Volume 3, Issue 12, December 2013 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

A Survey on Discovery of Part-Whole Relations with Knowledge-Base

Claudia Reynolds, G. Naveen Sundar
Department of Computer Science and Engineering,
Karunya University, India

Abstract— Mining the data from large-plain text enhances the retrieval of information from resources. This paper describes about the automatic discovery of part-whole patterns from the texts using knowledge. The Parts are found by learning semantic constraints and linking documents to the knowledge-base. Knowledge Discovery in text is a potential method, which automatically extracts the concepts and conceptual relationships from text. The importance of part-whole relation in various application domains has been analyzed. In addition, performance can be improved by using Wikipedia as a knowledge-base, and the effect of semantic-drift is eliminated.

Keywords: data mining, part-whole relation, knowledge-base, text mining

I. INTRODUCTION

Data mining is the process of uncovering the hidden patterns, by applying methods to the large data sets. It is an approach, in which the data is stored and indexed in databases, for executing and discovering the algorithms, which is quite efficient. Web Data Extraction is the retrieval of data from web, to be utilized in future. Information Extraction algorithms objective is to discover part–whole relations from large, broad-coverage, general-purpose corpora. Supervised approach generates a training corpus with a possible number of positive and negative examples of part-whole relations and the classification rules induced over the training data. In the absence of labeled training data, supervised approaches are unsuitable for the task of mining domain-specific part-whole relations from specialized texts.

The minimally-supervised, is an approach for learning part–whole relations from texts. Using Wikipedia as a source of knowledge-base, a set of reliable patterns acquired which express the part–whole relations [5]. Then, we use that patterns acquired to extract part–whole relation triples from a collection of sparse and domain specific texts. This paper is organized as, Section 2 includes related works of this paper and section 3 gives a comparative study.

II. RELATED WORKS

A. Web Mining

Web Mining is the data mining application, helps to find the patterns from the web. It can be categorized into Web Content Mining, Web Structure Mining and Web Usage Mining. Web Content Mining is the extracts the contents from the web documents, Web Structure Mining is used to analyzes the structure of the link and preferable structure of the document from the web and the Web Usage Mining is to find the useful information from the behavior of the users, when interacting with the Internet.

B. Text Mining

Automatic extraction of information from a large source of different unstructured textual resources, which is unknown, can be termed as Text Mining.

C. Template Extraction

Templates are provided with reliable structures and the web contents can be accessed easily. It improves the performance of search engines, clustering and web documents. Clustering of the similar templates from the heterogeneous documents can be useful in extraction of templates simultaneously.

D. Knowledge-base

Knowledge-base improves the performance of minimally-supervised algorithms and plays a major role in alleviating data-sparsity issue. Most of the Information Extraction (IE) algorithms are based on the Knowledge sources. Wikipedia as a knowledge-base gains importance in IE community.

III. COMPARATIVE STUDY

This section includes a study on the part-whole relation using knowledge base to improve the performance and the extraction of relation triples from a large set of corpus.

A. Finding Parts from Corpora

The method for extracting parts from wholes [2]. In a very large corpus, this method finds out the part words, for about top 50 words with 55% accuracy, as per the system. The part list is scanned by the user and added to an existing ontology i.e., WordNet or, a part of semantic lexicon. The part of relation is found between words. WordNet is a part-list, which can be used by the user to scan and mark the list of part words.

The first goal of the method is that finding the lexical patterns that are likely to indicate part-whole relations. The LDC North American News Corpus (NANC) is a compilation of the several US newspapers wire output and it is about 100,000,000 words of the total corpus. Run the program on the entire data set, which takes approximately four hours on network. The program consists of three phases: First identifies and records all occurrences of patterns. Next, it filters out all words ending with ing, ness, or, ity. At last, order the achievable parts by the likelihood that they are considered to be true parts, according to some suitable metric. The program developed, had some affinity to find qualities of objects. More of the data allows to produce better lists. It is more accurate and the larger numbers would result in finding the other reliable indicators.

B. Learning Part-Whole Relations

A method for learning part–whole relations [1], is important in many domains, but in general receives less attention than the subsumption relation. A method to find the part-whole relations is by finding out the phrase patterns for both explicit and implicit part-whole relations. Applying these patterns, the part-whole relation instances can be achieved. Part-Whole is the central most structuring principle in artifact design. The main aim is to develop a method for learning the part-whole relations from existing vocabularies and text sources. A method to learn part-whole relations by first learning phrase patterns which connects the parts to wholes from training set of known part-whole pairs using a search engine. The patterns are used for finding the new part-whole relations with the help of search engine.

