NTNU: An Unsupervised Knowledge Approach for Taxonomy Extraction

Bamfa Ceesay
Department of Computer Science and
Information Engineering
National Taiwan Normal University
No. 88, Tingz Chou Road, Section 4,
Taipei 116, Taiwan, R.O.C.
bmfceesay@csie.ntnu.edu.tw

Wen Juan Hou
Department of Computer Science and
Information Engineering
National Taiwan Normal University
No. 88, Tingz Chou Road, Section 4,
Taipei 116, Taiwan, R.O.C.
emilyhou@csie.ntnu.edu.tw

Abstract

Taxonomy structures are important tools in the science of classification of things or concepts, including the principles that underlie such classification. This paper presents an approach to the problem of taxonomy construction from texts focusing on the hyponym-hypernym relation between two terms. Given a set of terms in a particular domain, the approach in this study uses Wikipedia and WordNet as knowledge sources and applies the information extraction methods to analyze and establish the hyponym-hypernym relationship between two terms. Our system is ranked fourth among the participating systems in SemEval-2015 task 17.

1 Introduction

Taxonomies are essential tools for many Natural Language Processing (NLP) applications and the backbone of many structured knowledge resources. Taxonomies specific to a domain are becoming indispensable to a growing number of applications (Velardi *et al*., 2013). Several state-of-the-art approaches already exist to extract taxonomies to characterize the domains of interest from the corpus using the information extraction techniques. Recently, attention has been devoted to inducing the taxonomy from a set of keyword phrases instead of from a text corpus (Liu *et al*., 2012). Such approaches enrich the set of keyword phrases by aggregating search results for each keyword phrase into a text corpus to overcome the lack of explicit relationships between keyword phrases from which the taxonomy can be induced.

This approach faces a key challenge of extracting explicit relationships among keyword phrases. However, semantic relatedness between concepts in a domain is an important clue to extracting their taxonomy relationships. An important contribution in relation to this is reported by Gabrilovich *et al*. (2007) that present an explicit semantic analysis using the natural concepts and propose a uniform method of computing relatedness of both individual concept and arbitrarily long text fragments. Lexical databases such as WordNet (Miller, 1995) encode relations between words such as synonymy and hypernymy. Quite a few metrics have been defined that compute relatedness using various properties of the underlying graph structures of these resources. The obvious drawback of this approach is that the creation of lexical resources requires the lexicographic expertise as well as a lot of time and effort, and consequently such resources cover only a small fragment of the language lexicon. Specifically, such the resources contain few proper names, neologisms, slang, and domain-specific technical terms. Furthermore, these resources have strong lexical orientation and mainly contain information about individual words but little world knowledge in general.

Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015), pages 938–943,
Denver, Colorado, June 4-5, 2015. ©2015 Association for Computational Linguistics

With the advent of new information sources, many new methods and ideas are developed for the large scale information extraction taking advantages of huge amounts of unstructured available resources. Barbu and Poesio (2009) propose a novel method for acquisition of knowledge for taxonomies of concepts from the raw Wikipedia text. Their approach uses the learning process to derive concept hierarchies from WordNet and maps them to Wikipedia pages for extraction of appropriate knowledge. Most state-of-the-art approaches for the domain-specific taxonomy induction use the text corpus as its input and some information extraction methods to extract ontological relationships from the text corpus, and finally apply the relationships to build the taxonomy. Other automatic approaches to taxonomy construction from texts include a statistical method to compare the syntactic context of terms for taxonomic relations identification (Tuan *et al.*, 2014).

There have been a number of handcrafted, well-structured taxonomies publicly available online, including WordNet (Miller, 1995). However, such taxonomies are also not perfect since human experts are liable to miss some relevant terms.

In this study, we consider the challenging problem of deriving taxonomies of a set of concepts under a specific domain of interest. Consider for illustration, the domain *vehicle* containing concepts such as *car*, *bicycle*, *Toyota, automobile*, *bus*, *Toyota_cambire*, *cruiser* and *Motorcycle.* Establishing hyponym-hypernym relationships among concepts is a difficult task if no other information is provided. We propose an approach to the taxonomy extraction task in SemEval-2015 (Bordea *et al*., 2015) with the following contributions:

- To derive the statistical information about individual concepts in a given domain, the study uses WordNet and Wikipedia to find the definition for the concept.
- Using the definitions of concepts, the statistical information derived from these definitions is used to determine concept relationships and to represent the concepts in a domain with a Bayesian Rose Tree (BRT).
- The study finally extracts taxonomies for domain concepts using the BRT tree and WordNet type binary relations.

Bayesian hierarchical clustering algorithm (BRT) is used to cluster concepts having hyponym-hypernym relationships (Blundell *et al.*, 2012). Figure 1 presents our level approach to constructing the taxonomy for the domain concepts.

