TUTA1 at the NTCIR-12 Temporalia Task

Ning Liu
Tokushima University, Japan
c501547005@tokushima-u.ac.jp

Mengjia He
Tokushima University, Japan
c501437038@tokushima-u.ac.jp

Chao Li
Tokushima University, Japan
c501447002@tokushima-u.ac.jp

Xin Kang
Tokushima University, Japan
kang-xin@is.tokushima-u.ac.jp

Fuji Ren
Tokushima University, Japan
ren@is.tokushima-u.ac.jp

ABSTRACT

Our group submitted task for Temporal Intent Disambiguation (TID) Subtask (Chinese) of NTCIR-2012. We using word2vec to model query String into feature vector, and using cos function to measure the similarity between query string and training corpus SougouCA. Our results shows the approach is efficient for solving thoes Task.

Team Name

TUTA1

Subtasks

Temporal Intent Disambiguation (TID) Subtask (Chinese)

Keywords

Temporal search, Relation extraction,Similarity computing

1. INTRODUCTION

Successful search engines are supposed to consider temporal aspects of information because of the crucial role in estimating information relevance of time [1]. With successful solutions, search engines could then treat temporal queries accordingly to their underlying temporal classes[2].And an analysis [3] shows that users searching for fresh information also seek for information in past as well as future. But the results of search engines like Google, is now greatly depended on the searching history of users or the searching times of keywords. Recently, a news about searching for "Bismarck" by Google has become a hot topic on internet in Japan. In steading of the pictures of the "Otto von Bismarck", Google Japan will give more pictures about a game character because of the high popularity of the game in Japan that, times for searching about the character are far more than the ones for the real person. For getting correct results from Google, the keywords you should try may like "Bismarck 1815" for "Otto von Bismarck", or "Bismarck 1936" for "the battleship Bismarck". So it is important to add time into search engines which has been known as a long time.
Information, "Otto von Bismarck (1815-1898)" or "the battleship Bismarck (1939-1941)", contains temporal aspects which are supposed to be considered in search engines. The information retrieval based on temporal information is proved to be better than the information retrieval of simple text-keyword search in the capabilities of query expression and query processing [4].

Similar with English, some kinds of adverbial of time, "昨天(yesterday)", "现在(now)" and "明天(tomorrow)", are also be used in Chinese.According to the conventional research[5].these kinds of temporal expressions in Chinese can be classified into some classes as "PosDate" expressions ["去年春天(last spring)"], "TempWord" expressions ["春节(Chinese new year)"], "Composite" expressions ["1999年4月28日(April 28, 1999)", "两年后(two years later)"], "Set" expressions ["每年(every year)"], or "EventAn" expressions ["当他演讲时(when he was speaking)"]. However, just like the phrases of verbs "来过(have come) ", "来了(came) " or "将要来(coming)", no tense will be found in Chinese verbs as a considerable feature of this language that is different with English. Otherwise, a message at 15:00 like "今天晚饭吃什么？ (What shall be my dinner today?)", shows that to discriminate the message as a future event, comparison between "15:00" and "晚饭(affirmed to be 15:00 later)" is needed in Chinese while "shall" can tell the answer in English. Different approaches are anticipated to be necessary in Chinese sentences without adverbial of time or messages like the example therefore, which can be a valuable challenge as Chinese is a new processing language in the task, and our work is supposed to be a good beginning.

In this year's TID Subtask of Chinese, we provide an extending method of computing the similarity keywords of object query string in our training corpus sougouCA. For a better vector representation of a word, we use word2vec tool to training a 200 dimension vector. Then we extend the query string by calculating the most similarity ten words in sougouCA. Finally, we take this query string sequence instead of the single query word in Task.

The remainder of this paper is structured as follows. In $2, we will describe our approach proposed for TID subtask. In $3 and $4, we will give our experiment results and some analysises.

