Light Verb Construction and Entailment Recognition

Changli Li[1] and Zixian Zhang[*2, 3]

[1]School of Management & Communication, Capital University of Physical Education & Sports, Beijing, China

[2]School of Mechanical, Electronic & Information Engineering, China University of Mining & Technology, Beijing, China

[3]School of Foreign Languages, Liaocheng University, Liaocheng, China

1lichangli01@126.com,2zhangzixian@lcu.edu.cn

Keywords: natural language processing, entailment, light verb construction, entailment rule, answering system

Abstract: Light verb construction is a collocational phrase that follows the formula "take a+noun" (to take as an example in the research). From a light verb construction like "take a break", we can get the entailment rule "take a break→break (or synonyms of break)". On the other hand, entailment relation recognition is an important task for natural language processing. So entailment from the viewpoint of light verb construction is focused for study. Light verb construction appears in a certain context, thus statistic method is applied to find the possible subsequent constituents after the light verb construction in a sentence. Consequently, at first insight for light verb construction recognition is provided. It is found that there is a higher possibility for the existence of light verb construction if an adverb or a clause follows it. And ten most probable nouns are found that appear in the light verb construction with the pattern "take a + noun phrase". Second, based on the above analysis, entailment rules with one argument open are found and a corpus of sentence pairs with entailment relation is made. This provides insight for such natural language processing tasks as answering system and information extraction.

1. Introduction

There are three kinds of relations between two texts or sentences, namely paraphrase, entailment and elaboration. Among the three relations, entailment is a main topic for research in the field of natural language processing. And the most widely studied area in this field is Recognizing Textual Entailment (RTE) (I. Dagan and O. Glickman, 2004). In the task of Recognizing Textual Entailment there are two pieces of texts, i.e. T(abbreviation for Text) and H(abbreviation for Hypothesis), and the task aims to find out whether H can be inferred from T. To put it formally, the task aims to find out if the formula T⇒H exists. From the year 2005 until now, many rounds of conferences focusing on RTE have been held, with a view to providing methods for Recognizing Textual Entailment. And the results of the research are universal and can be used in many natural language processing tasks, such as Question Answering(QA), Information Retrieval(IR), Information Extraction(IE), Automatic Summarization(AS), machine translation evaluation, etc (M. Wang, S. Yu, and X. Zh, 2015:151-156).

All aspects of language need to be studied for the purpose of Recognizing Textual Entailment. Among them, multiword expression is the mostly studied part of language. The different parts of a multiword expression are arranged not in a random way but rather in a systematic method. In fact the way they are arranged needs to be made clear and thus ambiguity can be got rid of. Therefore, multiword expression has always been a heated topic of discussion in natural language processing. In a sentence the verb is usually regarded as the main part and tells the action indicated in the sentence. It is generally acknowledged that there are three kinds of verbal phrases: light verb construction, phrasal verb construction and embedded verb construction. In terms of meaning, light

  DOI: 10.25236/icecbn.2018.035

verb construction is semi-compositional, and its meaning comes mainly from the noun object in the phrase whereas the verb acts as the predicate and makes the sentence consistent and sound natural. We can take the following collocation as examples: “have a look”, “make an offer”, “take a bath”, etc. This is usually studied in the frame of templates (F. M. Zanzotto, M. T. Pazienza, Marco Pennacchiotti, 2005:37-42). In the above examples, the light verb here is somewhat unimportant in meaning. Such larger linguistic units as sentences are also studied with a view to recognizing entailment relations, but this is usually based on word- or phrase-level entailment relation.

In this research light verb construction is focused to analyze entailment relation between sentences, and methods for entailment recognition are provided. Based on previous literature, an algorithm for making a corpus of sentence pairs with entailment relation is made. The research results will provide insight for recognizing entailment tasks in natural language processing, and the corpus thus made can be used in various natural language processing tasks.

2. Related Work

Entailment refers to the semantic relation between two texts. If a hypothesis H can be inferred from a text T, entailment is said to exist between them. It is universally acknowledged that there are three kinds of entailment: positive, negative and non-entailment. In positive entailment, H can be inferred from H, in negative entailment T and H contradicts with each other, whereas in non-entailment there is no relation of one kind or another between T and H. From a hierarchical point of view, there are three layers of research: lexical, syntactic and semantic. Among the three layers, most researchers do the research from the lexical or syntactic point of view.

