the PRECE technique

The extraction of the set L^C is performed by traditional term extraction techniques. First, since we are interested in extracting concepts, and they are, most of the times expressed by nouns and noun phrases, words like prepositions, adverbs, pronouns and verbs are removed. This is done by checking the part of speech tags identified during the pre-processing.

At last, terms are weighted and only terms with weights above a certain value are selected. By *term* we mean a token or a noun phrase extracted during the preprocessing. Currently, the terms are weighted using TF-IDF (Term Frequency - Inverse Document Frequency) scores. Term frequencies are computed based on term lemmas. For instance if a document has one occurrence for the term *"cell"* and one for the term *"cells"*, the document frequency for the lemma *"cell"* has value 2.

The output of the term extraction process is the set L^C and annotations indicating which tokens are present in which documents.

Concept Identification is performed by inference in Markov Logic. In order to do that, the user must specify the number of concepts to be extracted. This corresponds to the number of groundings of the $Concept(concept)$ predicate in the evidence file used as input for the inference process. For n concepts, there will be n groundings $Concept(c_i)$ where $1 \leq i \geq n$ and c_i is an arbitrarily defined unique identifier. 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 the $Depends(term, term, dependency)$ predicate, which means that a term governs another term through the specified syntactic dependency.

To help to clarify the technique, take as an example a corpus containing two documents, namely *DOC1* and *DOC2*. For the sake of simplicity, *DOC1* and *DOC2* contain each one a single sentence as shown in Table 1.

Table 1. Two sample documents

DOC1	*DOC2*
A country's capital is the most important city of the nation.	*Brazil's capital, Brasilia, is a modern city.*

The evidence file with the groundings of the $HasTerm$ and *Depends* predicates is depicted in Figure 2. It is important to bear in mind that the value of this example is merely didactic since PRECE technique requires a considerably larger amount of text to work properly.

```
HasTerm(DOC1,"country")      HasTerm(DOC2,"city")
HasTerm(DOC1,"capital")      Depends("capital","country",POSS)
HasTerm(DOC1,"city")         Depends("capital","brazil",POSS)
HasTerm(DOC1,"nation")       Depends("city","capital",NSUBJ)
HasTerm(DOC2,"brazil")       Depends("capital","brasilia",APPOS)
HasTerm(DOC2,"capital")      Concept(1)
HasTerm(DOC2,"brasilia")     Concept(2)
```

Fig. 2. A sample file containing evidence ground predicates

It is also worth noting that documents are related to concepts. This relation, represented by the $HasConcept(document, concept)$ predicate, is the same as the relation between documents and latent topics in Probabilistic Latent Semantic Analysis (PLSA) [15]. Given the evidence that a term t is present in a document d and the concept c is a concept of document d, it is more likely that $F(c, t)$ holds. This is captured by assigning a positive weight to the following formula:

$$HasTerm(d, t) \wedge HasConcept(d, c) \Rightarrow F(c, t) \tag{1}$$

Also, two terms appearing in the same document are more likely to be related to the same concept. This is the case for the terms *country* and *nation* in document *DOC1*. This is captured by the following rule:

$$HasTerm(d, t_1) \wedge HasTerm(d, t_2) \wedge F(c, t_1) \Rightarrow F(c, t_2) \tag{2}$$

Two terms having the same syntactic dependency with the same term are more likely to denote the same concept. This can be observed in the given example by the terms *Brazil* and *country*. They both are related to the term *capital* through the dependency $POSS$. This can be expressed as follows.

$$Depends(t_3, t_1, dep) \wedge Depends(t_3, t_2, dep) \wedge F(c, t_1) \Rightarrow F(c, t_2) \tag{3}$$

To avoid the proliferation of term-concept assigning (i.e. a great number of groundings for the F predicate) the following clause is added to the MLN, stating that, given no evidence, a term does not denotes a concept.

$$\neg F(c, t) \tag{4}$$

Formula 5 states that a term is only related to one concept. The likelihood that a given term is polysemic is captured by the weight of this formula, automatically learned from the training data.

$$F(c_1, t) \wedge c_1 \neq c_2 \Rightarrow \neg F(c_2, t) \tag{5}$$

The concepts are extracted by performing probabilistic inference on the presented MLN. The evidence predicates $HasWord$ and $Depends$ are closed world by default. These groundings are automatically extracted from the corpus during pre-processing and term extraction. The query predicate is the F predicate. Since we do not know in advance the truth assignments for the groundings of the $HasConcept$ predicate, it is open world.

The goal of the probabilistic inference process is to infer the probabilities of the possible groundings of the F and $HasConcept$ predicates. In the example used here, the goal is to find the probabilities that each one of the ground atoms in Figure 3 is true given the evidence file from Figure 2, and the MLN composed by the formulas introduced in this section and their respective weights, automatically learned with using discriminative learning and EM. From these results it is possible to extract the lexical realization of concepts as follows. Be

a word w and a concept c, if $P(F(c,t)) > 0.5$ the w is a lexical realization of c. PRECE uses the MC-SAT algorithm for probabilistic inference described in [13].