In a training set, a search query is constructed for each part-whole pair. The phrases are collected from the query result and the patterns are obtained. Sort the patterns in order by frequency. Each pattern is filled with parts from a set and the phrases are collected, from which the part-whole pairs are extracted. The pairs are constrained with wholes and sorted by frequency.

C. Textractor

Textractor a framework [3] is used for extracting significant domain concepts from irregular corporate textual datasets. Various information extraction (IE) systems used for corporate exist. Conversely, none have target the product development and/or customer service domain, in spite of its significant application potentials and benefits. This domain poses new scientific challenges i.e., lack of exterior knowledge resources, and irregularities like ungrammatical constructs in texts, which negotiate successful information extraction. The Textractor is developed to address these issues. Textractor is an application, for extracting related concepts accurately from irregular textual narratives in datasets..

Evaluations on real-life corporate data exposed that Textractor extracts domain concepts, as a single or, multi-word terms in ungrammatical texts, with high precision. IE techniques could serve as a forerunner to business intelligence (BI) by extracting applicable concepts that pertaining to soft failures from textual data. The pre-processing of data is proceeded by identifying, selecting and extracting the candidates of the datasets.

The approach begins with data pre-processing, which lowers the noise level, standardizes the languages for multi-lingual texts, and performs basic operations on the textual data. Candidate Concept Identification recognizes terms in text based on contiguous sequences of parts-of-speech (PoS)-tags, called term signatures.

Concept Selection and Extraction stage applies two statistical filters to address the short-comings of the previous linguistic filter, to categorize between valid and invalid terms, and to measure the terms relevancy. This knowledge allows us to develop better quality products and ensuring financial returns in an organization. The algorithms are generic to be applicable in other corporate or, open domains, for dealing with extracting multiword terms from ungrammatical texts, such as from online forums and blogs.

D. Extraction of Part-Whole Relations using knowledge-base

It is an approach for the extraction of high-quality domain-specific part–whole relations from sparse texts [5]. The core is based on applying a minimally-supervised algorithm to a large corpus, which is used as knowledge-base. From this knowledge-base, a set of patterns is acquired which express part-whole relations. Then, from a domain-specific collection of texts all triples consisting of the acquired patterns and the instance pairs they connect are extracted, given in Fig. 1. These relation triples are the domain-specific part–whole relations. It is as an intense example of domain-adaption, since neither the source corpus, that is broad-coverage knowledge-base, nor the target corpus, that is domain-specific texts have been labeled with part–whole relations. The aim is to improve the quality of the product using pertinent information extracted from corporate data sources.

A domain in which the part–whole relation is of primary importance is the business or, corporate domain of Product Development/Customer Service (PD-CS). In this domain, part–whole relations extracted from customer complaint texts or repair notes of service engineers encode valuable operational knowledge that organizations can exploit for product quality improvement.

A minimally-supervised algorithm is used for improving the performance which relies on domain-specific knowledge-bases. This alleviate the need for labeled training data. The extraction is initialized with instance pairs, called

seeds, which denote part–whole relations, e.g., engine-car. This instance pairs is utilized to acquire patterns expressing part-whole relations, and in turn, employ the patterns to extract other instance pairs. Thus, it accurately extracts part–whole relations from sparse, domain-specific texts with minimal supervision.

