International Journal of Recent Engineering Research and Development (IJRERD)
Volume No. 01 – Issue No. 03, ISSN: 2455-8761
www.ijrerd.com, PP. 22-26

Extraction of Textual contents by applying Part Whole Relation with Domain Adoption Technique

O D Joshi[1], U A Nuli[2]

[1]Assistant Professor, Department of Computer Science and Engineering,
Sharad Institute of Technology,College of Engineering.Yadrav
Maharashtra-416121, India

[2]Assistant Professor, Department of Computer Science and Engineering,
DKTE's Textile and Engineering Institute, Ichalkaranji
Maharashtra-416115, India

Abstract: Extraction of knowledge from unstructured textual data is still a key challenge in information retrieval. Important limitation of traditional and widely used Keyword search method is it cannot extract semantically relevant data.This paper discusses the approach of text extraction techniques applicable for any specific domain and has capability to produce more semantically relevant data.

This paper is focused at design of a text extraction method based on part whole relation extraction technique, which provides reliable and semantically relevant text extraction. The required text is extracted by identifying the proper word and its relation with the statements..

Keywords:NLP, Part-whole relation, Data Mining, Abstraction, Text Mining, Text Analysis

1. Introduction

Information Extraction algorithms have been introduced for extracting occurrences of the particular words from the textual statements. There are many approaches are available to extract the information from the textual statements like part whole relation extraction from the statementsby which the extraction of textual information is taken place. The extraction of part whole relation with the pattern are known as triple formulation (part-pattern-whole) for e.g blown fuse found in set top box is called as triple relation (blown fuse(part)- found in (pattern)- set top box (whole)). It is equivalent to text analysis, considered as the process of analyzing high-quality information from text. Text mining usually involves the process of structuring the input text, deriving patterns within the structured or unstructured data, and finally evaluation and interpretation of the output. Text analysis involves information retrieval, lexical analysis to study frequency distributions, pattern recognition, tagging, annotation, information extraction, data mining techniques that includes link and association analysis, visualization, and predictive analytics.

The existing algorithms have mostly focused on generalized text collections i.e. not a domain specific textcollection.The relations between part and the whole in text mining for example the relation between the part"(blown) fuse" and it's whole "set-top box" establish the part whole relation by the pattern found in as blown fuse found in set top box.

The information expressed by part–whole relation is valuable to support a wide variety of corporate activities, such as Enterprise Intelligence or Product Development. The extraction of part–whole relations from domain-specific text is challenging because the unavailability of labeled domain-specific data for supervised algorithms. They are initialized with instance pairs known as seeds, which instantiate the target relation i.e. part–whole. It is proven that the performance of statistical NLP techniques, such as minimally-supervised algorithms, can be improved by leveraging upon knowledge-bases. However, domain-specific knowledge-bases are scarce in most disciplines. The WIKIPEDIA corpus deal with broad-coverage contents, their exploitation to support domain-specific relation extraction has not yet been investigated. Therefore there is need to develop and present an approach for extracting high-quality domain-specific part–whole relations from sparse texts, typically generated in specialized disciplines known as domain adoption.

2. Motivation

Many existing system predominantly focused on extraction procedure of part whole relations from the broad knowledge corpus. Due to lack of proper procedures for seed selection as well as domain adoption with same seed, the part whole relations extracted by the seed are deviate expected results. So, there is need to develop a system that will focus on formation of quality seed that results into formation of expected part whole relations and help to achieve domain adoption with same seed as well as part whole relations. This is achieved by system which classifies the words from the statements of WIKIPEDIA. This will aims to improve the quality of seed, that improves the quality of part-whole relation for broad as well as particular domain knowledge corpus.

