Classification and Extraction of Appositives based on English and Chinese Bilingual Corpus

Zhenzhen Xiang

School of Foreign Languages, Guizhou Minzu University, Guiyang, China

sandy.xiang@qq.com

Keywords: English-Chinese translation; appositive; semantic classification; extraction; automatic recognition

Abstract. This paper studies sentence translation in English and Chinese bilingual corpus, analyzing appositive mode with automatic identification and extraction from the perspective of computer linguistics. Firstly, based on semantic concept in linguistics, this paper classifies the basic semantic types of appositives and systematically introduces the diversity of these structures. Natural language and graphics are employed to describe the algorithm. Examples of English and Chinese sentences can be embodied in the use of the Deterministic Parsing method to determine the process of the translation and extraction of appositive structure. This paper aims to not only promote the development of linguistic concepts, but also provide models for computer programmers to deal with natural language and promote the development of computational linguistics.

Introduction

The establishment of the English-Chinese corpus [1] can provide a research platform to study the syntactic and semantic attributes of the English-Chinese sentences and the similarities and differences between English and Chinese. The abundant language information resources and corpus extraction in corpus not only provide language materials for both English and Chinese teaching [2, 3, 4], but also promote the development of machine translation [5]. The summarized appositive modes of English and Chinese structures have direct application value in English and Chinese translation. In particular, the research can help Chinese English learners to summarize the characteristics and patterns of appositives, which in turn help them to extract appositives automatically and analyze the language chunks next to appositives. The use of natural language processing technology can help people save time and energy to further study comparative linguistics. In this way, the traditional English-Chinese comparative linguistics research methods are supplemented and developed further. Corpus becomes more and more mature in natural language processing research. Based on traditional corpus linguistic method, it would be more objective and effective when computational linguistic method is introduced in corpus research.

This paper aims to compare and analyze the similarities and differences between English and Chinese appositive both in form and content to establish appositive modes. It is done by randomly selecting English-Chinese translated corresponding sentences in an English Chinese bilingual corpus in which part-of-speech of the corpus has been labeled. According to the established appositive mode, natural language processing technology is used to automatically identify the corresponding appositive structures and extract the information in English-Chinese bilingual translation sentences.

The Appositives in English and Chinese

English appositives are noun modifiers. Appositives can be words, phrases, or clauses. An appositive usually goes after a modified antecedent, specifying nouns or noun phrases of identity, job title, appellation, etc. Appositives are usually embodied in the following way: the appositives are separated from their antecedents by the commas, indicating the ordinary appositive relationship. And sometimes the dash or the colon can be used with appositives to emphasize the role of appositive or to make longer pause. Sometimes certain leading words precede appositives to show the special meaning between appositive components. The appositive structures can be nouns and their phrases (sometimes separated by comma), nouns (separated by comma), infinitives (sometimes separated by comma), adjectives and their phrases (separated by comma), prepositional phrases (sometimes separated by comma), nominal clauses (the preceded antecedents are usually with definite articles).

Appositives refer to two nouns or a noun and a pronoun which are in the same grammatical position among the sentences, and is semantically anaphora. The "appositive" focuses on the form, demanding that the two parts must be in the same functional position; while the anaphora is more as a semantic concept, which is related to the pragmatic factors. Until the late 1970s and the years after, structuralism dominates Chinese grammar and focuses more on language forms. Terms related to the “appositive” phenomenon became more and more, such as “co-occurrence phrase”, "appositive structure”, “appositive phrase” and so on. “Noun and Noun” appositive is often divided into the upper front type and lower front type. The upper front type is neutral in expressions, which is not influenced by other factors in expressions. The lower front type has specific functions in expressions and their organization is influenced by other factors in expressions. The upper front type has certain functions as a modifier. It has the following characteristics: First, in Chinese, there is no “de” between two nouns of this type of expression. Because “de” functions more as an affiliation word instead of a connection word. Second, the modifier in such appositive structure always modifies the preceding word instead of the entire co-ordinate structure. Third, both two nouns can be expanded and are separated by comma. Fourth, under certain conditions, the antecedent and the latter item of such appositive structures can change place with each other. Fifth, this type of appositive structure can be nested to use. The lower front is also divided into the following three types: condensed type, reduced type and class type.

Based on the above comparison, the English and Chinese appositive structures have obvious similarities in the following ways: 1) cities, rural areas, mountains, rivers, lakes, wharfs, hotels, shops in both English and Chinese can be used as appositives. 2) a route, a movement, a period, starting and ending points, interacting points and other central units in both English and Chinese can be used as appositives. 3) animal names in both English and Chinese can be used as appositives. In a word, their constituents in above appositive structures have the relationship of superordinate and hyponymy. Therefore, composition of superordinate and hyponymy is the most obvious commonalities in English and Chinese appositives.