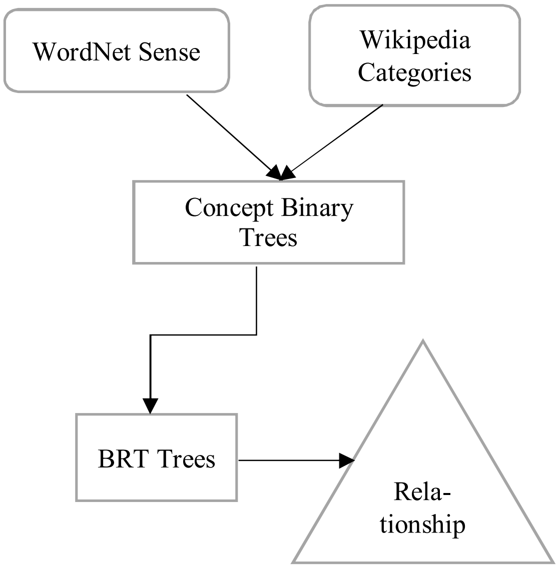

Figure 1. Approach to Taxonomy Extraction.

In Figure 1, resources WordNet and Wikipedia are first used to help the extraction of the definitions of the concepts. Then, information extracted from WordNet sense and Wikipedia categories are utilized to build the concept binary trees. With the concept binary trees, the system can construct the BRT tree and furthermore generate the relationships in the taxonomy for the concepts. Details in each step are described in the following sections.

2 Concept Definition and Bayesian Ross Tree

Definitions for describing concepts can be extracted from a variety of sources: dictionaries, databases, corpora, web directories and others. Wikipedia and WordNet have drawn attentions on derivation of concepts for taxonomy construction (Barbu and Poesio, 2009; Song *et al*., 2011) and the syntactic conceptual taxonomy (Tuan *et al*. 2014).

In this study, to generate definitions for concepts and map concepts and keywords in definitions to a BRT tree for taxonomy extraction, we follow the steps below:

- First, given a concept, we use WordNet and Wikipedia to derive its definitions. In addition, the related WordNet synset and Wikipedia category are extracted for the taxonomy induction.
- Using the Wikipedia categories that describe the corresponding concept article, the WordNet sense, and the WordNet hyponym tree from the first step, the study uses a binary tree to represent each concept in the given domain. The left node represents the set of terms considered to be hypernyms and the right node represents the set of terms considered to be hyponyms.

Applying the above steps, the binary tree representation of concepts in the given domain is used to construct the BRT tree for the taxonomy construction. One example of the BRT tree is shown in Figure 2 as below.

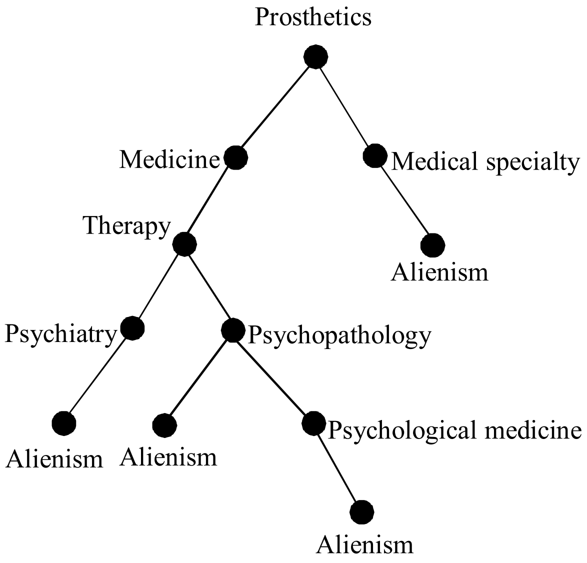

Figure 2. BRT Tree for a Term “Alienism,” in Science Domain.

Figure 2 illustrates the concept of BRT tree for a domain term, “*alienism*.” Each node represents a hypernym of the child node. For example, “psychiatry” is the hypernym of “alienism,” and “therapy” is the hypernym of “psychiatry” and “psychopathology” respectively.