2. OUR APPROACH

2.1 Annotation Corpus

In Temporal Intent Disambiguation (TID) Subtask, we will give our results for 300 different queries in temporal attribute.Our corpus for dry run and formal run is called SougouCA which was published by Sougou Labs[6].
This coprsu conrains 762809 news contents,which were col-

```
<doc id="c172394d49da2142-69713306c0bb3300">
<meta-info>
<tag name="host">http://news.sohu.com</tag>
<tag name="date">2012-06-12</tag>
<tag name="url">http://news.sohu.com/20120612/n345428229.shtml</tag>
<tag name="title">公安机关销毁10余万非法枪支 跨国武器走私渐起</tag>
<tag name="source-encoding">UTF-8</tag>
</meta-info>
<text>
中广网唐山6月12日消息(记者汤一亮庄胜春)据中国之声《新闻晚高峰》报道，今天(12日)上午，公安机关
</text>
</doc>
```

Figure 1: The structure of SougouCA

lected from 2012.06 to 2012.07. There are eighteen news tags in this corpus like: national news, international news, entainment, sports .Figure one shows the structure of SougouCA corpus. As can be seen from the figure, tag "id" is the id number of this news, the tag name "host" means the page website of this news, the tag "date" represents the published time of this news, the tag "url" means where the data came from, the name "title" give us the title of the page, and the <text> means the contents of this news.

According to those full of Temporal and named entities parirs news, we can get annotated sentences using "TemporaliaChTagger" [7, 8]. The news contents in SougouCA were annotated on sentences level, and using <SE> </SE> representing the sentenc's begin and end respectively, as show in Figure 2.
The sentence was annotated mainly into two categories: one is time represented as <T>***</T> and the other one is named entity represented as <E>***</E>. For category <T>,its type is "DATE", and which has four values, named "PRESENT","PAST","FUTURE" and a specific date format as "year-month-day". Four of the samples can be shew blew:

1. $<Ttype="DATE"value="PRESENT_REF">$现在(English: now)$</T>$
2. $<Ttype="DATE"value="PAST_REF">$近日(English: recently)$</T>$
3. $<Ttype="DATE"value="FUTURE_REF">$不久(English: soon)$</T>$
4. $<Ttype="DATE"value="2011-08-25">$２０１１年８月２５日(English: 2011.08.25)$</T>$

For category <E>, it has five types, named "ORGANIZATION", "GPE", "PERSON", "LOCATION", and "MISC"(by merging all but the four most dominant entity types into one general entity). The five Typies can be shew in the following annotation samples:

1. $<Etype="ORGANIZATION">$公安部治安局(English: Ministry of Public Security Bureau)$</E>$
2. $<Etype="GPE">$唐山(English: Tangshan)$</E>$

```
<doc id="c172394d49da2142-69713306c0bb3300">
<meta-info>
<tag name="host">http://news.sohu.com</tag>
<tag name="date">2012-06-12</tag>
<tag name="url">http://news.sohu.com/20120612/n345428229.shtml</tag>
<tag name="title">公安机关销毁10余万非法枪支 跨国武器走私渐起</tag>
<tag name="source-encoding">UTF-8</tag>
</meta-info>
<text><SE><E type="ORGANIZATION">中广网</E><E type="GPE">唐山</E><T ty
<SE>黄明:<T type="DATE" value="PRESENT_REF">现在</T>我宣布,全国缉枪制爆统
<SE>与此同时,在全国各省区市<E type="MISC">150</E>个城市,破案追缴和群众主动上
<SE><E type="ORGANIZATION">公安部治安局</E>局长<E type="PERSON">浏绍武</
<SE><E type="PERSON">浏绍武</E>:打击破案包括涉黑、涉恶的团伙犯罪、毒品犯罪,还有
<SE>在销毁现场,记者看到了被追缴和上缴的各式各样的枪支。</SE>
<SE><E type="PERSON">浏绍武</E>:也包括制式枪,有的是军用枪、仿制的制式枪,还有猎
<SE>按照我国的枪支管理法,这些都是严厉禁止个人非法持有的。</SE>
<SE><E type="GPE">中国</E>是世界上持枪犯罪的犯罪率最低的国家之一。</SE>
<SE><E type="GPE">中美</E>联手破获特大跨国走私武器弹药案<T type="DATE" val
<SE>在<E type="GPE">美国</E>抓获犯罪嫌疑人<E type="MISC">3</E>名,缴获各类
<SE>这是公安部与<E type="GPE">美国</E>移民海关执法局通过联合调查方式侦破重大跨
<SE><T type="DATE" value="2011-08-25">2011年8月25日</T>,<E type="GPE
<SE>经检验,这些都是具有杀伤力的制式枪支及其配件。</SE>
<SE>这引起了公安部和海关总署的高度重视。</SE>
<SE><E type="ORGANIZATION">公安部刑侦局</E>局长<E type="PERSON">浏安成</
<SE><E type="ORGANIZATION">上海市公安局</E>和<E type="ORGANIZATION">上海
<SE>专案组于<T type="DATE" value="2012-08-26">8月26日</T>在<E type="GP
<SE>王挺交代,他通过一境外网站上认识了上家<E type="PERSON">林志富</E>,<T typ
<SE>此案中,犯罪分子依托虚拟网络进行犯罪交易,隐蔽性强,涉案人员使用的身份、地址、联
<SE>浏安成说,此案体现了是新型犯罪,特别是现代犯罪的新特点。</SE>
<SE><E type="PERSON">浏安成</E>:他不受距离的限制、经常是跨国跨境,甚至是跨一个
<SE>这种犯罪手法的改变和新型犯罪的特点,要求我们各国警方充分合作。</SE>
<SE>作者:汤一亮庄胜春</SE>
</text>
</doc>
```