3. Literature Review

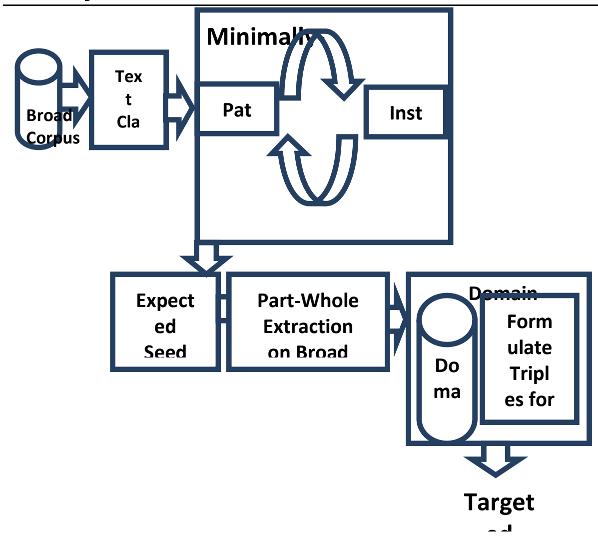
From last few decades, many approachesavailable for extraction of part whole relationfrom the text.But these methods are not adequate to extract the quality part whole relations as well as part-whole relation for specific domain knowledge base as every approach having some of its own drawbacks. The following discussion comprises about some of the approaches those are relate with part whole extraction techniques and domain extraction techniques so far,In Natural Language Processing, primary block of Semantic relations is more important, such as ontology learning/building described in [1] and question–answering systems for the betterment of text mining is described in [2,3].In the approach of R. Mihalcea et al[3], Existing Information Extraction (IE) algorithms have predominantly focused on discovering part–whole relations from large, broad-coverage (general-purpose) corpora, such as the L.A. TIMES collection. However, the rapid proliferation of textual data across virtually all disciplines has highlighted the pressing need for domain-specific IE (and NLP) techniques. A domain in which the part–whole relation is of fundamental importance is the business/corporate discipline 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.In the approach of C.Legg et al.[4] Minimally-supervised approaches alleviate the need for labeled training data. They are initialized with instance pairs, called seeds, which denote part–whole relations, e.g. “engine-car”. They use the instance pairs to acquire patterns expressing part–whole relations, and in turn, employ the patterns to extract other instance pairs.The principal distinction in the taxonomy of Keet and Artale [6] (Keet's taxonomy) is between transitive or mereological part–whole relations and their intransitive or meronymic counterparts [5,7]. Subsequent distinctions are made by enforcing semantic selectional restrictions, in the form of DOLCE ontology [8] classes, on the instances that participate in the different relation types. For example, in Keet's taxonomy, the sub-quantity-of part–whole relation can only connect instance pairs that belong to the class AMOUNT-OF-MATTER in the DOLCE ontology, such as the pair “alcohol-wine”.

Investigations on the part–whole relations have spanned over a many of the disciplines, including philosophy, cognitive science and conceptual modeling all the details about part-whole relations are discussed in [19-11]. In the approach of Girju et al. [12,13], patterns expressing part–whole relations between the various concept pairs are manually extracted from sentences of the Loss Angles TIMES and the SEMCOR corpora, and used to generate a training corpus with positive and negative examples of part–whole relations. Classification rules induced over the training data achieve a precision of 80.95% and a recall of 75.91% in identifying part–whole relations in previously unseen texts.In the approach of AshwinIttoo et al. [14], it extract the part whole relations from textual statements of Wikipedia as knowledge base, aims to form reliable patterns that express part whole relations. It applies the domain adoption technique on Wikipedia and transforms the broad knowledge corpus to the domain specific knowledge corpus.

By above discussion, it is clear that there are many approaches available for extraction of part whole relations from the broad knowledge as well as particular domain knowledge corpus. Minimally supervised approach help to provide the seed without a need of labeled data. Seed selected by minimally supervised approach extracts part whole relations from the broad knowledge corpus also going to applied on domain knowledge corpus, due to which results of part whole relation deviate from the expected result. Proposed system will going to classify the statements into words which helps to acquire the knowledge of statement. Due to the proper knowledge about the statement, it will help to improve the quality of the seed as well as the part whole relation extraction from the statement. While transforming the part whole relation from broad knowledge corpus to domain knowledge corpus, meaning of the part whole relation will change. So to overcome this problem, proposed system will going to verify the part whole relations extracted from the broad knowledge corpus after the domain adoption takes place. It ensures the quality of the part whole relations for domain adoption.

4. Proposed Work

For improving the quality of seed, it is necessary to classify the statement which improves the quality of part whole relation. The extracted part whole relation of broad knowledge corpus changes its meaning when they apply on domain knowledge corpus. There is need of system that will perform text classification, domain adoption technique and verification of part whole relations after completion of domain adoption. The proposed will comprises the following work, extraction of part-whole relations in easy manner, better knowledge of the statements for seed selection and domain adoption from broad coverage knowledge base. The proposed work comprises the following modules namely Text Classification, Seed Formulation, Final Seed, Domain Driven Tool, Domain Corpus and Triple Formulation.



