Research on the Information Processing Model of Source Language of Parallel Corpus

Changchun Hong

1School of Foreign Languages, Huangshan University, 39 Xihai Road, Tunxi District, Anhui Province, China
hongcc@hsu.edu.cn

Abstract: The rapid development of corpus-based machine translation brings opportunities and challenges to the construction of corpus. With the continuous development of information technology, the information processing model of the source language has always been an academic hot topic in natural language processing. In the current machine translation framework, this paper proposes a parallel corpora source language processing model for translation equivalent trees. Firstly, the phrases in the source language and the target language sentence are tagged with parts of speech respectively; Secondly, the word alignment tool is used to align the phrase pairs of the source language and the target language; Syntax analysis is used to parse the source language sentence; Then, a translation equivalence tree is constructed based on the source language syntax tree, target language sentence and word alignment results. Finally, phrase translation pairs are extracted from the translation equivalent tree.

1. Introduction
Since the 21st century, the development of machine translation has entered a new era. The important sign of this new era is the introduction of corpus methods in rule-based technologies, including statistical methods, instance-based methods, methods for transforming corpora into language knowledge by means of corpus processing and so on. In recent years, corpus-based machine translation systems have developed rapidly and have achieved outstanding results. In 2006, Google developed and launched Google Translate. In 2011, Baidu launched Baidu Translate, which supports translation in 27 languages. These machine translation systems are increasingly used in people's daily work and life, and the quality of translations is getting higher and higher.

2. Natural Language Processing and Machine Translation
Natural language processing begins with machine translation. Machine translation is one of the core components of natural language processing. Historically, the development process of natural language processing has been almost the same as machine translation (Feng Zhiwei 2011) and these two are complementary. For example, the year of 1990 was a watershed in the process of "sequencing" natural language processing. It was preceded by a so-called "rationalist" methodology based on rules, and later it became a so-called "empirical" methodology based on statistics.

Machine translation is a new discipline that uses computers to translate one language (the source language) into another language (the target language). In 1954, the world's first machine translation experiment was completed by the “IBM-701” computer in the United States. Subsequently, machine translation studies have experienced high expectations (1954-1966), depression (1967-1976), recovery (1976-1989), and development (1989-). During the period, machine translation technology evolved
from the early (70s to the late 1980s) syntactic and rule-based machine translations to empiric corpus-based machine translation. Among them, the corpus-based methods can be subdivided into Example-based Machine Translation and Statistical Machine Translation. Statistical-based machine translation does not use translation examples directly, but abstracts the implicit translation knowledge in bilingual corpus into a statistical model through a training process. The statistical method is favored because of its good mathematical model, unsupervised learning ability, and good robustness. At present, most of the multilingual automatic translation systems on the Internet use statistical machine translation technologies such as Baidu, Google, Bing, and Youdao. However, these systems still need improvement in long-range ordering capabilities and translation logic. Especially in the translation of content or model translation that requires free translation, the accuracy of statistical machine translation results is not high. In response to these problems, some scholars proposed syntactic-based machine translation methods based on word-based and phrase-based machine translation techniques. This method aims to solve long-range ordering and increase the readability of the translation. It can avoid the difficult problems that traditional rule-based machine translation methods have to encounter in deep language analysis.

3. **Source language processing of parallel corpus**

1) Corpus Phraseology

At present, as an important part of natural language computer processing, corpus linguistics mainly deals with the collection, storage, retrieval, statistics, grammar annotation, syntactic semantic analysis and memory of machine-readable natural language texts. The corpus with the above-mentioned functions are widely used in language teaching, quantitative analysis, vocabulary research, collocation research, dictionary compilation, grammar research, language and cultural studies, legal language research, work style analysis, natural language understanding, and machine translation [3]. Words have always been the most universal unit of meaning in linguistic theories. However, with the continuous development of information technology, the boundaries of meaning-carrying units have been expanding. In the 1960s, the linguist Sinclair proposed that the most important unit bearing the meaning of the text is collocation, not words. After nearly 50 years of development, Phraseology has been established as a specialized subject area of linguistics and applied to such fields as language acquisition, teaching, and natural language processing [1]. Especially with the help of artificial intelligence technology (such as natural language processing, machine translation), the development of the phraseology has attracted the attention of cross-disciplinary disciplines such as linguistics, translation studies, and information science, and its importance has become increasingly prominent.

