Mining Ontological Knowledge from Domain-Specific Text Documents

Xing Jiang and Ah-Hwee Tan
School of Computer Engineering, Nanyang Technological University
Nanyang Avenue, Singapore 639798
{jian0008,asahtan}@ntu.edu.sg

Abstract

Traditional text mining systems employ shallow parsing techniques and focus on concept extraction and taxonomic relation extraction. This paper presents a novel system called CRCTOL for mining rich semantic knowledge in the form of ontology from domain-specific text documents. By using a full text parsing technique and incorporating both statistical and lexico-syntactic methods, the knowledge extracted by our system is more concise and contains a richer semantics compared with alternative systems. We conduct a case study wherein CRCTOL extracts ontological knowledge, specifically key concepts and semantic relations, from a terrorism domain text collection. Quantitative evaluation, by comparing with a state-of-the-art ontology learning system known as Text-To-Onto, has shown that CRCTOL produces much better precision and recall for both concept and relation extraction, especially from sentences with complex structures.

1. Introduction

An ontology is an explicit specification of a conceptualization [2], comprising a formal description of concepts, relations between concepts, and axioms about a target domain. Considered as the backbone of the Semantic Web [1], domain ontologies enable software agents to interact and carry out sophisticate tasks for users.

To reduce the effort of building ontologies, ontology learning systems have been developed to learn ontologies from domain relevant materials. However, most existing ontology learning systems focus on extracting concepts and taxonomic (IS-A) relations. For example, SymOntos [5], a symbolic ontology management system developed at IASI_CNR, made use of shallow NLP tools including a morphologic analyzer, a part-of-speech (POS) tagger and a chunk parser, to process documents and employed text mining techniques to produce large ontologies based on document collections. The concept extraction method was however domain-dependent and had limited applicability.

Text-To-Onto [4], also based on shallow NLP tools, was able to extract key concepts and semantic relations from texts. Selection of concepts was based on the tf/idf measure used in the field of information retrieval. Semantic relations were extracted using an association rule mining algorithm and predefined regular expression rules. However, as tf/idf was designed primarily for IR, the system extracted both domain-specific and common concepts. Also, the identification of semantic relations is based on POS tags, limiting the accuracy of the relations extracted.

Rajaraman and Tan [7] extracted knowledge in the form of concept frame graph (CFG) from text documents. Semantic relations between concepts were extracted through analyzing the POS tags of the sentences using a library of extraction rule. As the CFG system extracted concepts and relations from all sentences, it tended to extract a large number of concepts and relations, many of which had no real significance. Also, the CFG system was designed to extract non-taxonomic relations only.

In this paper, we present a novel system, known as Concept Relation Concept Tuple based Ontology Learning (CRCTOL) for mining rich semantic knowledge in the form of ontology from domain-specific documents. By using a full text parsing technique and incorporating statistical and lexico-syntactic methods, the knowledge extracted by our system is more concise and contains a richer semantics compared with alternative systems. We conduct a case study wherein CRCTOL extracts ontological knowledge, specifically key concepts and semantic relations, from a terrorism domain text collection. Quantitative evaluation, by comparing with the Text-To-Onto system, has shown that CRCTOL produces much better accuracy for concept and relation extraction, especially from sentences with complex structures.

The rest of the paper is organized as follows. Section 2 presents the CRCTOL system's framework. The algorithms for concept extraction and semantic relation extraction are described in Section 3 and 4 respectively. In Section 5, a case study of this system is presented.

Proceedings of the Fifth IEEE International Conference on Data Mining (ICDM'05)
1550-4786/05 $20.00 © 2005 IEEE

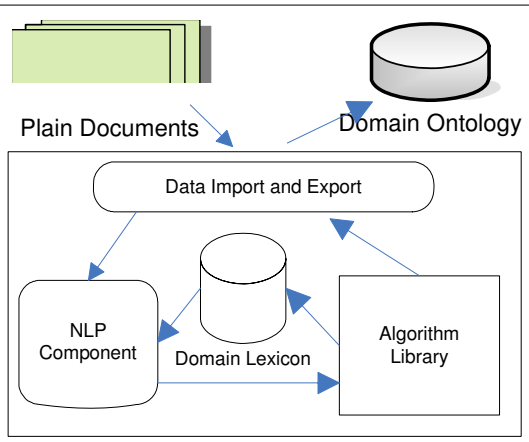

Figure 1. The CRCTOL system's architecture.