F(1,"country")	F(2,"country")	HasConcept(DOC1,1)
F(1,"capital")	F(2,"capital")	HasConcept(DOC1,2)
F(1,"city")	F(2,"city")	HasConcept(DOC2,1)
F(1,"nation")	F(2,"nation")	HasConcept(DOC2,2)
F(1,"brazil")	F(2,"brazil")	
F(1,"brasilia")	F(2,"brasilia")	

Fig. 3. Ground atoms which probabilities should be inferred

The concepts identified so far are sets of terms and no label is assigned to them. Since probabilistic inference is used for extracting concepts, the result is the probability of each grounding of the query predicate F. Each concept, is labeled with the term t with highest probability $P(F(c,t))$. In the case of a tie, the term with the highest score computed in the term extraction phase is chosen.

5 Evaluation

The PRECE technique was evaluated by comparing its output with a gold-standard. The training of the MLNs used by the PRECE technique was performed using the GENIA[1] corpus. This corpus is semantically annotated according to the GENIA ontology, developed by the Tsuji laboratory [16].

The *LonelyPlanet* corpus [17] 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 for comparing the evaluated techniques.

In this work, three concept 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:

- $\mathcal{C}_{PRECE}$ - concepts extracted using the PRECE technique. Learning and inference in MLNs are performed using the Alchemy software package [18]. The tasks related to the corpus pre-processing were performed using GATE [19];
- $\mathcal{C}_{PRECEdep^-}$ - concepts extracted using the PRECE technique but without considering the syntactic dependencies between terms i.e., using an MLN as described in section 4.2 but without the formula in equation 3;
- $\mathcal{C}_{PLSA}$ - concepts extracted using traditional PLSA. Each topic discovered by the PLSA process is a concept. Concepts are labeled with the term with highest probability given the respective topic.

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

These sets were compared with the $\mathcal{O}_{Tourism}$ ontology, using the recall, precision and f-measure, as defined in [20]. The recall is defined as the total of concepts that are present in both concept sets divided by the amount of concepts in the gold standard. The precision is defined as the total of concepts that are present in both ontologies (their intersection) divided by the amount of concepts in the learned ontology. The F-measure is an harmonic mean of both. The match between the induced concepts and the concepts in the gold standard ontology was computed by simple string comparison.

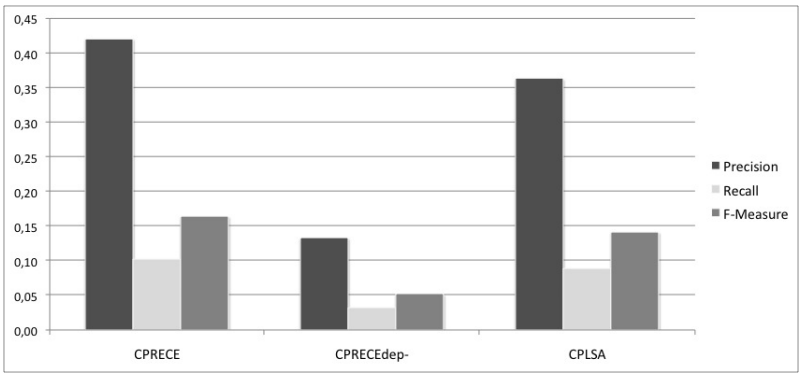

Fig. 4. Results of the experiments on the *LonelyPlanet* corpus

The results of the experiments are shown in Figure 4. This figure shows that PRECE outperforms pure PLSA for concept extraction. The difference between the $\mathcal{O}_{PRECE}$ and $\mathcal{O}_{PRECEdep^-}$ ontologies provides a strong evidence that the usage of syntactic dependencies can yield improvements on the effectiveness of ontology learning techniques.

6 Related Work

Work has been done on learning ontologies from text. Much of the work on concept learning is not decoupled from hierarchy learning. The extraction of the intension of a concept from text can be approached using Inductive Logic Programming (ILP) [21] or Formal Concept Analysis (FCA) [22].

Learning of linguistic realizations is usually carried out through clustering techniques. The main techniques used for this task are based on unsupervised hierarchical clustering [23]. 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 [3], distributional similarity [24] and co-occurrence [4] or on the semantic distance in structures such as the WordNet [25]. In [5] probabilistic latent semantic analysis was used for extracting ontology learning from texts, thus learning concepts as probability distributions over terms.

7 Concluding remarks

This work introduced PRECE, a technique for learning ontology concepts from text. The approach proposed here makes use of ideas from different state-of-the art methods for concept learning together with new statistical relational learning techniques and Probabilistic Latent Semantic Analysis.

The experiments conducted in this work showed that by considering the syntactic dependencies between terms, the effectiveness of the PRECE technique was greatly improved, thus giving strong evidence that statistical relational learning is a suitable approach for ontology learning, since it was able to remove the independent and identically distributed assumption by considering the relationships between terms.