Due to the different emphasis, different scholars have different expressions for phrases in the corpus in addition to "chunk", "semi-fixed patterns", "speech formulate" and so on. For example, Lewis uses “lexical chunks”, Nattinger selects “Lexical phrases”, and Cowie chooses “ready –made complex units”, Pawley and Syndey use "lexicalized sentence stems". The research of phrase units in corpus linguistics generally adopts a corpus-driven research method, which uses concept definition→database construction→confirmation of working methods→extraction of multiple word structures→analysis from linguistic level (grammatical level, semantic level, pragmatic level)→functions clustering (conceptual function, program function, interpersonal function).

The first issue in machine translation is to determine the translation unit. Based on the linguistic knowledge of Chinese and English and the practice of Chinese-English translation, the phrase is currently the unit of translation for instance-based machine translation. Therefore, corpus phrasing is also the theoretical basis for constructing Chinese-English parallel corpora. Corpus phrasing is a corpus (driven) phrasing based on co-occurrence frequency of lexical items, supported by computer technology and quantitative analysis. Its main research objects are divided into the following two types: First, there are many fixed or semi-fixed forms. Word structures, such as strong tea, want to, etc. Second, grammatical phrases (non-grammatical relations) whose specific terms often appear in a specific grammatical structure. It is also worth noting here that there is a certain distance (also referred
to as n-gram) between discontinuous multi-word structures, ie. term components. In different linguistic schools, there is also a great difference in the level of phrasing. For example, the generative linguistic school pays little attention to the phraseology. Until the end of the 20th century, the phraseology was only taken seriously. According to Geries's point of view, the phraseology subverts the analysis-generation mechanism of the conversion-generation school using word-based units. It expands the unit into a modular organization, which is very similar to the core concepts related to the "Algorithmic Grammar and Lexicon" in natural language processing [2]. The development of such research has a greater impact on cognitive linguistics, especially construction grammar and corpus linguistics. With the maturity of natural language processing technology, corpus-driven research on phrase units has broken through the field of pure language research.

2) Source language processing of parallel corpus

In the process of the development of natural language processing, the construction of language resources plays an important role, because the language processing technology and its applications are inseparable from the support of corpus, knowledge base, and development platform. Corpus can be further divided into parallel corpus, translation corpus, and analog corpus, which refer to corpus consisting of translated texts in one language and its counterparts, corpus consisting of different translations of texts in the same source language and corpus composed of texts in different languages of similar time, domain, context, content, and scale. Parallel corpora can be unidirectional or bidirectional or multidirectional.

Phrase correspondence refers to any recognizable corresponding segment in the source text and the translation in a parallel corpus. It is not just a theoretical concept, it is an operational concept. According to certain working principles, if a text processor considers two bilingual segments in a parallel corpus to be corresponding in the translation process, it can be confirmed. This process certainly makes sense, but it is free. As long as the text operator is a person with knowledge of the relevant language, the choices he makes with his own language experience are acceptable. Professor Li Wenzhong of Beijing Foreign Studies University once suggested that the concept of “corresponding unit” is based on the existing linguistic theory, but it is not a linguistic level concept in itself, but is used at the operational level. It does not make the language processor at a loss, but at the same time meets the purpose of serving the machine-assisted translation.