2. System Architecture

The CRCTOL system consists of three core components, namely *Natural Language Processing*, *Algorithm Library*, and *Domain Lexicon* (Figure 1).

Natural Language Processing (NLP): This component incorporates NLP tools, such as Eric Brill's POS tagger for attaching words with POS tags and Michael Collins's syntactic parser for parsing sentences. With the NLP component, we could utilize the full text parsing technique for text analysis. It distinguishes our system from alternative systems which only use shallow NLP techniques.

Algorithm Library: The algorithm library consists of a statistical algorithm that extracts key concepts from a document collection; a rule based algorithm that extracts relations between the key concepts; and a modified generalized association rule mining algorithm that builds the ontology.

Domain Lexicon: The domain lexicon contains terms specific to the domain of interest. These terms are used in the NLP component for analyzing documents. The domain lexicon is manually built and can be updated during the process of ontology learning.

The overall procedure for ontology learning is summarized as follows.

- *Data Preprocessing*: The CRCTOL system assumes that the input documents are in the plain text format. Text files in other formats are converted to plain text before processing.
- *NLP Analysis*: The input files are processed by the NLP component. Syntactic and POS tags are assigned to individual words in the documents.
- *Concept Extraction*: Concepts are identified by a statistical algorithm from text. These concepts are called the key concepts of the target domain.
- *Semantic Relation Extraction*: Semantic relations of the key concepts are extracted from the text. These include taxonomic and non-taxonomic relations.
- *Ontology Building*: An ontology is built in this step by linking concepts and relations extracted. The final ontology is presented in the form of a semantic network.

3. Concept Extraction

Traditional ontology learning systems for concept extraction were based on words. First, keywords were identified from the text. These words are typically single-word terms and will be seen as the concepts. Then, possible multi-word terms were formed by combining these keywords. As a result, the multi-word terms generated were not natural and most concepts extracted were only single-word terms.

When using the NLP component to process documents, we found most noun terms in the text were multi-word term. As it was also shown that 85% of the terms in text were multi-word terms [6], traditional systems focusing on single-word term extraction will thus miss many concepts. We adopt a different strategy for concept extraction. First, multi-word terms are induced from text directly. Then, single-word terms are extracted if they appear frequently in the multi-word terms or they are found related to the multi-word terms through certain semantic relations. This strategy reduces the chance of missing important concepts. The procedure for concept extraction is described below.

1. Extract all multi-word terms using the predefined regular expression rules. As concepts are nouns or noun phrases in texts, only word patterns with the NP tag are collected.
2. Remove articles and descriptive adjectives such as "a", "many" and "several" from the terms extracted.
3. Generate all possible sets of two or more words in each extracted term as candidate terms. For instance, generate *terrorist attack* from *international terrorist attack*.
4. For each term t, compute a linear combination

$$\textit{TIM-DRM}(t) = \alpha \textit{TIM}(t) + (1-\alpha)\textit{DRM}(t), \quad (1)$$

where $\alpha \in [0,1]$ is a weighting parameter. The *TIM* and *DRM* scores[1], are the statistical measures for evaluating terms. Terms with high *TIM-DRM* values are selected to form an initial concept list T.
5. Let V be the set of single-word terms appearing in the T as the syntactic head of a term t. For instance, in *(NP (JJ terrorist) (NNS attacks))*, *attack* is the syntactic head. We compute for each single-word term in V the occurrence frequency in T. Those with frequency above a threshold δ are added to the list T.

1 http://erlab.ntu.edu.sg/TIM_DRM.htm

4. Semantic Relation Extraction

We extract semantic relations between multi-word terms as well as relations between multi-word terms and single-word terms from the text collection.

Verbs are hypothesized to indicate semantic relations between concepts. A semantic relation of the ($Concept$, $Relation$, $Concept$) tuple thus has a lexical realization in text in the form of ($Noun_1$, $Verb$, $Noun_2$), where $Noun_1$ and $Noun_2$ are noun terms in text and concepts in the ontology, $Verb$ is the verb term in text, $Noun_1$ is the subject of $Verb$, and $Noun_2$ is the object of $Verb$.