The PRECE technique also has its drawbacks. Since it is a supervised approach, it is very sensible to the quality of the training set. Besides that, gathering such a training set is a hard task. This problem can be alleviated by investigating unsupervised techniques for learning Markov Logic Networks.

Another limitation of the proposed approach is that it learns only ontology concepts. The next step is to extend this approach for learning also the concept hierarchy and the non-taxonomic relations. Once the concepts are known, the goal is to determine whether the taxonomic relation, given by the set H^C exists between two given concepts.

References

1. Buitelaar, P., Cimiano, P., Magnini, B.: Ontology learning from text: An overview. Frontiers in Artificial Intelligence and Applications Series **123** (2005)
2. Drumond, L., Girardi, R.: A survey of ontology learning procedures. In de Freitas, F.L.G., Stuckenschmidt, H., Pinto, H.S., Malucelli, A., scar Corcho, eds.: WONTO. Volume 427 of CEUR Workshop Proceedings., CEUR-WS.org (2008)
3. Harris, Z.: Mathematical Structures of Language. Wiley (1968)
4. Maedche, A., Staab, S. In: Ontology learning. Springer (2004) 173–190
5. Zavitsanos, E., Paliouras, G., Vouros, G., Petridis, S.: Discovering subsumption hierarchies of ontology concepts from text corpora. In: Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence, IEEE Computer Society Washington, DC, USA (2007) 402–408
6. Hearst, M.: Automatic acquisition of hyponyms from large text corpora. In: Proceedings of the 14th International Conference on Computational Linguistics. (1992) 539–545
7. Snow, R., Jurafsky, D., Ng, A.: Learning syntactic patterns for automatic hypernym discovery. Advances in Neural Information Processing Systems **17** (2005) 1297–1304
8. De Raedt, L., Kersting, K.: Probabilistic logic learning. ACM SIGKDD Explorations Newsletter **5**(1) (2003) 31–48
9. Richardson, M., Domingos, P.: Markov logic networks. Machine Learning **62**(1) (2006) 107–136
10. Maedche, A.: Ontology Learning for the Semantic Web. Kluwer Academic Publishing (2002)

11. Kautz, H., Selman, B., Jiang, Y.: A general stochastic approach to solving problems with hard and soft constraints. The Satisfiability Problem: Theory and Applications (1997) 573–586
12. Gilks, W., Richardson, S., Spiegelhalter, D.: Markov chain Monte Carlo in practice. Chapman & Hall/CRC (1996)
13. Poon, H., Domingos, P.: Sound and efficient inference with probabilistic and deterministic dependencies. In: PROCEEDINGS OF THE NATIONAL CONFERENCE ON ARTIFICIAL INTELLIGENCE. Volume 21., Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999 (2006) 458
14. de Marneffe, M., MacCartney, B., Manning, C.: Generating typed dependency parses from phrase structure parses. In: LREC 2006. (2006)
15. Hofmann, T., Puzicha, J., Jordan, M.: Unsupervised learning from dyadic data. Advances in Neural Information Processing Systems **11** (1999)
16. Ohta, T., Tateisi, Y., Kim, J.: The GENIA corpus: An annotated research abstract corpus in molecular biology domain. In: Proceedings of the second international conference on Human Language Technology Research, Morgan Kaufmann Publishers Inc. San Francisco, CA, USA (2002) 82–86
17. Kavalec, M., Svatek, V.: A study on automated relation labelling in ontology learning. Ontology Learning from Text: Methods, Evaluation and Applications (2005) 44–58
18. Kok, S., Richardson, M., Singla, P., Poon, H., Lowd, D., Wang, J., Domingos, P.: The alchemy system for statistical relational ai. Technical report, Department of Computer Science and Engineering, University of Washington, Seattle, WA (2009)
19. Cunningham, H.: GATE, a general architecture for text engineering. Computers and the Humanities **36**(2) (2002) 223–254
20. Dellschaft, K., Staab, S.: On how to perform a gold standard based evaluation of ontology learning. In: Proceedings of ISWC-2006 International Semantic Web Conference. (2006)
21. Lavrac, N., Dzeroski, S.: Inductive Logic Programming: Techniques and Applications. Ellis Horwood (1994)
22. Cimiano, P., Hotho, A., Staab, S.: Learning concept hierarchies from text corpora using formal concept analysis. Technical report, Institute AIFB, Universität Karlsruhe (2004)
23. Caraballo, S.: 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 (1999) 120–126
24. Lin, D.: Automatic retrieval and clustering of similar words. In: Proceedings of the 17th international conference on Computational linguistics, Morristown, NJ, USA, Association for Computational Linguistics (1998) 768–774
25. Resnick, P.: Semantic similarity in a taxonomy: an information-based measure and its application to problems of ambiguity in natural language. Journal of Artificial Intelligence Research **11** (1999) 95–130