A Topic-Oriented Syntactic Component Extraction Model in Social Media

Yanxiang Xu[1], Tiejian Luo[1], Guandong Xu[2], Rong Pan[3]

[1]School of Information and Engineering, Graduate University of Chinese Academy of Sciences, Beijing 100049, China
[2]Centre for Applied Informatics, Victoria University, Australia, PO Box 14428, Vic 8001, Australia
[3]Department of Computer Science, Aalborg University, DK-9220 Aalborg, Denmark

{Xuyanxiang,tjluo}@gucas.ac.cn
guandong.xu@vu.edu.au
rpan@aau.dk

Abstract. Topic-oriented understanding is to extract information from various language instances, which reflects the characteristics or trends of semantic information related to the topic via statistical analysis. The syntax analysis and modeling is the basis of such work. Traditional syntactic formalization approaches widely used in natural language understanding could not be simply applied to the text modeling in the context of topic-oriented understanding. In this paper, we review the information extraction mode, and summarize its inherent relationship with the "Subject- Predicate" syntactic structure in Aryan language. And we propose a syntactic element extraction model based on the "topic-description" structure, which contains six kinds of core elements, satisfying the desired requirement for topic-oriented understanding. This paper also describes the model composition, the theoretical framework of understanding process, the extraction method of syntactic components, and the prototype system of generating syntax diagrams. The proposed model is evaluated on the Reuters 21578 and SocialCom2009 data sets, and the results show that the recall and precision of syntactic component extraction are up to 93.9% and 88%, respectively, which further justifies the feasibility of generating syntactic component through the word dependencies.

Keywords: Text Understanding, Topic-oriented Parsing, Syntactic Component Extraction, Text Modeling, Natural Language Understanding.

1 Introduction

With the rapid development of internet, a large volume of contents resulting from various network applications have been created, while the existing information processing models could not fully handle the information extraction in natural language processing. In business domain, traditional business surveys costing a lot of

manpower and resources is used to retrieve market trends and customer opinions; however, it is often hard to get a high sampling coverage. Therefore, companies aim to directly understand and obtain the customer's intention and find the product demand and potential market from user feedbacks to improve the products and. These business surveys with the intention of forecasting tasks can be attributed to the topic-oriented understanding in text. The core concept within this context is the topic of concern to find the distribution pattern of information from the collection of the specified text to achieve at some kind of trend forecasting. The central idea of the "topic" generally refers to not only a conversation, a lecture or discourse, but also the talking about the theme, or a concern of a specific object. The "topic" in this paper is especially concerned about a specific event, resource or action. A topic can be reflected by one or a certain amount of the textual information.

The data source in topic-oriented understanding is the massive user-generated content on the target site, and it is usually a short form of natural language text. A piece of text is an ideological expression recorded in the form of human natural language and its processing requires the application of natural language understanding technology.

The topic-oriented natural language understanding deals with the "analysis and forecasting" which aims at using computers instead of human beings to process the large amount of text, define the extraction models from the viewing point of semantic formalization, and thus the induce the thematic distribution information from the extracted information, interpret the current observation and predict the future trend via statistical approaches.

The core of topic-oriented information extraction model is the derivation of the pattern of "Subject – Description". The "Subject" refers to the objects, events and activities. "Description" covers the advices, comments, evaluation, intents and demands made on a specific topic. The "Subject" and "Description" extracted from the text have a clear meaning for the syntactic elements. Syntactic elements play a structural role in the organization of words, phrases or sentences, typically including the "subject" (the subject of a statement by the sentence) and the "predicate" (used to describe which action the subject does the action or at which state it is on) and so on.

In understanding systems we need to conduct syntactic analysis, and then choose the desired syntactic component from it. Syntactic analysis is to model the internal structure of the sentence by using word as the basic elements, to reveal the relationship between words (such as: dependency, the Lord from the phrase), the attributed role a word or phrase plays in a sentence, as well as the structural characteristics (such as: complex and compound sentence) and other information.

In summary the whole framework of a topic-oriented understanding system is shown in Figure 1.

