The Web Information Extraction for Update Summarization Based on Shallow Parsing

Min Peng[1], Xiaoxiao Ma[1], Ye Tian[1], Ming Yang[1], Hua Long[1], Quanchen Lin[1], Xiaojun Xia[2]

[1] Wuhan University
Wuhan, China
{pengm, xxma, tianye, yangming, hlong, quanchen9}@whu.edu.cn

[2] North Dakota State University
USA
Xiaojun.xia@ndsu.edu

***Abstract*—Traditional text information extraction methods mainly act on static documents and are difficult to reflect the dynamic evolvement of information update on the web. To address this challenge, this work proposes a new method based on shallow parsing with rules. The rules are generated according to the syntactic features of English texts, such as the tense of verbs, the usages of modal verbs and so on. The latest novel information in English news texts is extracted correctly, to meet the needs of users for accessing to updated information of the developing events quickly and effectively. Performance results show the improvement of the proposed scheme in this work.**

Keywords: shallow parsing; web texts; updated information; information extraction

I. INTRODUCTION

A huge amount of information is being added and updated continuously on the World Wide Web (WWW). Users want to get useful and latest information on it quickly, but information overload is always a problem. However, though Information Retrieval (IR) systems such as Google, Yahoo etc. addresses the overload problem by identifying documents relevant to the user's query, ranking them and presenting them as an ordered list, the number of search results is very high and information pertaining to a query might be distributed across several sources without consideration of information update. It's very necessary to filter and aggregate updated information relevant to the user's query, and present it as a digest or summary. The technology of text extract and automatic summarization will be useful for readers to learn the latest development and the forecast of happened events from experts and commentators. However, the research poses significant challenges such as maintaining integrity, coherence and efficiency.

In this paper, we propose an approach to extract updated information for text summarization. It expands the traditional time-based sentences abstraction strategy, considering the tense feature in each sentence for texts parsing. According to the parsing results, the updated information is grouped and related sentences are extracted.

The paper is organized as follows. Section II starts from the state-of-the-art update summarization and presents a survey of the latest work on update information extraction. Next, Section III details the proposed shallow paring solution. Experiments and performance analysis and comparison are then presented in Section IV. Finally, conclusions and future research directions are highlighted in Section V.

II. RELATED WORK

A. Update Summarization

Summarization approach can be classified as abstraction and extraction. The work here is focused on extraction-based update summarization. Update summarization is an important branch in the research of automatic summarization and even NLP (Natural Language Processing), which is to produce a short update summary from newswire articles to inform the reader new information about a particular topic. Furthermore, it expands and outspreads the traditional text summarization, considering the chronological order of the texts and aiming to create a summarization under the premise that users have already known or don't care the historical information [1].

Update summarization is the task to generate summaries of raw texts while minimize redundancy with previously read documents. Actually, users prefer to track fresh information from news or documents which has the similar topics as the preceding history information. Meanwhile, they also go for the information with higher novelty in the timeline, don't satisfy with the information which is pretty similar to the preceding content. No doubt, real-time information in time-sensitive documents or news of happening events is more important and than other parts for readers.

Hence there is a strong demand to get the most novelty information from the web news or other documents, the

techniques of update information extraction are very important in text summarization. Update summarization was included in DUC (Document Understanding Conference) in 2007 and TAC (Text Analyze Conference) in 2008, 2009 as theme tasks [1][2]. The aim here is to address the key technology in update summarization, e.g., how to extract time-sensitive sentences from documents based on sentences syntax parsing.

B. Web Update Information Extraction

To address the challenges in update text summarization, a host of extraction schemes have been proposed. The general methodology is to identify time-relative references within news documents to construct a timeline for all events in news. Under some ideal circumstance (such as technology literatures or news with specific time-stamps), it works well. However, in most cases, it is too complex to yield very good results, since time-relative references highly depend on their surrounding elements in texts and also require prior conditions, e.g., an understanding of ontological and logical foundation of temporal reference construction [3] which involves the recognition of the all different forms of time references. What is more, the standardization of these references is another bigger difficulty. Overall, the above research works based on temporal reference cause time consuming and unsuitability for computer processing.