Figure 2: The annotation file of SougouCA made by TemporaliaChTagger

3. < *Etype* = "*PERSON*">刘绍武(English: Shaowu Liu)< /*E*>
4. < *Etype* = "*LOCATION*">欧美(English: Europe and America)< /*E*>
5. < *Etype* = "*MISC*">1 5 0< /*E*>

2.2 Word2vec Tool

Word2vec is a tool that takes a text corpus input and produces the word vectors as ootput [9].Given enough data, usage and contexts, Word2vec can make highly accurate guesses about a word ' s meaning based on past appearances.Those guesses can be used to establish a word ' s association with other words,or cluster documents and classify them by topic [10]. We use this tool to produce the word vectors of SougouCA. Table 1 shows the ten similariry words of query string "滑雪(English:Skiing)"of id 037 computing by the model trained by word2vec.

Table 1: Ten similariry words of query string "滑雪"

Word	Similarity	Word	Similarity
登山(English: Climbing)	0.7583161	探险(English: Expedition)	0.51729447
露营(English: Camping)	0.6478083	柳汉奎(English: Ryu Hangyu)	0.5031028
冲浪(English: Surfing)	0.54753125	徒步(English: Hiking)	0.4999488
热气球 (English:Fire Ballon)	0.54753125	冰雪(English: Ice and Snow)	0.497199
观峰(English: Sightseeing)	0.52005875	水上(English: Overwater)	0.492894

When training the word vectors, the main parameters for using word2vec are "-skip-gram -size 200 -windows 5". For every query string $S_{query} = (w_1, w_2, \cdots, w_i, \cdots, w_n)$, in which $w_i$means the word in query string. In our method, the feature vector for query string V_{query} can be represented as formula(1):

$$V_{query} = \sum_{j=1}^{n} v_{word}(j) \tag{1}$$

$v_{word}(j)$ means the vector of the word in query string. n means the number of words in query string.

3. OUR EXPERIMENT SYSTEM FOR TASK

As show in figure 3 and Algorithm 1, for the query string S_{query} in task, using word2vec tool to train word vectors about SougouCA news corpus. For the sentences $S_{<SE>}$ in SougouCA annotated by TemporaliaChTagger, defined the feature vectors of query and training sentences as formular (2) and (3):

$$V_{query} = \sum_{j=1}^{n} v_{word}(j) \tag{2}$$

$$V_{<SE>} = \sum_{k=1}^{m} v_{<SE>}(m) \tag{3}$$

For every query, computing the similarity of $V_{<SE>}$ and V_{query} using *cos* function. Selecting the total ten highest similarity sentences in SougouCA, and using tag "<T,type="

Algorithm 1: Query String Temporalia Computing

Input: Query $S_{query} = (w_1, w_2, \cdots, w_i, \cdots, w_n)$, and annotated sentences $S_{<SE>} = (C_1, C_2, \cdots, C_k, \cdots, C_m)$ in training corpus.
Output: The Temporalia Query series results $T_{query} = (P_{past}, P_{recency}, P_{future}, P_{atemporal})$.

```
index = 0; Num_<SE> = Iterator();
repeat
    V_query = Σ_{j=1}^{n} v_word(j);
    V_<SE> = Σ_{k=1}^{m} v_<SE>(m);
    R_Similarity = cos(V_query, V_<SE>);
    for each sentence <SE> in R_Similarity within ten
    do
        Compare T_<tag> in <SE> with S_date in
        S_query;
        if (T_<tag> < S_date) then
        end
        Num_past+=1;
        if (T_<tag> > S_date) then
        end
        Num_future+=1;
        if (T_<tag> = S_date) then
        end
        Num_recency+=1;
        if Found = FALSE then
        end
        Num_atemporal+=1;
    end
    if Num_*!=Null then
        T_query = (P_past, P_recency, P_future, P_atemporal);
    else
        T_query = (1, 0, 0, 0);
    end
until Num_<SE>.hasnext();
```