In the current NLP domain (including machine translation), the commonly accepted approach to phrase alignment is the alignment of Chinese-English phrases based on translated equivalence trees. The main workflow is as follows: First, the phrases in the source language and the target language sentence are tagged with parts of speech respectively; Second, the word alignment tool is used to align the phrase pairs of the source language and the target language; Syntax analysis is used to parse the source language sentence; Then, a translation equivalence tree is constructed based on the source language syntax tree, target language sentence and word alignment results. Finally, phrase translation pairs are extracted from the translation equivalent tree. The entire process is shown in Figure 3 below.
The phrase-aligned machine translation system training is significantly higher in accuracy of translation than word-matching machine translation training, and the effect is better than “analysis-analysis-matching” strategy when extracting phrase translation pairs based on translation equivalent tree extraction methods. The reason is that in the "analysis-analysis-matching" strategy, the entire alignment process is constrained by the accuracy of the target sentence analysis and the inconsistency of the bilingual grammar body, and the quality of the obtained phrase pair is low; In the translation equivalence tree method, the alignment result is not constrained by the accuracy of the target sentence method and the inconsistency of the bilingual grammar system, and thus the effect is good, and the experiment proves that the accuracy can reach 80.97%. This is also the significance of the use of phraseology in the field of NLP practice. Corpus phrasing adopts a large number of principles and operations in the field of NLP to expand its theoretical boundaries and develop more doctrinal approaches. The ultimate goal of corpus construction is to serve theoretical research and practical applications. Large-scale, comprehensive, multi-purpose national-level parallel corpora can meet both needs. As far as practical application is concerned, the existing method is mainly based on a database in which the corpus runs in the background, and a network retrieval platform is established. Database size and processing depth determine the content and complexity of the search information.

4. Conclusion
The globalization of the world economy and the rapid spread of computer networks have brought unprecedented challenges to machine translation and natural language processing technologies, and at the same time, they have also created a great deal of space for their development. The complexity of machine translation is determined by the complexity of human natural language itself, which is unavoidable. Because there are too many factors (including rhythm, lexical, syntactic, semantic, etc.) that affect human natural language, these factors are intertwined with one another, which determines the non-determinism of NLP and greatly restricts the progress of machine translation research. The corpus-based machine translation system has a series of obvious advantages. It will achieve a substantial increase in machine translation quality based on the phrase-based machine translation system, combined with the latest artificial intelligence technology. However, more scholars have realized that if you want to go further, you need to conduct more in-depth research on the source language information processing model, especially the problems of personality and commonality of source language and target language.

Acknowledgments
This paper is financially supported by the following: Anhui Social Sciences Key Project Funding
References:
[1] Halliday, Mark Linguistics and machine translation [A]. In J. Webster (ed.)Computational and Quantitative Studies [C]. Beijing: Peking University Press, 2007.26
[2] Hausser, Roland. Foundations of computational linguistics : Man-machine communication in natural language [M].Berlin Heidelberg New York : Springer, 1999.34-35
[3] Granger, Paquot. Disentangling the phraseologica lweb [C] //Phraseology: An Interdisciplinary Perspective. Amsterdam/Philadelphia: John Benjamins, 2008: 27—49.
[4] Geries. Corpus Linguistics and theoretical linguistics: A Love—hate relationship? Not necessarily [J]. International Journal of Corpus Linguistics, 2010, 15 ( 3):327.
[5] Deerwester, Dumaisst, Furnas . Indexing by latent semantic analysis[J]. Journal of the American Society of Information Science, 1990, (41).
[6] Hofmann T. Probabilistic latent semantic indexing [C]. Proceedings of the 22nd annual international ACM SIGIRconference on Research and development in information retrieval, 1999.369-372
[7] Hirschberg, Julia Advances in natural language processing [J]. Science, 2015, 349 (6245)
[8] Li, Wenzhong. Extracting the multiword units from China's English news articles. Paper presented at the BAAL Corpus Linguistics SIG/OTA Workshop: Identifying and Researching Multiword Units, Computing services, Oxford University. 2005, 273-276
[9] Wray, Alison. Formulaic language: Pushing the boundaries [M]. Oxford: Oxford University Press, 2008: 87
[10] Anastasiou D, Gupta R. Comparison of Crowdsourcing Translation with Machine Translation [J] Journal of Information Science, 2011, 37( 6) :198-201