Recent efforts on improved time-sensitive references strategy have been done. Typically, Wan [4] put forward TimeTextRank algorithm which placed more emphases on new information extraction dynamically by putting the temporal dimension into consideration. In particular, this work took time gap between publication time and current time in document as a novelty measure weight, namely, the document published latest would get highest attention. However, there is an obvious defect in this work, i.e., the newer information can't be distinguished from older ones within one document, which will be difficult to present novelty information accurately.

Due to the obstacles on the temporal-based research on web update information extraction, diverse research works are basically focused on digging out the disparate content compared to the history information. For example, Li and Wei [5] proposed a PNR^2 model to weigh the novelty by calculating the similarity between documents.

In addition, some works address novel information detection in the sentence level. Time expression consisted in sentences is adopted as time stamp to improve fresh information extraction quality, such as the work in Suo [6]. It was supposed that the sentences containing repeated time-stamps are more novelty and significant, therefore, these sentences were selected as main candidates to be picked out as updated information. Still, the time-tamps were used to build the timeline of events, and the real-time information would be ranked in the front of the timeline (with higher sentence score). But there were still many noise sentences without real-time information in the extracted information.

Meanwhile, time-stamps were applied in Elena [6] as well. The difference was the time-stamp here was either time-points or time-intervals, and was assigned to each clause in the text.

As mentioned earlier, for a time-sensitive text, the recognition and the standardization of time expression are the bottleneck to construct a timeline to extract update information. Our work is inspired by Elena's work [6] in a certain extent. Their work proposed a tense-based implicit time references method, while we expand tense to syntax analysis to identify time-oriented novelty information. Meanwhile, we remove the relatively complicated part of identifying explicit time references which mainly refers to temporal adverbials. Note that timeline constructing is not considered in our work for the following concerns: (1) It is too complicated to establish timeline; (2) The sentences with the latest time references do not always contain the most updated information; (3) Generally, instead of the exact time, readers prefer to know what is happening recently and what will happen in the near future; (4) From past experiences, it is noted that quite a part of necessary statements in time-sensitive text concern updated information but don't contain an explicit time-point or even an implicit time-interval.

III. METHODOLOGY

Through the observation and analysis, we believe that novelty of an event is much more related to novelty level of executive action than time-stamps. To a specific text, the novelty of an executive action can be weighed by the novelty degree of the verbs.

In English texts, there are lots of special features (such as verb tenses, phrases and grammars) which can be applied to tell the novelty degrees. By means of the natural advantage in English text, we propose a syntax parsing based method to identify and extract update information, and avoid the tedious work of time references analysis greatly.

A. Shallow Parsing

Paring is one of the most important technologies in NLP, and is divided into shallow parsing and fully parsing according to the parsing level. Shallow parsing is also named as partial parsing or chunk parsing. Fully parsing requires a complete parsed tree through a series of parsing process for each sentence, while shallow parsing is to identify some specific and simple ingredients within sentences (chunk), such as noun phrases, verb phrases, etc [8].

To date, in terms of parsing technology, there are two primary approaches, e.g., statistics-based and rule-based. We adopted the latter one, which means some manual writing rules have to be provided in advance, to define phrases' borders and types.

According to the labeled strategy, rule-based parsing is classified into two categories:

1) Incremental / constructive approach: Some syntactic markers such as the boundary and type of a phrase are inserted into the sentence string in this approach.

2) Reductionist approach: On the contrary, illegal markers are deleted from multiple candidate syntactic makers.

In this paper, the incremental approach is mainly employed to parsing step based on a rules set which was made ahead and represented by a series of regular expressions. The rules set here is used for training a text parser, and the trained parser will be applied to parsing to get novelty degree for each sentence. Then sentences will be classified according to their novelty degree.

B. Updated Information Extraction

1) Working Flow

Note here that, since the main work in this paper is sentence parsing instead of yielding an update summarization, in texts level, we apply TimeTextRank [4] to find the latest text among a series of time-sensitive reports to simplify the process and handle the single text object in our work.