Architecture for proposed work

The architecture of the proposed system comprise following modules,

4.1 Text Classification
4.2 Seed Formulation
4.3 Final Seed
4.4 Domain Adoption Tool
   4.4.1 Domain Corpus
   4.4.2 Triple Formulation.

The following section will discuss the proposed modules in detail.

4.1 Text Classification

The Syntactic patterns, derived from the dependency trees of syntactically parsed sentences, have been proposed as a solution to overcome the limitations of surface patterns. To provide relation in the form of triples, minimally-supervised is used. It uses pattern selection and instance selection procedures to form the particular seed. In the dependency tree of a sentence, nodes correspond to the tokens. They are connected by edges, depicting their syntactic associations, e.g. subject-verb-object This module consider Wikipedia statement as input and provide text classification of the words from the statement will be classified according to the statement like noun, verbs, adjectives etc as output. This will results into improvement of seed selection procedure.

4.2 Seed Formulation

Pattern Induction and formalization transforms the sentences of WIKIPEDIA knowledge base into corresponding relation triples, consisting of a pairs of instance and patterns that connect them in sentence. After understanding the exact meaning of the words from statement, it is beneficial for the pattern formalization procedure to find out the seed. Seed is nothing but the pattern from which the actual relation of part whole will be established. For e.g. the relation between the part "(blown) fuse" and its whole "set top box", established by pattern "found in", as in blown fuse found in set top box. In Pattern Induction and Formalization, it transforms the sentences from WIKIPEDIA knowledge-base into corresponding relation triples. It consists of a pair of instances and the patterns that connect them in the sentences. This module converts the unstructured texts of WIKIPEDIA statements into a structured statements. In Instance Selection phase, most reliable patterns previously identified, and extract the instance pairs that they sub-categorize in our WIKIPEDIA knowledge-base, these instances are likely to instantiate part–whole relations. They will be selected and fed back to the Pattern Selection phase for identifying other reliable patterns useful for creating part-whole-relation. Therefore, to extract the most accurate patterns, here it ensures that only the valid and most reliable part–whole instance pairs are selected. After analyzing the each statement, the seed generated on broad coverage corpus will applied on each statement to formulate the triple.

4.3 Formulation of Triples

Pattern Selection and instance selection are the procedures, by which the relations are extracted from the sentences. By considering the seed from last module, the same seed will be applied on the each statement of WIKIPEDIA to form the part and whole relations. To overcome the issue of semantic-drift, there is need of approach that determine whether an instance pair instantiates a valid part–whole relation. Only those valid part–whole pairs are then selected and fed back to promote the selection of reliable part–whole patterns. In this approach, it is achieved by "instance_pair_purity measure", It differs from the pair-pattern association strength. It does not merely take into account the association of instance pairs with patterns, some of which may be ambiguous as discussed previously. Instead, it estimates the degree to which an instance pair genuinely instantiates a valid part–whole relation, i.e. the "purity" of the pair. This principle posits that pairs, which are sub-categorized by the same patterns, instantiate the same semantic relations. This module will provide the part whole relations on the broad coverage corpus.

4.4 Domain Adoption Tool

Textractor algorithm is used to transform the broad knowledge corpus to the domain specific knowledge corpus. The Domain Adoption tool in the proposed system comprises three modules namely, Domain corpus, which is responsible for the conversion of broad knowledge corpus to the domain specific knowledge corpus. Second, after finalizing the seed from the minimally supervised algorithm that seed need to applied on domain corpus to form the part whole relations and finally the output generated after triple formulation, is going to verify by the instance selection. It is because the seed that select at initial time is on broad coverage knowledge corpus, now same seed applied on domain driven knowledge then it may impact on triple formulation.