Appositive Mode Analysis Based on Automatic Identification and Information Extraction

It is obvious that the English language belongs to the typical inflectional language with rich morphological changes. But Chinese lacks the strict morphological changes. In English, the constitution of appositive is more extensive and broader. Therefore, some structures are considered as appositive structure in English, while the corresponding one may be only an adjective phrase in Chinese instead of an appositive. From the perspective of computational linguistics, English appositives can be divided into three kinds as follows: they are appositives with a hyphen, quoted appositives and appositives without a hyphen.

From the attributes of nouns, modified nouns and appositive nouns are divided into two categories: “common noun + common noun” and “common noun + proper noun”. And then based on the accordance of gender and category of the two nouns constituting appositives, appositives with hyphen are classified into the same gender and the different gender between the appositive noun and the modified noun. N is used to represent nouns, the upper right corner of the l and 2 respectively stand for the modified noun and appositive noun, the lower right corner i (i = l, 2, 3, 4, 5, 6) stands for the case of the noun. The form “adjN” indicates that the adjective form is corresponding to the noun meaning. Take “common noun-common noun” as an example, the model is N^1_i-N^2_i, N^1 and N^2 are the same gender, N^1_i-N^2_i→N^2N^1 (English-Chinese conversion, N^1_i-N^2_i is English appositive type, and N^2N^1 is the corresponding Chinese word order). Another example is proper noun-common noun. Proper nouns generally indicate the names of people or things, by the form of “proper noun+ common noun”, they constitute the appositive. Common noun appears as appositive, indicating what the proper noun is. The model is also N^1_i-N^2_i type, N^1_i-N^2_i→N^1N^2.

Such quoted appositives do not need to be changed with the case of the modified noun. In the quotation, the appositive can be a single noun or it can be two or a phrase with more words. Whether they are single nouns, phrases, or complete sentences, when they are in a quotation mark and act as an appositive component in a sentence, they simply represent the name or the exact meaning of the noun that is modified before the quotation marks. And the appositive within the quotation mark is never changed. An appositive phrase in a sentence can be regarded as a N_iZ_i type, which can be subdivided into two types: $N_iZ_i \rightarrow ZN$ and $N_iZ_i \rightarrow NZ$.

Appositives without hyphen are generally divided into two categories. And the order of the two nouns are “common noun + proper noun”. Therefore, this type can be divided into the following categories: $N^1{}_iN^2{}_i$ type， $N^2{}_iN^1{}_i$type and $N^2{}_iN^l{}_i \rightarrow N^2N^1$ type.

Automatic Identification of Appositives

The statistical method used in this paper is based on Hidden Markov Models [6] to recognize the part-of-speech of appositives. As a simple and effective statistical tool, Hidden Markov Models has been widely used in many fields, such as natural language processing [7, 8], speech recognition [9, 10] and bioinformatics [11, 12]. Compared with such statistical methods as Hidden Markov Models, the traditional rule-based part-of-speech analysis usually has the following disadvantages:

(1) Ambiguity: How to choose an optional structure from a large number of ambiguous structures? Rule-based methods often cannot make a satisfactory result;

(2) Sketchy judgment: Rule-based methods cannot make a satisfactory guess for incorrect grammatical sentences;

(3) Rule conflicts: When the rules are increased, the conflicts between the rules become very serious. The rule debugging is very difficult. The latter rules often offset the effect of the previous rules, making the overall system performance hard to improve.

When the hidden Markov model of appositive classification method is employed on the English-Chinese bilingual corpus in this paper, several points should be noted:

(1) Consistent data structures (such as word maps) should be used to facilitate the convergence of the phases. This data structure should be able to deal with redundant expressions and indicate the results of a variety of segmentation tagging;

(2) Make sure that several possible results are presented at each stage. Some ambiguities cannot be ruled out at a certain stage and may be easily solved in the next stage. Providing multiple possible outcomes will help to reduce the overall error rate.

(3) Try to ensure that the probabilistic score obtained in the previous steps can be effectively used in the later stages. And try to establish a unified probabilistic model for each step to obtain the overall optimal results.

The steps based on the Hidden Markov model to identify appositive automatically are as follows:

When the Hidden Markov model is employed to identify appositive, the key problem is to mark each word in a word cluster, for example, to mark its part of speech. According to statistical rules, the distribution probability of each part of speech is only related to the part of speech of the previous word i.e. the binary grammar of part-of-speech. And the distribution probability of each word is only related to its part of speech. If we already have a corpus that has been marked with part-of-speech, then we can get the following two matrices by statistics. In fact, there is an initial matrix of part-of-speech distribution probability.