The entire working flow is shown in Figure 1, main steps are as following:

a) Sentence segmentation

Generally, most of sentences in English texts are long and carry lots of information. In order to accelerate the processing, we split the sentences into the minimum units with comma. Some other preprocessing works are also necessary. For example, the too short sentences should be filtered out, since they just play a decorative and supplementary role.

b) Word segmentation and POS tagging

First, all of the filtered sub-sentences will be segmented into tokens then. POS tagging here is necessary too, as we will pay the most attention on verbs in this paper. In our opinion, only the sub-sentence which consists of at least one verb has the possibility to carry the information of the happening events. We define it as "*event sub-sentence*". The tagging step is very vital and it must be done in advance. Next, each "*event sub-sentence*" form a data structure in the style: $S[(w_1,t_1),(w_2,t_2),(w_3,t_3).....(w_i,t_i)]$, w_i stands for the words in sentence and t_i for the POS tag of w_i.

c) Rules making and parser training

All the rules designed for parsing are mainly handmade, represented by a set of regular expressions. All the tagged sub-sentences will be sent to the trained parser to produce a collection of parse trees. Every tree is corresponding to a sub-sentence.

d) T-chunk detection and information grouping

Finally, traverse each parse tree and detect the chunks suggesting chronological information in the sub-sentence.

We simply name these chunks "*T-chunk*". They can be employed to divide the whole sub-sentences collection into different groups, namely, future information, present information and past information.

Further details about paring and grouping will be discussed in the next part.

2) Parsing and Information Grouping

By exploring a certain amount of English texts we find that some existing syntax features, such as the tense of verbs, can indicate the chronological information of sentences they are located. Particularly, via the study of the modal verbs' usage, we conclude that the most important function of them is to express the prediction of the future situation. Therefore, we stipulate that the verbs attached to modal verbs also keep a higher temporal novelty. Inspired by this, expanded parsing rules are formulated below.

As the ultimate goal is to extract the updated information in texts, the starting point of getting rules should be to pick out these sentences which will be grouped into the future information set.

Main part of the parsing rules is listed as following:

a) Tense of notation rules

With the help of the words' POS tag, the tense of notation verbs can be identified and then used to judge whether the action happened in the past or is happening at present;

b) Syntactic rules

Three kinds of *T-chunk* are designed for groups matching in the parsing process, to judge if the action will happen in the future.

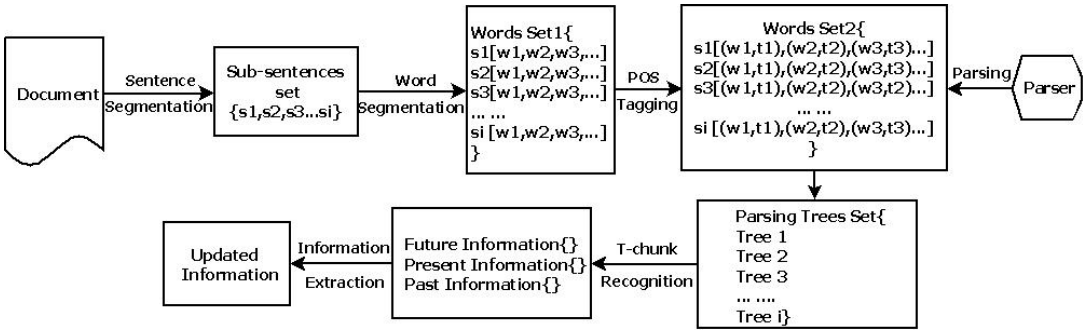

Figure 1. Working Flow

- Chunks shaped like [modal verb...+ base verbs], such as "will buy" or "shall not go" and so on. This sentence structure is mainly used to express prediction and speculation of events take place in the future;
- Chunks in [be going to...+base verbs] style, sentences in this style convey the message of "prepare to do..." or "mean to do ...";
- Chunks in [be to...+base verbs] format which delivering the information of what is most possibly going on.

Next step, all the *event sub-sentences* go through the trained parser containing all the rules, and then a set of parse trees with equivalent amount of sub-sentences will be produced. Now clearly, many *T-chunks* classified into three groups also will be re-generated during the growth of each tree, i.e., each time, once one of the rules is satisfied, there will be a *T-chunk* acting as a sub-tree. During the traversing, the sentences can be grouped into three sets according the type of *T-chunk*.

The future information set is the focus, since our original aim is to extract the updated information. In this paper, the judgment of future information gets a priority, which will be done before all the sentences are identified for the second time, to guarantee that if a sentence possesses the feature of future information, it will be definitively settled into future group preferentially and rightly. Past information set is not emphasized either on parsing or judging.