3 Concept Binary Tree Construction

This study defines a set of concepts derived from the Wikipedia category structure, a set of terms in the WordNet hypernym sense, *whyp*, and a set of terms in the WordNet hyponym, *whypo*, for a given concept in the domain.[1] For the set of terms in the Wikipedia category, a syntax-based method is employed by referencing the research of Tuan *et al*. (2014) to derive the taxonomy structure for the category terms. In our case of study, *is_a* relationship is an identification of the hypernym and hyponym relationship between terms in the category set. That is, “**X** *is_a* **Y”** is translated as “**X** is a hyponym of **Y**.” However, this only shows a relationship among terms in the category set. To identify the hypernym-hyponym relationship between the domain concept and the terms in the category, the study uses the semantic relatedness approach proposed in the research of Wu *et al*. (2009). Finally, set operations are used to collect hypernyms and hyponyms and we use these features to construct a binary tree with the domain concept as the root, if no category term is a hypernym of domain concept. If there exists a category term that is a hypernym of the domain concept, the term becomes the root. For a given concept in the domain, a set of hypernym, *hyper*, and a set of hyponym, *hypo* are defined. After deriving the category taxonomy, *wt*, for Wikipedia categories, the following operations are defined:

$$hyper = whyper \cap wt.hypernym \tag{1}$$

$$hypo = whypo \cap wt.hyponym \tag{2}$$

[1] Wikipedia can be downloaded at *http://downlaod.wikimedia.org*. This study uses the English Wikipedia database dump

where *wt.hypernym* represents all hypernym terms connected to the category taxonomy *wt* and *wt.hyponym* represents all hyponym terms connected to the category taxonomy *wt*.

The multiple binary trees are used to construct BRT trees for the taxonomy extraction. However, it is worth to note that a cascading binary tree can be used instead of a BRT tree. For efficiency and computational purposes, a BRT tree is used, since a concept hypernym (parent node) can have more than two hyponyms (child node). The objective is to find relatedness between root concepts and the assigned parent node to the root concept of the binary tree. Figure 3 shows an illustration presenting concepts in our example domain in Section 3.

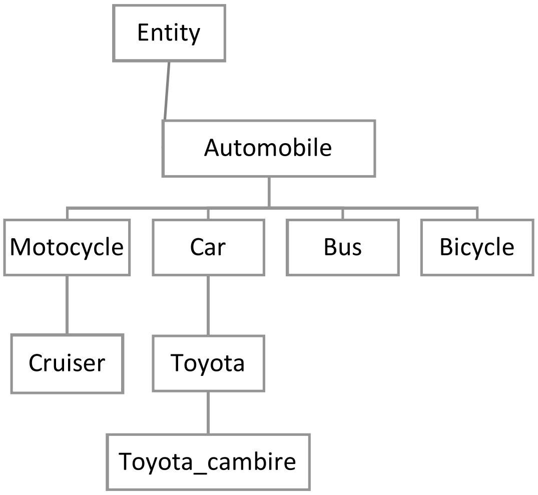

Figure 3. An Illustration of Concept Representation.

4 Extraction of Taxonomy Relationships

To extract the hypernym-hyponym relation between concepts, our approach uses the taxonomy de-scribed in Section 3. The concepts from the given domain are replaced by their concept IDs to distinguish them from the rest of the concepts in the BRT tree. The root of the BRT tree is an empty node, label entity. We use the Breath First Traversal algorithm to extract concepts and their corresponding hyponyms from the BRT tree.[2]

For a given concept in the BRT tree, we consider the concepts in the immediate child nodes, and extract the corresponding hypernyms and the hyponyms. Consequently, we can build the relationships of hypernyms and hyponyms for the concept.

5 Evaluation Matrix and Result

Our system is ranked fourth among the comparative evaluation final ranking of the task participant.[3] The table below shows the performance of participants' system based on average precision (Avg. P), recall (Avg. R), and average F-score measure (Avg. F) for the taxonomy extraction.

Participant	Rank	Avg. P	Avg. R	Avg. F
INRIASAC	1	0.1721	0.4279	0.2427
LT3	2	0.3612	0.6307	0.3886
ntnu	**4**	**0.1754**	**0.2756**	**0.2075**
QASSIT	5	0.1563	0.1588	0.1575
TALNUPF	6	0.0720	0.1165	0.0798
USAARWLV	3	0.2014	0.3139	0.2377

Table 1. Comparative evaluation results for SemEval-2015 Task 17, showing our system result in bold letters.

The evaluation tool measures a system-generated taxonomy against the gold standard taxonomy by comparing the following items:[4]

- The overall structure of the taxonomy against a gold standard, with an approach used for comparing hierarchical clusters.
- Structural measures.
- Manual quality assessment of novel edges.

In comparison against the gold standard data, the system's average performance under certain domain terms (chemical (CH), equipment (EQ), food and science (SC) domains) with respect to vertices in common, edge coverage and ratio of novel edges are shown in the table below.

[2] BFT is efficient in traversing a tree level by level and from left to right *http://en.wikipedia.org/wikiTree_traversal*

[3] http://alt.qcri.org/semeval2015/task17/index.php?id=evaluation

[4] http://alt.qcri.org/semeval2015/task17/index.php?id=evaluation

Features	CH	EQ	Food	SC
Vertices in coverage	0.3149	0.3144	0.3165	0.4390
Edge coverage	0.2803	0.2331	0.2603	0.3287
Ratio of Novel edges	0.4198	1.3419	1.0264	0.8584

Table 2. System's comparison against gold standard data.

