ATLANTIS PRESS
Advances in Engineering Research (AER), volume 130
5th International Conference on Frontiers of Manufacturing Science and Measuring Technology (FMSMT 2017)

Retrieving Collocation Frameworks for Entity Attribute Knowledge Acquisition

Hong-lin Wu[1,a], Ruo-yi Zhou[2,b] and Ke Wang[1,3,c]

[1] College of Computer Science and Engineering, Northeastern University, Shenyang, China

[2] School of Information Engineering, Zhengzhou University, Zhengzhou, China

[3] Research Center for Artificial Intelligence, Shenyang Linge Technology Co., Ltd., Shenyang, China

[a]wuhl@mail.neu.edu.cn, [b]zhouruo.yi@qq.com, [c]flyingegg.ke@gmail.com

Keywords: Retrieving, Collocation, Entity Attribute.

Abstract. The key problem in the acquisition of the entity attribute knowledge for natural language understanding lies in the connections between the entity attributes. These connections could be represented by entity attribute collocations. It is impossible to get these entity attribute collocations manually. This paper proposed a method of retrieving collocation frameworks for entity attribute knowledge acquisition, which could acquire the entity attribute collocations from real corpus automatically. Because the collection framework template is actually the simplest syntactic sub-tree which retained the core verbs and the brother branch of the entity word and the attribute around the core verb. The proposed method obtained the entity attribute collocations based on the pruning of the syntactic tree. The experimental result showed that the proposed method performance well on the real corpus.

Introduction

The key problem in the acquisition of the entity attribute knowledge for natural language understanding lies in the connections between the entity attributes. These connections could be represented by entity attribute collocations. It is impossible to get these entity attribute collocations manually. This paper proposed a method of retrieving collocation frameworks for entity attribute knowledge acquisition, which could acquire the entity attribute collocations from real corpus automatically. Because the collection framework template is actually the simplest syntactic sub-tree which retained the core verbs and the brother branch of the entity word and the attribute around the core verb. The proposed method obtained the entity attribute collocations based on the pruning of the syntactic tree.

Syntactic Analysis

Syntactic analysis is the key technology of deep processing and analysis of language in natural language processing tasks, which has been a major area of research within computational linguistics for decades. The main task of syntactic analysis is to determine the grammatical units and the relationship between these units contained in the given legal sentence according to the given grammar system, and give a formal representation of these units and their relationships. There are two main tasks of syntactic analysis. One is to establish a language model for grammatical disambiguation. The other is to output the most probable structure of all possible sentence structures. In general, the sentence structure of the syntactic analysis is expressed and stored in the form of a syntactic tree as showed in Fig. 1.

Fig. 1 showed a syntactic tree of a sentence that describes the relationship between the computer entity and its attributes. After segmentation and entity/attribute recognition, the sentence had been put into the Berkeley Parser to obtain the syntactic tree. It can be seen that the syntactic analysis assigned a definite grammatical appellation to each word in the input sentence and gave a hierarchical relationship between these grammatical units in the form of grammar tree.