IV. EXPERIMENT AND EVALUATION

To evaluate the method proposed in this paper, we choose a news document to go through the parser. The news title is "China will send delegations to US, EU to correct trade imbalance", published in "China Daily" on the web in March 7, 2011. Two of the parse trees will be selected for testing. The performance of update extraction based on shallow parsing is tested by Python Natural Language Processing Toolkit [9].

We can learn the specific meaning of defined *T-chunk* by observing the parsing trees. In the collection of *event sub-sentences*, the 4th one is "that China is going to send many more trade delegations to developed nations in 2011 than in previous years." After parsing, the result is shown in Figure 2.

Same, the parsing result of 5th "The delegations will buy goods" is shown in Figure 3.

We chiefly present the future information group here which is closely related to updated information. The final grouped information of this news is shown in Table I.

Table I shows a work making sense to get the "*event sub-sentence*" set (seen in the second blank), the mark numbers is the order of original text. From the second blank we can see that most of them are containing event. Furthermore, a significant grouped information is produced (seen in the third blank) with the same number marked earlier. Here, as expected, the result shows that most of the extracted sub-sentences for future information can represent the updated and novel information mostly, which states that our work can yield a good result.

Next, a comparison is supplemented here to evaluate our update information extraction method. The comparative experiment is designed to act on a larger corpus: 10 news text documents published in the "China Daily" on web in March and April 2011. The hits rate of the shallow parsing method in this paper is compared to the Timeline constructing method. The hits rate here is the rate that the amount of sub-sentences got from each method to the total number obtained manually. According to the respective comparison results of the hits rate and the average performance shown in Table II, we can conclude that the performance of our method by shallow parsing for novel information extraction is much better.

Though the performance is better, it is also observed that there is still a little missed information in the process of parsing. The reasons are as below: Firstly, the precision of parsing is restricted by the precision of tagging; Secondly, the handmade rules are not comprehensive enough to cover whole *T-chunks*, which causes some sub-sentences can not be grouped or can not be grouped rightly.