In the table, the feature "vertices in coverage" represent the ratio of number of vertices in common with the gold standard taxonomy to the number of the gold standard vertices. The feature "edge coverage" is the fraction of number of edges in common with the gold standard over the number of edges in the gold standard. The ration of the product of the number of taxonomy edges and the number of edges in common with the gold standard to the number of gold standard edges is represented by "Ratio of Novel edges" in the result in Table 2.

From Table 2, it can be observed that, the system has the best and the worst performance in taxonomies for the science and equipment domains respectively. The bases of these differences in the system's performance are its precision for individual domain against its gold standard. For instance, from 452 vertices for the gold standard science domain from the taxonomy of fields and their subfields, the system was able to extract 338 vertices. Furthermore, the system's cumulative measure of the similarity against the gold standard is affected by the precision rate. For instance, in the worst performance for the gold standard domain of material handling equipment combined with IS-A relations from WiBi (Flati *et al.*, 2014), our system has a precision of 1.61% as shown in the evaluation result[5] while SC has good results in edge and vertex retrieval due to the good cumulative results.

6 Conclusion

In this paper, we present an approach for the unsupervised knowledge extraction for taxonomies of concepts using WordNet and Wikipedia as the sources of information. We first induce the construction of binary tree structures for each term in the domain using the extracted hypernym and hyponym. From the set of binary trees, we attempt to construct a BRT tree for the taxonomy extraction.

We regard this work as initial, as there is some improvement space to be made as well as many related areas to look into. First, in any future work we will investigate what better evaluation framework we can propose for the system. Second, we would like to give more attention to optimize the system result to a more formalized taxonomy. Third, we would like to include more concepts of relatedness extraction to obtain the stronger features.

Acknowledgements. Research of this paper was partially supported by Ministry of Science and Technology, Taiwan, under the contract MOST 103-2221-E-003-014.

References

Barbu, Eduard and Poesio, Massimo. (2009). Unsupervised Knowledge Extraction for Taxonomies of Concepts from Wikipedia. *RANLP-2009*, pp. 28-32. Borovets, Bulgaria.

Blundell, Charles, Teh, Yee Whye, and Heller, Katherine A. (2012). Bayesian Rose Trees. *DBLP, abs/1203.3468*.

Bordea, Georgeta, Buitelaar, Paul, Faralli, Stefano, and Navigli, Roberto. (2015). Semeval-2015 task 17: Taxonomy Extraction Evaluation (TExEval). *Proceedings of the 9th International Workshop on Semantic Evaluation.*

Flati, Tiziano, Vannella, Daniele, Pasini, Tommaso, and Navigli Roberto. (2014). Two Is Bigger (and Better) Than One: the Wikipedia Bitaxonomy Project. *ACL 2014*, Baltimore, Maryland, USA June 22-27, 2014.

Gabrilovich, Evgeniy and Markovitch, Shaul. (2007). Computing Semantic Relatedness Using Wikipedia-based Explicit Semantic Analysis. *IJCAI-07*, pp. 1606-1611.

Liu, Xueqing, Song, Yangqiu, Liu, Shixia, and Wang, Haixun. (2012). Automatic Taxonomy Construction from Keywords. *ACM.*

Miller, George A. (1995). WordNet: A Lexical Database for English. *Communications of the ACM*, 38(11), 39-41.

[5] http://alt.qcri.org/semeval2015/task17/index.php?id=evaluation

Song, Yangqiu, Wang, Haixun, Wang, Zhongyuan, Li, Hongsong, and Chen, Weizhu. (2011). Short Text Conceptualization Using a Probabilistic Knowledgebase. *Proceedings of the Twenty-Second international joint conference on Artificial Intelligence*, Vol. 3, pp. 2330-2336, AAAI Press

Tuan, Luu Anh, Kim, Jung-jae, and Kiong, Ng See. (2014). Taxonomy Construction Using Syntactic Contextual Evidence. *EMNLP-2104*, pp. 810-819.

Velardi, Paola, Faralli, Stefano, and Navigli Roberto. (2013). OntoLearn Reloaded: A Graph-Based Algorithm for Taxonomy Induction. *Computational Linguistics*, 39(3), 665-707.

Wu, Fang, Lu, Zhao, Yan, Yu, and Gu, Junzhong. (2009). Measuring Taxonomic Relationships in Ontologies Using Lexical Semantic Relatedness. *ICADIWT'09. Second International Conference on the. IEEE, 2009*, pp. 784-789.