Study on English Translation Based on Probabilistic Syntax

Xueqin Du

School of Humanities, Jiangxi University of Traditional Chinese Medicine, Nanchang, 330008, China

Key words: Probabilistic syntax, English translation.

Abstract. As international exchange and cooperation deepens continuously in China, English translation has become an essential link in international language exchange in recent years. It plays a crucial role in various fields and activities of international social development. As international exchange continues to strengthen, English translation requirements also improve. Thus, seeking a scientific and efficient English translation method to improve translation level is especially important. English translation based on probabilistic syntax us a translation method which is proposed under such international background, is applied in translation practice and has gained a good effect. It helps students exclude the ambiguity of words or phrases fast, improve translation accuracy and enhance translation efficiency and quality. Based on brief introduction to the principle of probabilistic syntax, this paper analyzes the application of probabilistic syntax in English translation.

Introduction

Globalization is deemed as the most significant characteristic of the 21st century. To be specific, the contact among various countries becomes increasingly close and then international exchange becomes more and more frequent. Beyond all doubt, language is the well-deserved medium and platform for international communication. English as a worldwide language is most frequently used currently. The demand of international community for English talents of non-English speaking countries becomes bigger and bigger. As globalization goes deep continuously, higher and higher requirements are proposed for English talents. The famous educator Li Lugui mentioned in Comprehensive English Grammar for High Schools that, “English and Chinese belong to two different language families. They are different in essence. These differences bring about numerous challenges and troubles for Chinese students to learn English.” Although the differences between the two language families bring about pressure for language learners, they also bring driving force for them. Linguistics experts concentrate on studies all the time in order to seek a method to help students eliminate English study troubles and obstacles.

In the seeking process, the opposition between rationalism through and economism appears for rule method and statistical approach during handling natural language. Some scholars always hold very extreme opinio0ns. For example, as early as 1956, the famous American linguistician Noam Chomsky said, “We must realize the concept of sentence probability. In any known explanation of this term, it is a totally useless concept (translation).” Chomsky entirely ignored the concept of sentence, and scoffed at statistic approach. Fred Jelinek - the chief responsible person of Voice Research Group at Watson Research Center of IBM Corporation rodomontaded and said in 1988 that, “If linguisticians leave our research group, the voice recognition rate will improve (translation).” The opinions of the above two famous scholars are not fair. But happily, many scholars combined rule method and statistical approach. At present, very significant results have been achieved. Their researches mainly involve two aspects: Probabilistic Context Free Grammar (PCFG), and lexicalized PCFG. The detailed introduction is as follows:

Introduction to probabilistic syntax

Probabilistic model

Classic probabilistic model - PCFG

Definition

Probabilistic syntactic analysis is a syntactic analysis method based on context-free grammar of probabilistic rule set. The rule set mainly aims at words and phrases. But, the facts reflect that the property of a word is same, but the vocabulary is different. The common syntactic rules are different generally. Hence, ambiguity problem may be caused easily if such rule is applied to analyze sentences. As a result, countless parse trees will be worked out for a sentence. To describe the rule in amore detailed and accurate manner, multiple grammar rule systems which derive from context-free grammar are brought in this field. PCFG is a grammar rule system which is formed through introducing probability in context-free grammar rule system. It makes probability provide basis for ambiguity resolution in the analysis process.

The symbol system of PCFG includes the following elements:

1) A set of terminal symbols, $\{w^k\}$,k=1,…,V

2) A set of non-terminal symbols,$\{N^1\}$， i=1,…,n

3) A begin symbol, N^1

4) A set of rules, $\{N^1 \rightarrow \zeta^j\}$, (wherein, ζ^j is a sequence of terminal symbols and non-terminal symbols)

A PCFG is a quintuple (N,∑,S,R,P)

1) A non-terminal symbol set N

2) A terminal symbol set ∑

3) A begin non-terminal symbol S∈N

4) A production set R

5) For any production r∈R, the probability is P(r)

In terms of probability, PCFG gives the probability distribution of various possible symbol sequences which derive from a non-terminal node, i.e.

$$\forall i \quad \sum_{j} P(N^i \rightarrow \zeta^j) = 1$$

To figure out the probability $P_{(t)}$ of a parse tree, necessary independence hypothesis is required. Classic PCFG holds that the probability of applying each rule is independent from the context and ancestor node. In other words, if a non-terminal symbol node is given, the probability of derivative child node sequence is decided by grammar. In this way, we suppose all rules in t form a multiset R of rules. Then,

$$P_{YtY} = \prod_{(r \in R)} P(r|LHS(r))$$

Assumed conditions

PCFG includes the following three assumed conditions:

1) Location invariance: the probability of subtree is independent of the location of words governed by the subtree in the sentence;

2) Context-free property: the probability of subtree is independent of the words beyond the control range of subtree;

3) Ancestor irrelevance: the probability of subtree is independent of ancestor node of subtree.