As texts have been fully parsed by the NLP component, syntactic and POS tags are assigned to sentences. We thus adopt a rule based method similar to the CFG for extracting ($Noun$, $Verb$, $Noun$) tuples from texts. These tuples represent the semantic relations between the concepts extracted. The Noun and Verb terms are identified by the regular expressions below:

$$Noun : (DT)?(JJ)^*(NN|NNS|NNP|NNPS)^+$$
$$Verb : (VB|VBD|VBN|VBZ)^+$$

where JJ represents an adjective, NN, NNS, NNP, and NNPS represent nouns, DT represents an article, and VB, VBD, VBN and VBZ represent verbs.

Compared with alternative methods such as [3], more sentence level information is involved in our method for semantic relation extraction. The non-taxonomic relations extracted are thus more likely to be accurate.

5. Experiments

We conducted experiments to compare the performance of CRCTOL with Text-To-Onto (Version 1.0, released 09/11/2004) , based on a case study on the terrorism domain. Several considerations were involved in selecting Text-To-Onto as the yardstick of comparison. Firstly, it was publicly available, allowing a fair comparison. Secondly, it was one of the few systems that were able to extract semantic relations.

Documents of the US state department report "Patterns of Global Terrorism (1991-2002)" were downloaded from its website as the test corpus. Contrasting corpora were collected from the TREC collection, covering the commercial, computer, energy, and general domains.

5.1. Concept Extraction

Two experiments were set testing the CRCTOL's ability for concept extraction. One evaluated the efficiency of the full text parsing technique for extracting multi-word terms from text and the other compared the performance of CRCTOL with Text-To-Onto for concept extraction.

Filter	Recall	Prec.	F-Measure
Text-To-Onto	96.2%	99.1%	97.6%
CRCTOL	99.3%	99.7%	99.5%

Table 1. The performance of Text-To-Onto and CRCTOL for multi-word term extraction.

5.1.1. Multi-word Term Extraction Experiments were conducted to evaluate the CRCTOL's term extraction performance against Text-To-Onto, which used a linguistic filter defined by the regular expression: "$ADV^*ADJ^*NOUN^+$", where ADV is an adverb, ADJ is an adjective and NOUN is a noun.

Documents of the PGT corpus (1991) were used as the test corpus. Manual annotation of the document set identified 600 multi-word terms, used as the target list for evaluation. The linguistic filter of Text-To-Onto extracted 577 terms from the text, five of which were wrong. CRCTOL extracted 596 terms from texts, with two errors. The performance, in terms of precision, recall and F-measure, is summarized in Table 1.

We see that CRCTOL generally performs better than Text-To-Onto in multi-word term extraction. The lower precision score of Text-To-Onto is mostly due to its deficiency in separating modifiers from terms. For instance, it identified word strings such as "*difficult law enforcement effort*" as a term, although "*difficult*" is in fact a modifier of the term "*law enforcement effort*". The CRCTOL system works better for this problem.

5.1.2. Domain Concept Extraction Documents of PGT (1991-1992) were selected as the test corpus. Manual annotation of the documents identified 82 single-word terms and 104 multi-word terms as domain-specific concepts. Terms were selected in Text-To-Onto using the tf/idf measure, whereas TIM-DRM was used in CRCTOL. We compared their performance by evaluating the top 127 terms extracted by CRCTOL and Text-To-Onto respectively.

As shown in Table 3, CRCTOL produced much better precision and recall than Text-To-Onto in identifying domain concepts. The poor performance of Text-To-Onto can be attributed to several factors. First, the tf/idf measure is simply not suitable for domain concept extraction. In fact, our experiments found that selecting terms with low tf/idf scores performed even better than selecting terms with high tf/idf. Also, Text-To-Onto extracted single-word terms and multi-word terms together, but tended to miss a large number of multi-word terms. Our approach focused on multi-word extraction. Although our system also missed some single-word terms, the result on the whole is much more satisfactory. The top ten terms extracted by CRCTOL and Text-To-Onto are illustrated in Table 2.