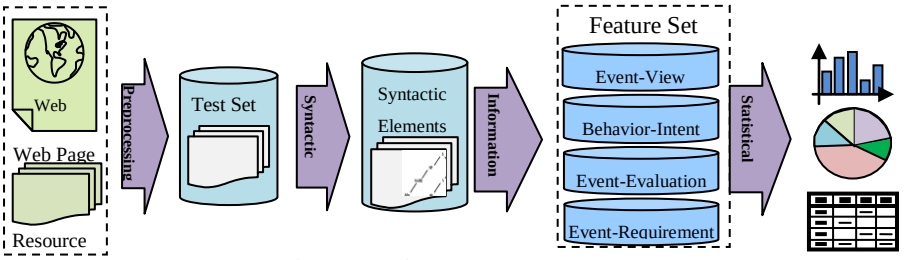

Fig. 1. The framework of Topic-oriented Understanding System

The whole framework consists of four steps: First, crawl and preprocess the text from the web and resource, and form the text collection based on the single content unit; secondly, analyze the text collection, annotate syntactic elements contained in the text; thirdly, based on the requirements of the different understanding tasks, extract specific syntactic elements, standardize them and form the feature sets from the information; at last, conduct statistical analysis on the feature sets of information, and present the results in the statistical form of charts and tables.

In this framework, the text pre-processing part is realized by the use of conventional crawling and text processing tools commonly used in text processing systems, the statistical analysis is well studied in database and statistics. However, there are still open questions to extract feature information for syntactic component generation. It is not only because of the capability of effective natural language processing and understanding; but also the requirement that the derived syntactic components could be used in the topic-oriented understanding of extracted information.

In this paper, our main contributions are to:

- propose a reduced syntactic component extraction model to guide the syntactic analysis of text to generate syntactic elements, which can be used for information extraction;
- devise an algorithm of mapping the extracted components with the targeted syntactic component; and
- implement the syntactic component generation algorithm based on existing POS tagging and parsing techniques, and validate the effectiveness of the proposed model and algorithm by experiments.

The remainder of this paper is organized as follows. In Section 2 we discuss the related work. Section 3 introduces the reduced syntactic component extraction model, and then Section 4 proposes the Framework of the Model and Algorithm. Section 5 reports the evaluations of the experiment. Section 6 concludes the paper.

2 Related Works

The research for the formal grammar theory begins with Chomsky [1]. In the Chomsky's “Syntactic Structures” discusses the syntax of the language in the form of rules of composition and structure of the rules, priority rules. The first rule is rewritten as S (sentence) NP (noun phrase) and VP (verb phrase), the following rules further

to rewrite the NP and VP, until the formation of the final lexical items and grammatical elements combination. Generate results can be used with the syntax name (NP, etc.) to demonstrate the type of graph is called a labeled tree labeled trees.

Lucien Tesnière is the the founder of the modern dependency grammar and valence theory. He consists the fundamental elements of 'Structural syntax'[2], which is known as dependency grammar, and the main theoretical schemata, with price and interdependence. He stressed that "understanding a sentence, is to identify the linkages between the various words in the link sentence." Structural links are established dependencies between words.

For the Syntactic component analysis, the "subject – description mode" is the one kind of information organization, and there're different expressions for the various languages. If we use the "Subject – Predicate" structure to analysis the Chinese, the syntactic rules of language instance perhaps only 50% [3].

3 Syntactic Component Reduced Extraction Model

3.1 Framework of the Model

Syntactic component extraction model is the process to help transform from the natural language to extractable information features. Thus, the model should not only be able to satisfy the syntactic generality of the structure coverage of the target text, but also meet the requirements of topic-oriented information extraction.

The model is divided into two parts: the syntactic composition model and the reduced extraction model. The syntactic composition model is responsible for parsing the text, and the goal is to, without the change of semantics, to transform the natural language expressions in different syntax and diverse expressions into the symbolized components of sequence or tree forms, as the basis of information extraction. Semantic represents the inherent content meaning to express. The reduced extraction model will extract the required components according to the different requirements of topics from the syntactic component. The framework of the model is shown in Figure 2:

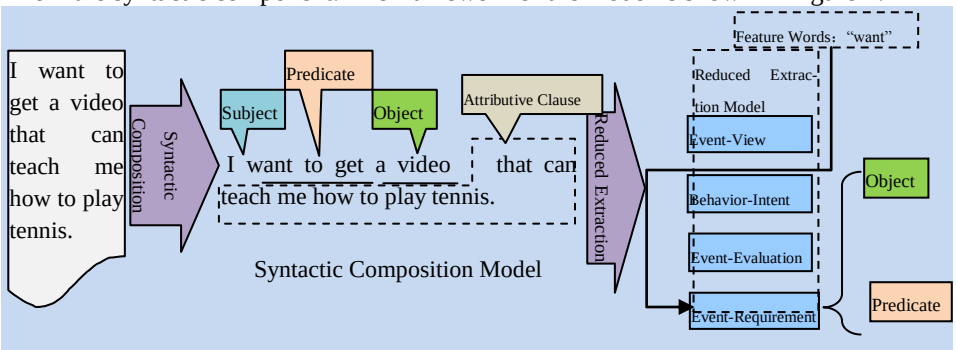

Fig. 2. The Framework of the Model

3.2 Syntactic Component Model

Syntactic component model is to carry out the dual tasks of syntactic coverage and information extraction. It includes three aspects: The first is able to handle the trunk syntax of component extraction and semantic capture; the second is the function of the complete description and the primary-secondary component separation of syntactic elements and; and the third is the summarization of complex sentences.