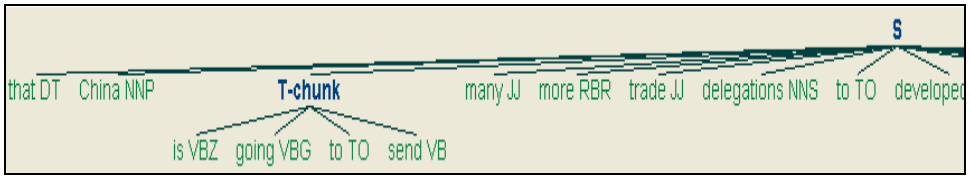

Figure 2. Parsing result of 4th sub-sentence

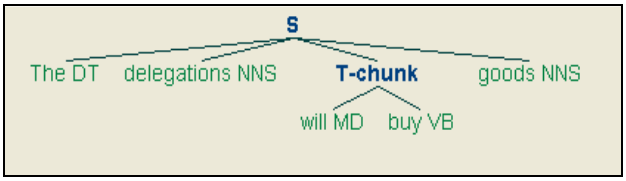

Figure 3. Parsing result of 5th sub-sentence

TABLE I. GROUPED INFORMATION

NEWS RAW DOCUMENT	BEIJING - China will send two large trade delegations to the United States and two to European Union nations this year to help stimulate its imports, said the head of China's trade promotion organization. Wan Jifei, chairman of the China Council for the Promotion of International Trade, told China Daily at the sidelines of the annual session of the Chinese People's Political Consultative Conference (CPPCC) National Committee, that China will send many more trade delegations to developed nations in 2011 than in previous years. The delegations will buy goods, especially energy-saving and environmentally friendly products, with the aim of balancing trade between China and those regions." During the first half (of the year), we will have one team each for the US and the EU - including Britain, Germany and France - and in the second half, one more each for both regions," said Wan, also a member of the CPPCC National Committee. And a few days after the two sessions of CPPCC and National People's Congress, the China-Japan Green Economy Forum will be held in China. During the forum, China will buy large amounts of high-tech, energy-saving and environmentally friendly products, the first time it will have ordered such a large volume of those types of goods from Japan, Wan said. During the forum, China will buy large amounts of high-tech, energy-saving and environmentally friendly products, the first time it will have ordered such a large volume of those types of goods from Japan, Wan said.
EVENT SUB-|SENTENCES	[①'BEIJING - China will send two large trade delegations to the United States and two to European Union nations this year to help stimulate its imports', ②" said the head of China's trade promotion organization.", ③" told China Daily at the sidelines of the annual session of the Chinese People's Political Consultative Conference (CPPCC) National Committee",④ ' that China is going to send many more trade delegations to developed nations in 2011 than in previous years.',⑤ 'The delegations will buy goods', ⑥' with the aim of balancing trade between China and those regions.', ⑦' we will have one team each for the US and the EU - including Britain',⑧ ' the China-Japan Green Economy Forum will be held in China.', ⑨' China will buy large amounts of high-tech', ⑩' the first time it will have ordered such a large volume of those types of goods from Japan'… …]
GROUPED INFORMATION	Future Information[①'BEIJING - China will send two large trade delegations to the United States and two to European Union nations this year to help stimulate its imports',④ 'that China is going to send many more trade delegations to developed nations in 2011 than in previous years .', ⑤'The delegations will buy goods', ⑦'we will have one team each for the US and the EU - including Britain',⑧ 'the China-Japan Green Economy Forum will be held in China .',⑨ 'China will buy large amounts of high-tech',⑩ 'the first time it will have ordered such a large volume of those types of goods from Japan' …. …]

TABLE II. COMPARISON RESULT

News Documents	Shallow Parsing	Timeline Constructing
1	12/15	7/15
2	5/6	2/6
3	6/10	3/10
4	7/9	2/9
5	7/10	6/10
6	6/9	3/9
7	4/6	4/6
8	5/6	1/6
9	3/4	2/4
10	3/5	3/5
Average Hit Rate	72.3%	41.7%

V. CONCLUSIONS

This paper studies the extraction of update information from web time-sensitive texts. We propose an approach based on shallow parsing with rules, to meet the needs of users for quick and effective access to updated information of happening events contained in the texts or news. We specially take use of tense of notation verbs, modal verbs and some other syntactic features of English texts to generate parsing rules. Detailed performance results show the improvement of the proposed scheme.

Future research will focus on two sides. First is to improve the tagging accuracy by adopting other advanced POS tagging policy. Another is studying a self-correct procedure based on machine learning and conversion rules, to promote the performance of the parser.

VI. ACKNOWLEDGEMENT

This research has been supported by the *National Science Foundation of China* (NSFC) under Award 61070083. The authors are very grateful for this generous support.

REFERENCES

[1] TAC2008 Update Task, http://www.nist.gov/tac/2008/summarization/

[2] DUC2007 Update Task, http://www-nlpir.nist.gov/projects/duc/duc2007/tasks.html

[3] Florian Boudin, Juan-Manuel Torres-Moreno, Marc El-El-B'eze DIRO, "Improving Update Summarization by Revisiting the MMR Criterion," Proc. of TAC 2010, April 21, 2010

[4] Xiaojun Wan, "TimedTextRank: Adding the Temporal Dimension to Multi-document Summarization," Proc. of 30th ACM SIGIR, 2007, pp 867-868

[5] Wenjie Li, Furu Wei,Qin Lu,Yanxiang He,"PNR2: Ranking Sentences with Positive and Negative Reinforcement for Query-Oriented Update Summarization," Proc. of COLING, 489-496

[6] Hongguan Suo, Yuhuan Liang, Yushu Liu ,"Automatic Multidocument Summarization Based on Time Stamp," Computer Engineering, Vol. 33 No. 16, 2007,8

[7] Elena Filatova and Eduard Hovy, " Assigning Time-Stamps to Event-Clauses," Proc. of ACL, 2001

[8] Honglin Sun and ShiWen Yu, "Shallow Parsing Overview" on Contemporary Linguistics, 2nd ed., vol.2, 2000, pp. 74-83

[9] Bird, S, and Loper, E. 2004. NLTK: The natural language toolkit. In Proc. of 42nd Annual Meeting of the Association for Computational Linguistics(ACL-04)