CRCTOL	Text-To-Onto	
	highest tf/idf	lowest tf/idf
international terrorism	bid	government
terrorist attack	embargo	terrorist
terrorist group	war	terrorism
terrorist incident	profile	year
state sponsor	goal	scientist
intelligence service	ally	bomb
terrorist organization	pair	level
property damage	percent	organization
peace war	plane	international terrorist
civil war	expulsion	number

Table 2. The top ten terms extracted by Text-To-Onto and CRCTOL respectively.

	Recall	Prec.	Concept Missing	
			single-word	multi-word
CRCTOL	63.4%	92.9%	67	1
Text-To-Onto (highest tf/idf)	5.9%	8.7%	79	96
Text-To-Onto (lowest tf/idf)	17.4%	26.0%	52	101

Table 3. The performance of Text-To-Onto and CRCTOL for concept extraction.

5.2. Semantic Relation Extraction

Text-To-Onto was a typical system using shallow NLP techniques for semantic relation extraction. It defined the non-taxonomic relation as $VCC(n)$ [3], which described that concept $C1$ and $C2$ had non-taxonomic relation V if $C1$ and $C2$ both occurred within n words from an occurrence of verb V. That is, the Text-To-Onto system also extracted ($Noun$, $Verb$, $Noun$) tuples from texts as the semantic relations. A lot of hand-tailored rules were used in the Text-To-Onto system for the non-taxonomic relation extraction.

A key challenge in relation extraction is the handling of sentences with complex structure. For instance, a verb can be modified with auxiliary verbs, such as "have to do", "should do" and "may do" etc. To study the impact of auxiliary verbs, we conducted two experiments, one based on sentences without auxiliary verbs and the other based on all sentences (i.e., with no restriction).

For the first experiment, the documents of the PGT corpus (1991-1997) were used. There were 111 sentences without auxiliary verbs containing 141 semantic relations. Text-To-Onto extracted 122 relations, of which 15 were incorrect. CRCTOL extracted 118 relations with only one incorrect. The performance of the two systems, in terms of precision, recall, and F-measure, is summarized in Table 4.

System	Prec.	Recall	F-Measure
Text-To-Onto	87.7%	75.0%	80.9%
CRCTOL	99.1%	82.9%	90.3%

Table 4. The performance for relation extraction on sentences without auxiliary verbs.

System	Prec.	Recall	F-Measure
Text-To-Onto	74.4%	25.7%	38.2%
CRCTOL	86.0%	57.1%	68.6%

Table 5. The performance for relation extraction on sentences without restriction.

For the second, documents of the PGT corpus (1991) were used. There were a total of 243 sentences in the documents containing 226 semantic relations. Text-To-Onto extracted only 78 relations, of which 20 were incorrect. CRCTOL extracted 150 relations and 21 was wrong, translating into a precision score of 86% (Table 5). The recall however was much poorer, due to the difficulty in handling complex sentence structure. Nevertheless, it was still much better than the 25.7% recall obtained by Text-To-Onto.

We see CRCTOL works much better than Text-To-Onto in semantic relation extraction, especially in handing sentences that may include auxiliary verbs. This indicates that a full text analyzing technique is more effective to handle sentences with complex structure for the purpose of mining text content.

References

[1] T. Berners-Lee, J. Hendler, and O. Lassila. The semantic web: A new form of web content that is meaningful to computers will unleash a revolution of new possibilities. *Scientific America*, 2001.

[2] T. R. Gruber. A translation approach to portable ontology specification. *Knowledge Acquisition*, 5:199 – 220, 1993.

[3] M. Kavalec, A. Maedche, and V. Svateck. Discovery of lexical entries for non-taxonomic relations in ontology learning. In *SOFSEM2004*, pages 249–256, 2004.

[4] A. Maedche and S. Staab. The text-to-onto ontology learning environment. In *Software Demonstration at ICCS-2000*, pages 14–18, August 2000.

[5] M. Missikoff, R. Navigli, and P. Velardi. The usable ontology: An environment for building and assessing a domain ontology. In *ISWC 2002*, pages 39–53.

[6] H. Nakagawa and T. Mori. Simple but powerful automatic term extraction method. In *COMPTERM 02*, pages 29–35, 2002.

[7] K. Rajaraman and A.-H. Tan. Mining semantic networks for knowledge discovery. In *Third IEEE ICDM*, pages 633–636, 2003.