On the word level, cosine value or word vector is usually used to compute entailment relation between two words, and sentence-level entailment is computed as the sum of word entailment. Weight is assigned to each word according to its frequency in a specific corpus. Entailment between sentences in the Chinese language is also studied (Z. Zhang, H. Zhou, D. Yao, X. Lu., 2016:169-174) through the method of word vector. Elements in the word vector are features of the target word, including its part of speech, argument structure, etc. On phrase or sentence level, entailment relation computation requires more time and space on the computer. On this level, the Chinese language is also studied. For example, “Lu Xun wrote influential works such as ‘*Crying Out*’ and ‘*Diary of a Madman*’ entails “Lu Xun is author of ‘*Diary of a Madman*’” (S. Ni. S. Ni., 2013:125-129). In recent years, the perspective of context is chosen as the target of study for entailment relation computation, but this needs knowledge and theory from linguistics. For example, a topic model can be made for text T and hypothesis H to determine the entailment relation between them (H. Ren, Y. Sheng, W. Feng, M. Liu, 2015:119-126). In the research of entailment relations in English texts, lexical resources such as WordNet are usually used to obtain entailment between words (P. Pakray, 2011). Entailment relation study is usually under the framework of a particular grammar structure to acquire the relation between words. Generally speaking there are two grammatical frameworks in linguistic study: dependency grammar and phrase structure grammar. The former focuses on the relation between words in a sentence with the main verb as the head, and it is often used in natural language processing tasks B. (Ofoghi, J. Yearwood, 2010).

In entailment research, textual entailment recognition is the main focus, and it is generally acknowledged that there are two kinds of methods: logic-based and graph-based. From the perspective of etymology, the word “entailment” comes from logic (M. Tatu and D. Moldovan, 2005:371-378). Therefore, it is natural to study entailment from the perspective of logic, and logical rules are usually made to carry out textual entailment recognition. In the field of logical inference, there are many methods for textual entailment recognition, and the ones often used by researchers are first-order logic (E. Akhmatova, 2005; A. Fowler, B. Hauser, D. Hodges, I. Niles, A. Novischi, and J. Stephan, 2005), discourse representation theory (J. Bos and K. Markert, 2005:628-635) and neo-Davidsonian-style quasi-logical form (R. Raina, A. Y. Ng, and C. D, 2005), etc. In this method there is a strict prerequisite: logical rules are required. But the precision and recall under this method are not very high. In the graph-based method, a graph is first made for both text and hypothesis under research. In computing the entailment relation, graph for the text is compared to

the graph for the hypothesis, thus textual entailment recognition is reduced to graph comparison (B. MacCartney, T. Grenager, M.-C. de Marneffe, D. Cer, and C. D. Manning, 2006). In this method, the comparison or alignment between T and H is a NP-complete problem, and it is impossible for this to be fully achieved. This is the reason why precision of the recognition by this method is relatively low as compared to other methods.

Basically speaking, the nature of language should be studied further in order to improve textual entailment recognition efficiency. In this respect, different layers of language should be analyzed and the respective rules should be obtained. To put it more specifically, language should be studied from three levels: word-, phrase- and sentence-level. In each level, the meaning rather than form should be focused on. This research starts from the light verb construction in English sentences. Based on light verb construction corpus, a corpus of entailment rules is constructed and method for entailment recognition based on light verb construction is provided.

3. Methods and Materials

3.1. Methods and corpus design

Entailment relation between two English sentences is focused for research, and light verb construction is the main part of the English sentences. Therefore, a corpus of English sentences with light verb construction as the core is designed as the first step of the research. The underlying idea is that sentence pairs with entailment relation and then entailment rule can be obtained from the English sentences with light verb construction. Consequently methods for entailment recognition regarding such sentences can be drawn. The corpus comes from Wikipedia articles (as written in January, 2013). Altogether 8500 articles were chosen, and the context features of light verb construction were studied from a statistical point of view.

3.2. Materials and experiment design

It is generally acknowledged the commonly used light verb in English are “do”, “get”, “give”, “have”, “make” and “take”. In this research “take” is chosen as the target of study and sentences with “take a”, “take an”, “taken a”, “taken a”, “took a”, “took an” are selected from the corpus mentioned in 2.1. In total 2000 sentences were chosen. Then all the sentences were tagged with “y” if there is a light verb construction in it and with “n” if there is no light verb construction in it. This work is done manually by postgraduate English majors in China. The tagging is done by two groups. If the two groups tag one sentence with the same sign, then the sign is given to the sentence. If the two groups give different signs to the same sentence, another English major will be found to tag the sentence. And experts will be found to decide the sign for very controversial sentences. At last 1000 sentences were chosen for study, and the sign for all of them was definite. Among the 1000 sentences, 500 were signed with “y”, and the other 500 were signed with “n”.

Consequently part of speech of the words after the light verb construction is analyzed. The tagging of the part of speech here is done manually. Automatic tagging is subject to errors. Manual tagging is free from such errors. After analysis it is found that words after the light verb construction in an English sentence will probably be assigned such part of speech as preposition, adverb, adjective, conjunction, article, determiner, and other possible elements are clause, infinitive, punctuation, gerund, or participle. The specific distribution is shown in Table 1.