Part-of-speech transition probability matrix: $A=\{a_{ij}\}$, $a_{ij}=p(X_{t+1}=q_j|X_t=q_i)$

Part-of-speech to word output probability matrix: $B=\{b_{ik}\}$, $b_{ik}=p(O_t=V_k|X_t=q_i)$

Here $q_1,...,q_n$ means part-of speech collection, $V_1,...,V_n$是indicates the collection of word.

As for the problem of part-of-speech tagging, when we transfer one element a_{ij} in the transition probability matrix, we will get the following result. If the part-of-speech of the previous word is q_i, then the probability of part-of-speech of the next word is q_i. The meaning of one element b_i in the output probability matrix is the corresponding word V_k for its part-of-speech q_i.

With these two matrices, any given sequence of observations (word cluster) can quickly obtain the most probable sequence of state values (part-of-speech cluster) by a Viterbi algorithm [13]. The complexity of the algorithm is proportional to the length of the sequence of observations (the number of words in the sentence). The Viterbi algorithm are not described in detail.

Conclusion

Based on Deterministic Parsing, this study conducts the dynamic operation at the application level. It not only classifies and analyzes the appositive structures of the corresponding sentences in English and Chinese translation, but also proposes the feasible algorithm as an example for English-Chinese contrast linguistics, bilingual translation practice and computer programmers in the follow-up studies. Appositive structure types in the corresponding sentences of English-Chinese translation are finally confirmed and the information is extracted in this paper. The research on the implicated information of sentence structure should attract the attention of linguistic researchers and computer programmers. And it should become an important development direction of computational linguistics.

References

[1] T. Liang, D. Wong, C. Chao, P. Quaresma, F. Oliveira, S. Li, Y. Wang, Y. Lu. UM-Corpus: A Large English-Chinese Parallel Corpus for Statistical Machine Translation.]. Proc. 9th International Conference on Language Resources and Evaluation, Reykjavik, Iceland, European Language Resources Association, 2014, p.1837-1842

[2] J. Xu, X. Li. Structural and semantic non-correspondences between Chinese splittable compounds and their English translations. A Chinese-English parallel corpus based study. Corpus Linguistics and Linguistic Theory, 2014, Vol. 10, No.1, p.79-101

[3] X. Xu, Z. Xiao. Recent Changes in Relative Clauses in Spoken British English. English Studies, 2015, Vol. 96, No.7, p.818-838

[4] L. Lu. A contrastive study of the passive voice in journal articles in theoretical and applied linguistics. Chinese Journal of Applied Linguistics, 2013, Vol.36, No.4, p.465-478

[5] Y. Liu, T. Xiao. Translation of English-Chinese person name based on dictionary, bilingual corpus and web mining. Proc. 10th International Conference on Natural Computation, Xiamen, China, IEEE Press, 2014, p.818-822

[6] B. Leonard, P. Ted. Statistical Inference for Probabilistic Functions of Finite State Markov Chains. The Annals of Mathematical Statistics, 1966. Vol. 37, No.6, p.1554–1563

[7] Y. Wu, H. Lu, Z. Zhang. Text-Independent Online Writer Identification Using Hidden Markov Models. IEICE Transactions on Information & Systems, 2017, Vol. 100, p.332-339

[8] H. Quan, F. Ren. Weighted high-order hidden Markov models for compound emotions recognition in text. Information Science, 2016, Vol.329, p.581-596

[9] V. Renkens, H. Van. Weakly Supervised Learning of Hidden Markov Models for Spoken Language Acquisition. IEEE/ACM Trans. Audio, Speech & Language Processing, 2017, Vol. 25, No. 2, p.285-295

[10] C. Champion, S. Houghton. Application of continuous state Hidden Markov Models to a classical problem in speech recognition. Computer Speech & Language, 2016, Vol. 36, p.347-364

[11] Y. Hsu, W. Chen, S. Chen, H. Kao. Using hidden Markov models to predict DNA-binding proteins with sequence and structure information. Soft Computing, 2014, Vol. 18, No. 12, p.2365-2376

[12] P. C. Elisavet, N. Limnios. Viterbi algorithms for Hidden semi-Markov Models with application to DNA Analysis. RAIRO - Operations Research, 2015, Vol. 49, No.3, p.511-526

[13] A. Hayashi, K. Iwata, N. Suematsu. Marginalized Viterbi algorithm for hierarchical hidden Markov models. Pattern Recognition, 2013, Vol. 46, No. 12, p.3452-3459