3.3 Reduced Extraction Model

The syntactic component extraction model should get the core backbone sentence, combined with modified ingredients, covered the complex sentence, and compound a variety of sentence patterns. The reduced extraction model for syntactic components consists of two parts: the first describes the mapping between the trunk sentence structure and the "subject – description mode"; and another part is the description of the guidance of how to extract required components from the analyzed syntactic components. The so-called term "reduced" reflects the process that when dealing with the mapping only the trunk sentence structure is mainly concerned.

4 Algorithm and Implementation

In this section we describe the details of Syntactic Component Reduced Extraction Model (SCREM) algorithm and its implementation.

4.1 Syntactic Preprocessing

We integrated the NLP program package [4] toolbox provided by the NLP research Group at Stanford University into our main system to complete sentence separation, word segmentation, POS tagging and dependency relationship recognition. This results in the collection of textual instances expressing the dependency relationships between the words.

The dependency is the triples describing syntactic relationships between a pair of words. For example, "nsubj (play, boy)" denotes the subject "(nsubj)" of "play" is "boy". The Stanford's NLP package defines 53 typs of dependency [5], covering the majority of syntactic elements.

4.2 Generating Syntactic Graph

The sentence trunk model is a tree structure with root of subject. Therefore, we can generate the syntactic component based on the dependency between randomly ordered words. The strategy of generating syntactic graph is extending from the root to its branch layer. Via iterative traversals over the derived word dependency relationships we can extract the demanded syntactic components within the sentences in a sequence such as "Subject - Predicate", "Subject - Verb – Predicative", "Predicate -

Object", "Predicate - Complement", "Clause" and other syntactic component relationships. Three main syntactic component generation algorithms developed for sentence trunk model are generation algorithms of "Subject - Predicate", recovering algorithm of "Subject - Verb – Predicative" and generation algorithm of "Predicate - Object". Due to the limitation of length, we won't give the detailed description of these algorithms.

4.3 Merging Word Sequence

Each node in the SCSEM graph represents a syntactic component, a node can contain a few words, there're three kinds of it: 1) for the predicate node, the auxiliary words in passive voice sentences, the negative words in negative sentences, modal verbs, modal particles and other components incorporated should be combined into the predicate node; 2) for the noun phrase, the article should be added before the noun node; 3) for a fixed phrase, it needs to be merged into the same node.

The edges of the SCSEM graph are constructed by the relationship between syntactic components which contain more than one word in the syntactic component. While traversing the dependency relationships, it needs to merge the words belonging to the same identified node. The merging of words is realized by referring to the attribute belongingness of the "pre-word" and "post-word". We implemented the merging of words by a data structure of bi-directional linked list.

4.4 Graph of Output Syntax

We implemented a prototype system by integrating the above algorithms with the Stanford NLP-core package, called "Syntactic Component Builder V1.0". The system is developed on the eclipse 3.4.2 integrated development environment, and the user interface is manipulated by the use of the Java Swing-based framework and the JUNG package is utilized to generate the SCSEM graph output. In later experiments, we carried out evaluations with this prototype system.

5 Evaluations

We have utilized two measures which are commonly used in information retrieval, namely Precision and Recall to evaluate the system. They are defined as Precision = the number of correctly identified syntactic components / the total number of identified syntactic components; and Recall = the number of correctly identified syntactic components / total syntactic components contained within the original text.

5.1 Experiment Datasets

We selected 3 articles from SocialCom2009 Proceedings [9] and Reuters-21578 [10] news dataset respectively to form the experimental datasets. For the articles from proceedings dataset, we remove the contents of the title, charts and equations to pre-

pare paragraphs of test sets containing around 1000 words. As for articles from news datasets, we remove the special punctuation, symbols, and other information to truncate paragraphs of about 200 words. The statistics of test sets is shown in Table 1. The predefined 1330 syntactic components is able to preliminary meet the experimental requirements for statistical significance and coverage. And the experiments are carried out upon the syntactic component collection rather than the articles themselves.