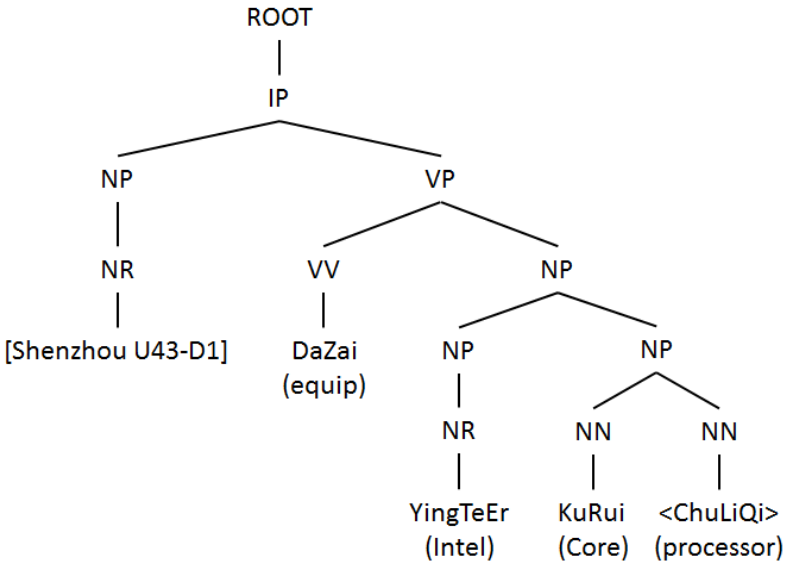

Fig. 1 An example of syntactic tree

Syntactic Tree Collocation Analysis

The orignal sentence in Fig. 1 is “ShenZhou U43-D1(a type of computer manufactured by Shenzhou Company) DaZai(equipped with) YingTeEr(Intel) KuRui(Core) ChuLiQi(processor)”. The entities and attributes in the sentence constitute the subject of the framework we expect to acquire. And the direct expression of such framework is the collocation relationship in the syntactic tree. The representation of the framework could be defined as a template with a semantic relationship. Such template is a reflection of the relationship of entities-attributes or attributes-attributes which based on treating the verb as the core.

According to the definition of the string-level template, we could apply the shallow lexical syntax analysis (such as: word segmentation and part of speech tagging) in the given sentence, and generate the template by the string-level matching driven by the attribute words. The template of the above example would be “[#entity DaZai/v YingTeEr/n KuRui/n #attri]”. Although the template contains the core verb that describes the relationship between the entity and the attribute, it also contains noise terms, such as: YingTeEr and KuRui. These noise terms are not related to the collection relationship between the entity and the attribute. This would reduce the versatility and the expression ability of the template.

Through the syntactic analysis, from the grammar tree we can clearly see the noise terms and attribute words in the same local sub-tree. The two components constitute a noun phrase together. The verbs that related to the entities and the attributes act directly on the noun phrases formed by the sub-tree. If the sub-tree of the noun phrase only retains the branch of the attribute word, that is, cut off the branch of the noise term, we can get only the meaningful statement that expressed the relationship between the entity and the attribute. This statement is not only grammatically legal, but also meets the requirements of the acquisition of the collection framework. It contains the core verb that restricted the relationship between entities and attributes. By re-traverse the pruning syntactic tree, we could get the template met the requirements: “[#entity DaZai/v #attri]”. So we would use the method based on the pruning of the syntactic tree to obtain the collection framework template of the entity and the attribute.

Retrieving Collocation Framework Templates from Syntactic Tree

The intention of the collection framework template we required is to reflect the description of the relationship between the entity and the attribute. The grammatical feature of the template is treating verbs as the core, and using the entity or attribute as the slot. Reflecting in the syntax tree, the main

characteristic of this grammatical feature is to obtain a simplest tree consisted of the noun phrases of the entity words or the attribute words which dependent on the core verb.

For the example showed in Fig. 1, syntactic tree consisted of two noun phrases. One is a noun phrase of entity words: “ShenZhou U43-D1 (a type of computer manufactured by Shenzhou Company)”. The other is a noun phrase of attribute words: “YingTeEr(Intel) KuRui(Core) ChuLiQi(processor)”. The core verb of the syntactic tree is verb “DaZai(equip)”. This syntactic tree is not the simplest tree, because the verb “DaZai” acts on the whole sub-tree of its right child branch. In this sub-tree, there is only one partial component “ChuLiQi(processor)” as the center of the local phrase. It is the object on which the core verb acted, and the other parts should be removed as the noise.

So, collection framework template we required is actually the simplest syntactic sub-tree which retained the core verbs and the brother branch of the entity word and the attribute around the core verb. The collection framework extracting process can be described as obtaining the template based on the pruning of the syntactic tree. Follow such idea, the collection framework template extracting procedure could be described by an example sentence “ZhuBan(mainboard) CaiYong(utilize) TeShu(special) De(a Chinese particle word, used in the back of the attribute) GongDian(power supply) SheJi(design)”. By the Berkeley Parser, we could get the syntactic analysis result as “((IP (NP (NR <ZhuBan>)) (VP (VV CaiYong) (NP (CP (VA TeShu) (DEC De)) (NP (NN GongDian) (NN <SheJi>))))))”. According to the syntactic analysis result the syntactic tree could be constructed as shown in Fig. 2. Based on the pruning of the syntactic tree, we could obtain the final template: “[ZhuBan/attr CaiYong SheJi/attr]”.