Table 1 Part of speech distribution after light verb construction.

pos	prep	adv	adj	conj	clause	article	det	inf	punc	Ving	Ved
total	689	31	5	56	27	1	4	62	101	15	9
positive//negative	313/376	19/12	2/3	19/37	16/11	1/0	2/2	21/41	37/64	5/10	3/6

4. Experiment Result Analysis

As shown in Table I, there are eleven kinds of elements after light verb construction. In terms of total, preposition takes up most of the examples with 68.9% of the whole training set. The numbers

of adjective, article and determiner are 5, 1, and 4 respectively, and the percentages are 0.5%, 0.1% and 0.4% respectively. We can find that the three classes take up so small a percentage that they are not included in the research. As for the other 8 kinds of elements, we computed the percentage between positive and negative examples as shown in Table 2. As shown in Table 2, when the element after the construction "take a+noun" (including variation of "take") is an adverb or a clause, the ratios between positive and negative examples are 61.29% and 59.26% which are relatively large values. Therefore, the feature of whether the following element is an adverb or a clause can be considered as judge for light verb construction. As for the case of "preposition", the total is larger than any other cases, but this feature is not discriminatory due to the balanced positive/negative ratio. However, due to the relatively large total of prep feature, it is necessary to do detailed research into the prep feature. It is found that among the 689 cases of prep, 313 cases are positive light verb construction. In the positive examples, the nouns in the light verb construction were analyzed, and the commonly used nouns are shown in Table 3.

Table 2 Positive and negative ration after light verb construction.

pos	prep	Adv	conj	clause	inf	punc	Ving	Ved
total	689	31	56	27	62	101	15	9
p/n	313/376	19/12	19/37	16/11	21/41	37/64	5/10	3/6
p/t ratio	45.43%	61.29%	33.93%	59.26%	33.87%	36.63%	33.33%	33.33%

Table 3 Most commonly used nouns in light verb construction (totalled 500)

noun	break	turn	trip	Approach	view	lead	tour	step	stand	Share
frequency	37	22	19	18	17	16	14	14	13	10

After analysis of Table 3, it can be concluded that if the noun in "take + a/an + noun phrase" is one of the nouns shown in Table III, the formula can be considered as a light verb construction. And when light verb construction appears, the meaning of the whole construction is determined by the noun in it whereas "take" is the predicate and its meaning is bleached. For example, in the light verb construction "take a break from performing", "break" is the semantic whole of the pattern "take a break", and the whole phrase can be rewritten as "break from performing". As for the ten nouns shown in Tab.3, if any one of them appears in the formula "take a+noun", the whole meaning of the formula can be determined by the noun in it. And the formula is a light verb construction. From the viewpoint of semantic category, the ten nouns in Table 3 are all deverbal, or to put it another way, they are transformation from verbs.

5. Database Construction

5.1. Database construction for entailment rules

The entailment relation between two sentences can be obtained through the recognition of entailment rules. An entailment rule works like this: X→Y, in which Y can be inferred from X. From Table III, we can get an entailment relation: take a break→ break, take a turn→turn, etc. From the relation here, we can get one-way entailment rules:

X takes a break→X breaks① (1)

X takes a turn→X turns② (2)

As for the rule (1), we get the synonyms for the noun ① in Nomplex: word11, word12, word13, then more entailment rules can be obtained from (1):
X takes a break→X word11, X takes a break→X word12, X takes a break→X word13…

As for the entailment rule (2), we carry out the same process, and more entailment rules are obtained as follows:
X takes a turn→X word21, X takes a turn→X word22, X takes a turn→X word23…Based on the entailment rules thus constructed, entailment relation between two sentences can be determined.

5.2. Database Construction for sentence pairs with entailment relation

All English sentences have either a subject-predicate structure or a subject-link-compliment structure. As for a sentence with subject-predicate structure, the predicate is the main component. Therefore, if the predicates of two English

Sentences have entailment relation, we can say the two English sentences have entailment relation. Based on the above logic, according to the entailment rules shown in part IV, sentence pairs with entailment relation can be obtained. In order to construct a database of such sentence pairs, we need to formalize the process. If there is a light verb construction L_i in sentence S_i and the object in L_i is O_i, we replace L_i with the verb OV_i which is derivative of O_i. The new sentence thus obtained is entailed by S_i. If L_i is not a light verb construction and the verb in S_i is V_i,we replace V_i with its synonyms. The new sentence thus obtained is also entailed in the original S_i. Based on the research in (Y. Tu, 2012), we made the specific algorithm for construction of sentence pairs with entailment relation as shown in Table 4.

As shown in Table4, several sentences may be entailed in one sentence. To put it another way, sentences that sum up to n may be entailed in S_i. The sentences entailed in S_i are paraphrase to each other. From a broad point of view, this paraphrase is also a kind of entailment.

Table 4 Construction algorithm for sentence pairs with entailment.