"date"> " to match the query timeline to get the final temporal probility of four time categories. For the query string "滑雪(English:Skiing)" in id 037, the results calculated by our method are shew in Table 2:

Table 2: Comparison between temporal Results and grand truth of query string "滑雪(English:Skiing)" in id 037

Item	Past	Recency	Future	Atemporal
Truth	0.1	0.0	0.2	0.7
Results	0.247104	0.274131	0.262548	0.216216

4. RESULTS AND CONCLUSIONS

As shows in table three and table four, our group submitted one run for Temporal Intent Disambiguation (TID) Subtask (Chinese). In the formal run, we proposed a similarity computing method for Temporal information query. In the total 300 queries, we got Averaged Per-Class Absolute Loss(APAL) of 0.271564197983333, and Cosine Similarity(CoS) of 0.607726046083874. For every sigle temporal categories, if the results of a query string contains a temporal category, we mark one time of the certain temporal category. After calculated all of the 300 queries, we get the precision of four temporal categories shown in table four.

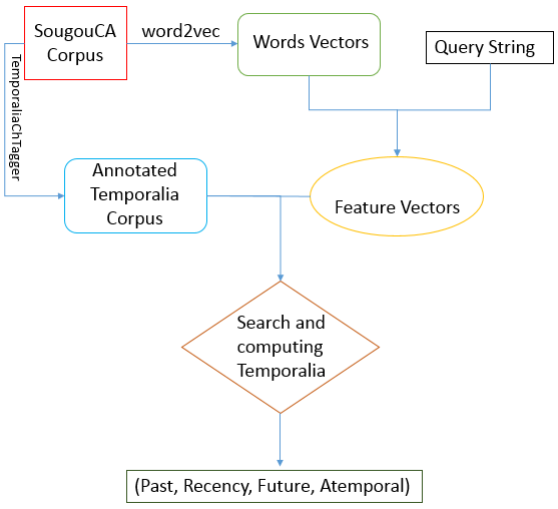

Figure 3: The flow chart for query system

As can be seen in table four , our method get higher precision of 92.7% in "atemporal" category.Through this task, we can found the effictiveness of using multi-label in searching queries, and get a new view about temporal extraction.

Table 3: Precision of four temporal categories

Item	Past	Recency	Future	Atemporal
Precision(%)	52.7	58	39.7	92.7

Table 4: Results of TID Task

Item	APAL	CoS
Avgeraged Value	0.271564197983333	0.607726046083874

5. ACKNOWLEDGMENT

This research has been partially supported by JSPS KAKENHI Grant Number 15H01712.

6. REFERENCES

[1] Hideo Joho, Adam Jatowt, and Roi Blanco. Ntcir temporalia: a test collection for temporal information access research. In *Proceedings of the companion publication of the 23rd international conference on World wide web companion*, pages 845–850. International World Wide Web Conferences Steering Committee, 2014.

[2] Hideo Joho, Adam Jatowt, Roi Blanco, Haitao Yu, and Shuhei Yamamoto. Overview of NTCIR-12 temporal information access (temporalia-2) task. In *Proceedings of the 12th NTCIR Conference on Evaluation of Information Access Technologies, June 7-10, 2016, Tokyo, Japan*, 2016.

[3] Hideo Joho, Adam Jatowt, and Blanco Roi. A survey of temporal web search experience. In *Proceedings of the 22nd international conference on World Wide Web companion*, pages 1101–1108. International World Wide Web Conferences Steering Committee, 2013.

[4] Zhe Wang and Chenggang Xu. The research of web information retrieval based on temporal information. In *National Conference on Information Technology and Computer Science (CITCS 2012)*, 2012.

[5] Wu Mingli, Li Wenjie, Lu Qin, and Li Baoli. Ctemp: A chinese temporal parser for extracting and normalizing temporal information. In *Natural Language Processing–IJCNLP 2005*, pages 694–706. Springer, 2005.

[6] Canhui Wang, Min Zhang, Shaoping Ma, and Liyun Ru. Automatic online news issue construction in web environment. In *Proceedings of the 17th international conference on World Wide Web*, pages 457–466. ACM, 2008.

[7] Jenny Rose Finkel and Christopher D Manning. Joint parsing and named entity recognition. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 326–334. Association for Computational Linguistics, 2009.

[8] TimeML Working Group et al. Guidelines for temporal expression annotation for english for tempeval 2010, 2009.

[9] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[10] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.