Motif-Based Hyponym Relation Extraction from Wikipedia Hyperlinks

Dr.Y. Kalpana and S. Mahalakshmi

Abstract--- Discovering hyponym relations among domain-specific terms is a fundamental task in taxonomy learning and knowledge acquisition. However, the great diversity of various domain corpora and the lack of labeled training sets make this task very challenging for conventional methods that are based on text content. The hyperlink structure of Wikipedia article pages was found to contain recurring network motifs in this study, indicating the probability of a hyperlink being a hyponym hyperlink. Hence, a novel hyponym relation extraction approach based on the network motifs of Wikipedia hyperlinks was proposed. This approach automatically constructs motif-based features from the hyperlink structure of a domain; every hyperlink is mapped to a 13-dimensional feature vector based on the 13 types of three-node motifs. The approach extracts structural information from Wikipedia and heuristically creates a labeled training set. Classification models were determined from the training sets for hyponym relation extraction. Two experiments were conducted to validate our approach based on seven domain-specific datasets obtained from Wikipedia. The first experiment, which utilized manually labeled data, verified the effectiveness of the motif-based features. The second experiment, which utilized an automatically labeled training set of different domains, showed that the proposed approach performs better than the approach based on lexico-syntactic patterns and achieves comparable result to the approach based on textual features. Experimental results show the practicability and fairly good domain scalability of the proposed approach.

Keywords--- Hyponym Relations, Taxonomy Learning, Motif-based Features, Lexico-Syntactic Patterns

I. INTRODUCTION

Wikipedia has become a popular data source in hyponym relation extraction research. Several such studies adopted the syntactic-pattern-based methods or textural feature- based machine learning methods. When shifting to a new domain, these methods require new syntactic patterns to be learned or new training samples to be manually constructed, which usually entail high labor costs. In addition, these methods do not fully utilize the topological structure of hyperlinks in Wikipedia article pages. Wikipedia article pages are a domain-specific term. Hyperlinks and article pages are consider as directed graph named Wikipedia article graph. Network motifs appear much more frequently in a specific network than in randomized networks. Generating training sets with minimal human involvement for extraction in the different domains is very difficult. Overcome these problem we propose Motif based hyponym Relation Extraction for hyponym relation discovery. This focuses on hyponym relation discovery from academic domains in Wikipedia, such as Data mining, Classical mechanics and Microbiology. We find thatthe navigation box, a type of structural information in Wikipedia article pages, contains plenty ofDomain-specific hyponym relations that can be automatically extracted. Navigation boxes and the Wikipedia category structure extracted from Wikipedia category pages

Dr.Y. Kalpana, Associate Professor, VELS University, Pallavaram, Chennai, Tamil Nadu–600117, India
S. Mahalakshmi, Research Scholar, VELS University, Pallavaram, Chennai, Tamil Nadu–600117, India

II. LITERATURE SURVEY

Kashtan et al., present a systematic approach to define ’motif generalizations’: families of motifs of different sizes that share a common architectural theme. To define motif generalizations, they first defined ’roles’ in a subgraph according to structural equivalence. For example, the feedforward loop triad, a motif in transcription, neuronal and some electronic networks, had three roles, an input node, an output node and an internal node. The roles were used to define possible generalizations of the motif. The feedforward loop can have three simple generalizations, based on replicating each of the three roles and their connections. They present algorithms for efficiently detecting motif generalizations.They find that the transcription networks of bacteria and yeast display only one of the three generalizations, the multi-output feedforward generalization. In contrast, the neuronal network of *C. elegans* mainly displays the multi-input generalization. Forward-logic electronic circuits display a multi-input, multi-output hybrid. Thus,networks which share a common motif can have very different generalizations of that motif. Using mathematical modelling, we describe the information processing functions of the different motif generalizations in transcription, neuronal and electronic networks.

Roberto proposed Word- Class Lattices (WCLs), a generalization of word lattices that were use to model textual definitions. Lattices are learned from a dataset of definitions from Wikipedia. This method is applied to the task of definition and hypernym extraction and compares favorably to other pattern generalization methods .

Andrew et al., proposed an approach and a set of design principles for such an agent, and described a partial implementation of such a system that had already learned to extract a knowledge base containing over 242,000 beliefs with an estimated precision of 74% after running for 67 days, and discussed lessons learned from this preliminary attempt to build a never-ending learning agent.

Marti A. Hearst, described a method for the automatic acquisition of the hyponymy lexical relation from unrestricted text. Two goals motivate the approach: (i) avoidance of the need for pre-encoded knowledge and (ii) applicability across a wide range of text. They identified a set of lexico- syntactic patterns that are easily recognizable, that oc- cur frequently and across text genre boundaries, and that indisputably indicate the lexical relation of inter- est. They describe a method for discovering these pat- terns and suggest that other lexical relations will also be acquirable in this way. A subset of the acquisition algorithm was implemented and the results were used to augment and critique the structure of a large hand-built thesaurus. Extensions and applications to areas such as information retrieval were suggested.