4.4.1 Domain Corpus

The part whole relations formed by considering broad corpus will change the meaning of actual result when they are applying on particular domain, therefore selected seed again need to apply on particular domain. This module accepts input as part whole relations on broad coverage corpus. The relation extraction approach in this module operates in the area of domain-adaptation. In domain-adaptation, the knowledge acquired from a source domain/corpus is applied to a target domain/corpus. In this approach, the source corresponds to WIKIPEDIA knowledge-base, the knowledge acquired refers to the reliable part–whole patterns that extract from its contents during Pattern Selection, and the target corresponds to expected domain-specific, sparse texts. By comparing previous studies in domain-adaption, it can be innovate in three fundamental aspects. First, employ domain-adaption for extracting part–whole relations. Second, domain-adaptation is completely unsupervised, i.e. annotated data is available neither from the source nor from the target domains. Finally, source and target corpora exhibit completely contrasting contents. To identify domain-specific part–whole relations, it use the patterns acquired from the WIKIPEDIA knowledge-base, and extract the instance pairs they connect in the target corpus. The extracted domain-specific part–whole relations are formulated as triples. After completing the procedure of domain adoption module, it will get only those relations that are applicable to the specific domain and rest of the part whole relations are eliminated from the list.

4.4.2Domain Based Triples

After formulating the domain specific part whole relations, this module will again verifying the part whole relations. The seed formed in the last module applied on the broad coverage corpus and now it is applied on domain driven corpus. It may possible that the result of broad coverage corpus and result of domain driven corpus may deviate with same seed. It applies the same seed on the domain driven corpus to form the triples for the domain knowledge corpus. Domain Driven tool will analyze the seed along with the part whole relation and get the triples in the format of part-seed-whole. Here seed selection is carried out by the Minimally supervised algorithm for domain knowledge base. But it may possible that seed which is suitable for broad knowledge corpus is not best suited for domain knowledge corpus. This module will formulate the triples from the seed that avail from the selection procedure of pattern selection and insert selection for particular domain.

5. Requirements

Environment: - Java,PHP,XAMPP.
Data source: - Text documents from WIKIPEDIA
Tools: - Jdk1.7

References

[1]. G.A. Miller, C. Leacock, R. Tengu, R.T. Bunker, A semantic concordance, In: Proc. of the 3rd DARPA workshop on Human Language Technology, 1993, pp. 303–308.
[2]. M. Berland, E. Charniak, Finding parts in very large corpora, In: Proc. of the 37th Annual Meeting of the ACL, 1999, pp. 57–64.
[3]. R. Mihalcea, A. Csomai, Wikify!: linking documents to encyclopedic knowledge, In: Proc. of 16th Conference on Information and Knowledge Management,CIKM'07, 2007, pp. 233–242.
[4]. AGROV Thesaurus, Available from: http://aims.fao.org/website/AGROVOC-Thesaurus/sub
[5]. O. Medelyan, C. Legg, Integrating Cyc and Wikipedia: Folksonomy Meets Rigorously Defined Common-Sense, In: Proc. of the AAAI Workshop on Wikipedia and Artificial Intelligence: An Evolving Synergy, AAAI Press, Chicago, USA, 2008.
[6]. C.M. Keet, A. Artale, Representing and reasoning over a taxonomy of part–whole relations, Applied Ontology 3 (1) (2008) 91–110.
[7]. A. Ittoo, G. Bouma, On learning subtypes of the part–whole relation: do notmix your seeds, In: Proc. of the 48th Annual Meeting of the ACL, 2010, pp. 1328–1336.
[8]. A. Gangemi, N. Guarion, C. Masolo, A. Oltramari, L. Schneider, Sweetening ontologies with DOLCE, In: Proc. of Knowledge Engineering and Knowledge Management: Ontologies and the Semantic Web, 2002, pp. 223–233.
[9]. P.D. Turney, The latent relation mapping engine: algorithm and experiments, Journal of Artificial Intelligence Research 33 (1) (2008) 615–655.
[10]. G. Bouma, I. Fahmi, J. Mur, G. van Noord, L. van der Plas, Linguistic knowledge and question answering, TraitementAutomatique des Langues 2 (46) (2006) 15–39.
[11]. R. Mihalcea, A. Csomai, Wikify!: linking documents to encyclopedic knowledge, In:

Proc. of 16th Conference on Information and Knowledge Management,CIKM'07, 2007, pp. 233–242.

[12]. M.E. Winston, R. Chaffin, D. Herrmann, A taxonomy of part–whole relations, Cognitive Sciences 11 (4) (1987) 417–444.

[13]. P. Gerstl, S. Pribbenow, Midwinters, end games, and body parts: a classification of part whole relations, International Journal of Human Computer Studies 43(1995) 865–890.

[14]. AshwinIttoo, Gosse Bouma, Minimally supervised extraction of domain specific part-whole relations using Wikipedia as knowledge base.