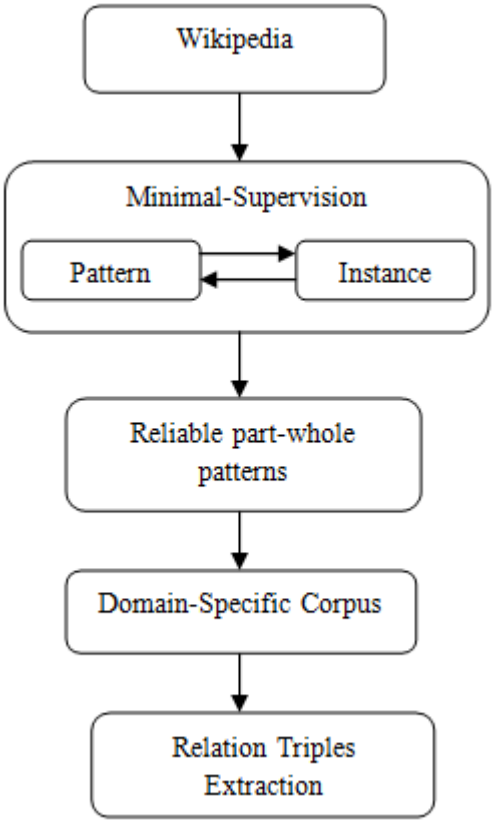

Fig. 1 Architecture of Part-Whole Pattern Extraction

E. Wikify

Wikify! Linking documents to encyclopedic knowledge [4] introduces the use of Wikipedia as a resource for the extraction of keyword and word sense disambiguation automatically. The important concepts are identified from the input document by the system and link these concepts to the corresponding Wikipedia pages. Evaluation shows that the automatic annotations are reliable and hard to distinguish from the manual annotations. The task Text Wikification extracts the most important words and phrases in the document automatically, and identifying each keyword that suitable link to a Wikipedia article.

The Wikify! System integrates the keyword extraction algorithm that automatically find the important keywords in the input document, and the word sense disambiguation algorithm that assigns each keyword with the acceptable link to a Wikipedia article. The unsupervised keyword extraction algorithm implementation works in two steps, namely, candidate extraction and keyword ranking. The candidate extraction is used to extract all possible n-grams by parsing over the input document and present it in a controlled vocabulary. Numeric value is assigned to each candidate in the ranking step and a valuable key phrase is given to the candidate.

The Disambiguation Algorithms are based on knowledge methods. It depends on the knowledge derived from dictionaries and the data-driven algorithms. The best word sense disambiguation system has a score of 83.21% of precision and recall.

TABLE I COMPARATIVE STUDY OF PART-WHOLE RELATION USING KNOWLEDGE BASE

FACTORS	**Berland (1999)**	**Girju (2003)**	**Ittoo (2010)**	**Mihalcea (2007)**	**Ashwin Itto (2012)**
GOAL	finding parts in very large corpora	discovery of semantic relations from text	to address the issues of ungrammatical constructs	keyword extraction and word sense	improving the performance of extraction

LEARNING PROCEDURE	supervised	supervised	-	unsupervised	minimally-supervised
APPROACH	automatic acquisition	semi-automatic	automatic	automatic	automatic
METHOD	statistical or, corpus based techniques	C4.5 decision tree learning	Textractor	Text wikification	extraction of domain-specific part–whole relations
INPUT	NANC	meronymic pairs	unstructured format	document	Wikipedia
OUTPUT	possible parts	semantic constraints	business intelligence	links to the web pages	domain specific texts
PERFORMANCE	finds part words with 55% accuracy for the top 50 words	precision of 83%, recall of 98%	precision of 85.6%, recall of 93.6% and f-score of 84%	precision and recall of 83.21%	precision of 80.95% , recall of 75.91%
MERITS	good accurate result	better accuracy of about 83%	better quality products	perform significantly	alleviate semantic drift
DEMERITS	data sparsity and tagger mistakes	learning accuracy	average score obtained	turing –like test is hard	multi-lingual relation extraction

IV. CONCLUSION

In this paper, the part-whole relation triples extraction and knowledge base have been discussed. A comparative study has been made regarding the techniques, methodology, merits and its demerits. It is an efficient model for the part-whole patterns using wikipedia as a knowledge-base.

ACKNOWLEDGEMENT

This work has been done with the help of our Department of Computer Science and Engineering. We would like to thank for providing us the facilities and support.

REFERENCES

[1] Willem Robert van Hage, Hap Kolb, Guus Schreiber, *A method for learning part–whole relations*, The Semantic Web, Lecture Notes in Computer Science, Vol. 4273, 2006, pp. 723–735.

[2] Matthew Berland, Eugene Charniak, *Finding parts in very large corpora*, 37th Annual Meeting of the ACL, 1999, pp. 57–64.

[3] Ashwin Ittoo, Laura Maruster, Hans Wortmann, Gosse Bouma, *Textractor: a framework for extracting relevant domain concepts from irregular corporate textual datasets*, Lecture Notes in Business Information Processing 47, 2010, 71–82.

[4] Rada Mihalcea, Andras Csomai, *Wikify! Linking documents to encyclopedic knowledge*, 16th Conference on Information and Knowledge Management, 2007, pp.233–242.

[5] Ashwin Ittoo, Gosse Bouma in *Minimally-supervised extraction of domain-specific part-whole relations using Wikipedia as knowledge-base, at SciVerse ScienceDirect*, Data & Knowledge Engineering, 2012, 85 57-79.