DongHyun Choi et al., proposed a novel algorithm to extract taxonomic (or isa/instanceOf) relations from category structure by classifying each category link. Previous algorithms mainly focued on lexical patterns of category names to classify whether or not a given category link is an isa/instanceOf. In contrast, this algorithm extracted intrinsic properties that represent the definition of given category name, and they used those properties to classify each category link. Experimental result showed about 5 to 18% increase in F-Measure, compared to other existing systems.

III. PROJECT DESCRIPTION

The proposed method analyzes sensitive information to the public in online social networking. Problem of this whether many vertices of the same degree tend to gather in the same dense sub graph(community). Note that if an attacker finds all the vertices of a particular degree appearing in a certain sub graph (community), he can obtain the privacy information such as the neighborhood and connectivity properties of a target.

IV. METHODOLOGIES

A. Admin

a. Authentication

If a new user is going to access the network then he/she has to register first by providing necessary details. After successful completion of sign up process, the user has to login into the application by providing username and exact password. The user has to provide exact username and password which was provided at the time of registration, if login success means it will take up to main page else it will remain in the login page itself.

b. File Upload

After complete the authentication process admin have to uplodas the file.this file Should be in domain oriented.This file is used for user search as welll as generating topics to search.

c. Hyperlink Generation

If admin uploads file the next process is admin have to generate the hyperlinks for user search.Select the file and the view all words from selected file and choose the domain relatd words to create hyper links.Admin Have to upload file for genmerated hyperlinks.

d. Hyponym Relation Extraction

Extract the informtions for selected hyperlinks and views the domain related informations. Filters the unwanted informations to show the user.

B. User

a. Query Search

Based on the user query in the basis of keyword shows the related contents with hyperlinks.And redirects related informations for the clicked hyperlinks.

b. Topic Search

All the Uploaded documents can show in the format of titles based on the user request shows the particular content with hyperliinks.

V. TECHNIQUE USED

A. Hyponym Relation Extraction

Automatically construct motif-based features from the WAG of a domain; Extract structural information from Wikipedia and heuristically label the training set of the domain based on the extracted structural information; learn the classification model from the training sets to discover hyponym relations.

B. Algorithm

Step 1: Extracting Category Forest and Navbox Forest;

Step 2: TS =Ø; TS =Ø;

Step 3: Selecting an edge e = ‹vx,vy› ϵ ES; ES=ES-{ e };

Step 4: If RDNF (υx ,υy) TSp = TSp∪{e};

Step 5: If RDCF (υx ,υy)=hyponym relation

Else if RDCF (υx ,υy) =other relation TSn = TSn∪{e};

Step 6: If |TSp∪TSn|< σ Goto STEP3;

Step 7: Return TSp, TSn.

VI. CONCLUSION

In this paper, the experiment results showed that our approach Works well in domains where there are enough hyponym relations. This approach may not work well in a domain where the hyponym relations among domain-specific terms are very sparse, such as human individuals or companies.

REFERENCE

[1] U.Alon, “Network Motifs: Theory and Experimental Approaches, ”Nature Reviews Genetics, vol.8,no.6,pp.450461, 2007.

[2] C.Andrew, B. Justin, K. Bryanetal., “Toward an Architecture for Never ending Language Learning,” Proc.24th National Conf. Artificial Intelligence (AAAI10),pp.13061313,2010.

[3] N.Y.Asuka Sumida, and K.Torisawa, “Boosting Precision and Recall of Hyponymy Relation Acquisition from Hierarchical Layouts in Wikipedia,” Proc. Intl Conf. Language Resources and Evaluation (LERC08),pp.24622469,2008.

[4] J. Carletta, “Assessing Agreement on Classification Tasks: the Kappa Statistic,” Computational Linguistics, vol. 22, no. 2, pp. 249254,1996.

[5] S. Chernov, T. Iofciu, W. Nejdl et al., “Extracting Semantic Relationships between Wikipedia Categories,” Proc. 1st Intl Workshop on From WikitoSemantics”,2006.

[6] D.Choi, E.K. Kim, S.A. Shimetal., “Intrinsic Property based Taxonomic Relation Extraction from Category Structure, ”Proc. 6th Workshop on Ontologies and Lexical Resources,pp.4857,2010.

[7] O. Etzioni, M. Banko, S. Soderland et al., “Open Information Extraction from the Web,” Communications of the ACM, vol. 51, no.12, pp.6874,2008.

[8] A. Fader, S. Soderland, and O. Etzioni, “Identifying Relations for Open Information Extraction,” Proc.Conf. Empirical Methods in Natural Language Processing (EMNLP11), pp. 15351545,2011.

[9] M.A.Hearst, “Automatic Acquisition of Hyponyms from Large Text Corpora,” Proc. 14th Conf. Computational Linguistics, pp. 539545,1992.

[10] D. Jurgens, and T.C. Lu, “Temporal Motifs Reveal the Dynamics of Editor Interactions in Wikipedia,” Proc. 6th Intl Conf. Weblogs and SocialMedia, 2012.