Input:sentence Si with positive or negative light verb construction sign Output: sentence nS_i which is entailed in S_i
1: for the light verb construction in sentence S_i, do
2: if isTrueLVC(L_i) == True then
3: search Wordnet and Nomlex, and find the verb OV_i that corresponds to L_i's object
4: search Wordnet and Nomlex, and find synonym set syn_i that corresponds to OV_i
5: for each $v_i \in syn_i$, $v_i \neq V_i$ do
6: $L_i \leftarrow v_i$
7: output sentence pair S_i and nS_i
8: endfor
9: else
10: search Wordnet and Nomlex, and find synonym set syn_i that corresponds to V_i
11: for each $_i \in syn_i$, $v_i \neq V_i$ do
12: $V_i \leftarrow v_i$
13: output sentence pair S_i and nS_i
14: endfor
15: endif
16: endfor

6. Conclusion

From the linguistic point of view, entailment relation recognition between sentences concerns such main categories as nouns, adjectives and adverbs. It is impossible to include all the aspects here into research. In this research light verb construction is studied for entailment relation cognition with verbs, a view to providing help for such natural language processing tasks as question answering, information extraction, etc. The innovation of this paper is three ways as follows. Firstly, recognition methods for light verb construction are provided based on the parts of speech of words after the light verb construction. "take a+noun" is taken as an example here. When light verb construction is recognized, entailment rules come into being. And this is the second profit of the research. This part needs such dictionaries as WordNet for synonyms. Thirdly, an algorithm is provided based on previous literature to construct sentence pairs with entailment relation. The corpus of sentence pairs thus constructed can be used in many natural processing tasks. Language is quite a complicated system, and in entailment relation research more linguistic categories should be included, such as adjectives, adverbs, etc. And the nature of language should be studied further in order to provide rules for the computer to process language. To put it more specifically, methods for

formalization of linguistic elements should be searched in future research. Only in this way can statistical methods show their efficiency.

References

[1] Abraham Fowler, Bob Hauser, et al, 2005. Applying cogex to recognize textual entailment. In Proceedings of the PASCAL RTE Challenge Workshop.

[2] Bill MacCartney, et al, 2006. Learning to recognize features of valid textual entailments. In Proceedings of HLT-NAACL.

[3] Bahadorreza Ofoghi, John Yearwood., 2010. Learning parse-free event-based features for textual entailment recognition,. AI 2010: Advances in Artificial Intelligence in Lecture Notes in Artificial Intelligence, Volume 6464.

[4] Elena Akhmatova, 2005. Textual entailment resolution via atomic propositions. In Proceedings of the PASCAL RTE Challenge Workshop

[5] Fabio Massimo Zanzotto, Maria Teresa Pazienza, Marco Pennacchiotti., 2005. Discovering entailment relations using textual entailment patterns. In Proceedings of the ACL Workshop on Empirical Modeling of Semantic Equivalence and Entailment, pp.37–42.

[6] Han Ren, Yaqi Sheng, Wenhe Feng, Maofu Liu., 2015. Recognizing textual entailment based on knowledge topic models. In *Journal of Chinese Information Processing*, 29(6), pp.119-126.

[7] Ido Dagan and Oren Glickman, 2004. Probabilistic textual entailment: Generic applied modeling of language variability. In PASCAL workshop on Text Understanding and Minin.

[8] Johan Bos and Katja Markert, 2005. Recognising textual entailment with logical inference. In Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing, pp.628-635.

[9] Marta Tatu and Dan Moldovan., 2005. A semantic approach to recognizing textual entailment. In *Proceedings of Human Language Technology Conference and Conference on* Empirical Methods in Natural Language Processing, pp. 371-378.

[10] Meng Wang, Shiwen Yu, and Xuefeng Zhu. 2015. Natural language processing and its applications in education. In *Mathematics in Practice and Theory*, 45(20), pp.151-156.

[11] Parthar Pakray., 2011. Answer validation using textual entailment. In Computational Linguistics and Intelligent Text Processing, PT II in Lecture Notes in Computer Science, Volume 6609.

[12] Rajat Raina, Anrew Ng, and Cristoper Manning., 2005. Robust textual inference via learning and abductive reasoning. In Proceedings of the Twentieth National Conference on Artificial Intelligence AAAI).

[13] Shengjian Ni., 2013. A survey of the study of textual entailment and its development tendency. In *Journal of Yunnan Nationalities University* (Social Sciences) 30(4), pp.125-129.

[14] Yuancheng Tu., 2012. English complex verb constructions: identification and inference, In Doctoral Dissertation, University of Illinois at Urbana-Champaign.

[15] Zhichang Zhang, Huixia Zhou, Dongren Yao, et al, 2016. Recognition of Chinese lexical entailment relation based on word vector. In *Computer Engineering*, 42(2), pp. 169-174.