Table 1. Dateset Descriptions

No	Source	Reference Description	Number of Sentence/ Words	Number of Syntactic Elements
1	SocialCom2009	SC-1[6]， Introduction	5/1296	409
2	SocialCom2009	SCA-31[7]， Introduction and the first 5 paragraphs in Chapter 2	42/1204	330
3	SocialCom2009	SIN-8[8]， Abstract, Introduction	37/1106	367
4	Reuters-21578	ID:7019，Full Paper	9/229	61
5	Reuters-21578	ID:12377，Full Paper	9/241	86
6	Reuters-21578	ID:15125，Full Paper	11/236	77
	Total		**160/4312**	**1330**

5.2 Result of Experiment

We use the Syntactic Component Builder described in section 4.6 to conduct text parsing and statistical analysis. The statistical results of all text syntactic extraction precision and recall rate are shown in Figure 3. We can see that the average extraction precision can reach up to 88% and the recall rate 93.9%.

We get the precision of 80.3% by utilizing the Stanford open source packages to derive the dependency relationship with a small sample of 10 sentence test. We believe that the improvement of precision and recall is mainly due to the focus on the syntactic components and the increased granularity of syntactic component. After further analysis, we find that the error of syntactic precision and recall rate is arising primarily from the annotation mistakes of dependence identified.

6 Conclusion

In this paper, we summarize the main difference between the topic-oriented text understanding and the traditional reasoning-based natural language understanding. We propose the framework containing the syntactic analysis, information extraction and the characteristic analysis process. The whole framework is based on the “subject -

Description" information extraction pattern and the main technical contribution is the syntactic component reduced extraction model.

We analyze, design and implement the topic-oriented syntactic component extraction model. The use of syntactic phrases as syntactic elements can prove the model is able to overcome the contradiction between the simple consistency and syntactic diversity of information extraction.

It is found that the syntactic component is corresponding to the semantic segment in language organization, which could be phrases but not to be limited to words. The use of syntactic analysis techniques, in particular, the syntactic dependencies between words, can effectively generate the syntactic elements for information extraction. We also conclude that by increasing the granularity of syntactic elements, the words become phrases, which can certainly improve the precision of the syntactic component extraction.

Table 2. Analysis of Experimental Results

No	**Syntactic Elements**	**Identified Elements**	**Correct Elements**	**Precision**	**Recall**
1	409	485	392	80.8%	95.8%
2	330	347	302	87.0%	91.5%
3	367	366	347	94.8%	94.6%
4	61	69	60	87.0%	98.4%
5	86	89	78	87.6%	90.7%
6	77	78	71	91.0%	92.2%

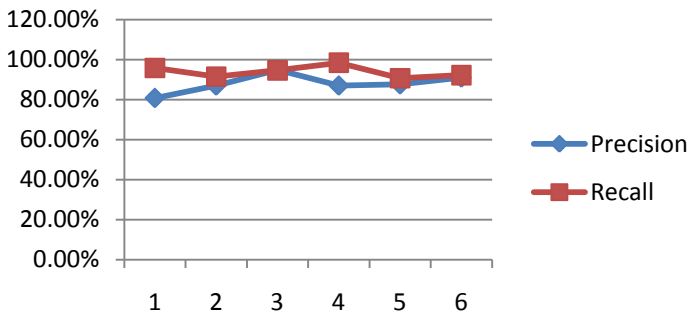

Fig. 3. Precision and Recall of Experiment Results

7 References

1 Chomsky, Noam (1957), Syntactic Structures, The Hague/Paris: Mouton
2 Tesnière, L. (1959) Eléments de la syntaxe structurale. Paris: Klincksieck.
3 Tongqiang Xu, Theory of language: the semantics of language structure and principles and research methods，Press of Northeast Normal University，1997.10, 9787560220505
4 http://nlp.stanford.edu/software/tagger.shtml
5 http://nlp.stanford.edu/software/lex-parser.shtml
6 Lijie Zhang and Weining Zhang, Edge Anonymity in Social Network Graphs, SocialCom 2009, Canada, August 29–31, 2009
7 Marc Smith, Derek L. Hansen, Eric Gleave, Analyzing Enterprise Social Media Networks, SocialCom 2009, Canada, August 29–31, 2009.
8 Philip Hendrix, Ya'akov Gal, Avi Pfeffer, Using Hierarchical Bayesian Models to Learn About Reputation, SocialCom 2009, Canada, August 29–31, 2009.
9 http://cse.stfx.ca/~socialcom09/
10 http://www.daviddlewis.com/resources/testcoll-ections/ reuters21578/