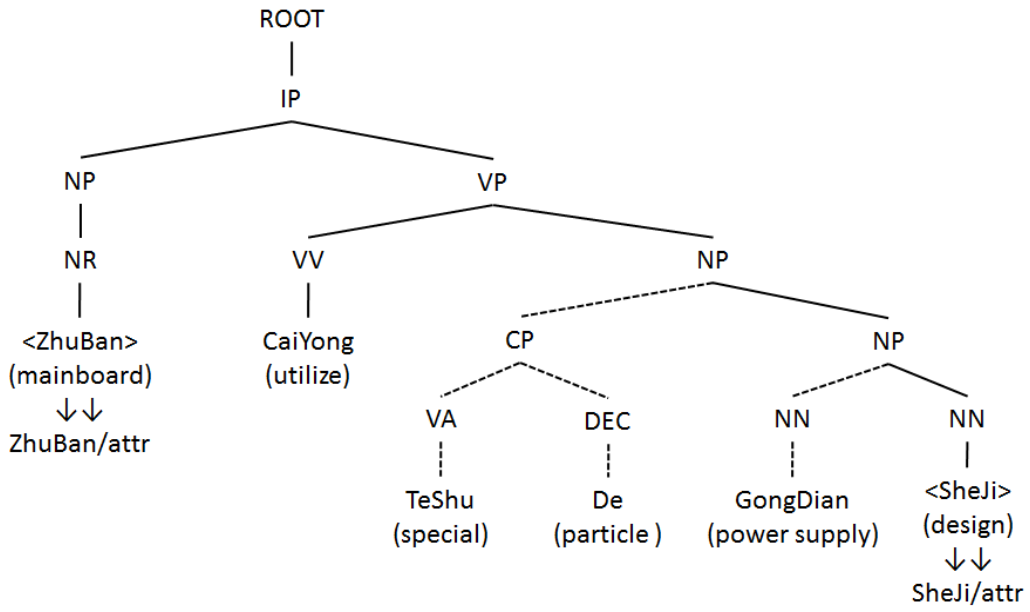

Fig. 2 An example of pruning a syntactic tree

Experimental Results

We collected a corpus with 100 computer commentary articles as the testing set. There are 364 sentences containing the entity words and the attribute words. In these 364 sentences, the system retrieved 276 collocation framework templates. Where, 214 of them are right templates. The system achieved precision of 78% and recall of 59%.

Conclusions

The key problem in the acquisition of the entity attribute knowledge for natural language understanding lies in the connections between the entity attributes. These connections could be

represented by entity attribute collocations. It is impossible to get these entity attribute collocations manually. This paper proposed a method of retrieving collocation frameworks for entity attribute knowledge acquisition, which could acquire the entity attribute collocations from real corpus automatically. Because the collection framework template is actually the simplest syntactic sub-tree which retained the core verbs and the brother branch of the entity word and the attribute around the core verb. The proposed method obtained the entity attribute collocations based on the pruning of the syntactic tree. The experimental result showed that the proposed method performance well on the real corpus.

Acknowledgements

This research was financially supported by the National Natural Science Foundation of China (61370155).

References

[1] F. Zhang, Z.M. Ma, J.W. Cheng, Enhanced entity-relationship modeling with description logic, Knowledge-Based Systems, 93(2015)12-32.

[2] F. Zhang, Z.M. Ma, J.W. Cheng, A survey on fuzzy ontology for the semantic web, Knowledge Engineering Review, 3(2016)1-44.

[3] Y. Yin, GDC: A robust tag recommendation algorithm, Journal of Computational Information Systems, 22(2015)8061-8069.

[4] F. Smadja, Retrieving collocations from text: Xtract, Computational Linguistics, 19(1993) 143-177.

[5] K. Church, P. Hanks, Word association norms, mutual information, and lexicography, Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics, (1987)76-83.

[6] H.L. Wu, R.Y. Zhou, K. Wang, Knowledge representation of entity attribute frame for natural language understanding, in press.

[7] K. Wang, H.L. Wu, Research on neologism detection in entity attribute knowledge acquisition, in press.

[8] H.L. Wu, R.Y. Zhou, K. Wang, Template based attribute value words acquisition in entity attribute knowledge base construction, in press.

[9] N. Chomsky, G.A. Miller, Introduction to the formal analysis of natural languages, Handbook of Mathematical Psychology, 2 (1962) 269-321.

[10] C.E Shannon, Communication theory of secrecy systems, Bell System Technical Journal, 28 (1949) 656-715.

[11] S. Abraham, K. Ferenc Kiefer, A Theory of Structural Semantics, Mouton & Co., Hague, 1967.

[12] C. Fellbaum, M. Palmer, H.T. Dang. L. Delfs, S. Wolf, Manual and automatic semantic annotation with WordNet, Proceedings of the NAACL 2001 Workshop on WordNet and Other Lexical Resources, (2001) 405-414.