Basic problems

PCFG includes the following three basic problems:

1) Give a sentence, predict and estimate the probability of sentence production;

2) Under the condition where syntactic structure of a sentence is ambiguous, how to choose the optimal syntactic analysis fast and accurately?

3) How to train grammatical parameters from the corpus?

Other probabilistic models

Because classic PCFG is actually established on the basis of some very ideal independence hypotheses, and these hypotheses do not quite conform to the actual conditions, actual effect of PCFG differs a lot with the expected effect. Some breakthrough probabilistic models of context-free hypothesis of PCFG are as follows:

P-PCFG - probabilistic model which takes into account of ancestor node

In P-PCFG, the symbol system remains unchanged, and the independence hypothesis is changed. It holds that the application probability of each rule is decisively influenced by the father node of left non-terminal symbol node. Accordingly, the model includes such probability distribution:

$\forall i,k \sum_j P(N^i \to \zeta^j | < N^i, N^k >) = 1$, wherein， $< N^i, N^k >$ means N^i is the child of N^k.

Probability calculation formula of parse tree:

$$P_{(t)} = \prod_{r \epsilon R} [\![P(r| < LHS(r), ParentOf(LHS(r))]\!] >)$$

The phrase structure is restricted by the supper-layer node. For example, the internal structures of NP phrase as the subject (NP is below S) and NP phrase as the object (NP is below VP) have obviously different probability distribution. P-PCFG adds this decisive factor into the probabilistic model.

PORD-PCFG - probabilistic model which takes into account of ancestor node and node position

PORD-PCFG considers that the application probability of each rule is decided by the father node of left non-terminal symbol node and the ranking of left non-terminal symbol node in the brother node. Accordingly, the model includes such probability distribution:

$\forall^{I,k,ord} \quad \sum_j P(N^i \to \zeta^j | < N^i, N^k, ord >) = 1$, wherein, $<N^i, N^k, ord>$ means N^i is the ord^{th} child of N^k

Probability calculation formula of parse tree:

$$\mathbf{P(t)} = \prod P(r| < LHS(r), ParentOf(LHS(r)), OrderOf(LHS(r)) >)$$

If two or more same non-terminal symbol nodes derive from a non-terminal symbol node and the sole difference between them is the location, location information is very crucial. For example, if a verb phrase VP has two objects NP−1 and NP−2, the structure of the two objects will differ significantly. PORD-PCFG takes into account of the influence of father node and ranking position on the object structure.

PRORD-PCFG - probabilistic model which takes into account of rules at the superior level

PRORD-PCFG holds that the application probability of each rule is decided by the implication rules of left non-terminal symbol node and its location in the implication rules. Accordingly, the model includes such probability distribution:

$\forall^{i,r,ord} \sum_j P(N^i \to \zeta^j | < N^i, r, ord >) = 1$,wherein , $<N^j, r, ord>$ belongs to the relation between the ord^{th} child at the right of rule r and N^j.

Probability calculation formula of parse tree:

$$\mathbf{PYtY} = \prod P(r| < LHS(r), ParentRuleOf(LHS(r)), OederOf(LHS(r)) >)$$

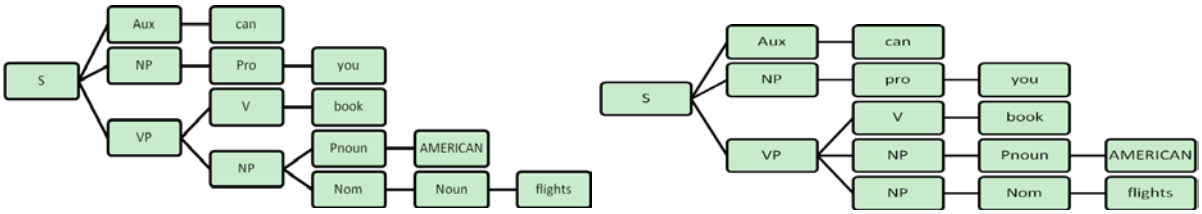
PRORD-PCFG totally considers the intervention of rules at the superior level in the derivation chain on the rules at this level. The rules at the superior level give a relatively closed local context. The functions of father node, brother and ranking location are included.

Study on English translation based on probabilistic syntax

PCFG

For example, “Can you book AMERICAN flights?” This sentence is ambiguous. One of the meanings is that, “Can you book flights run by AMERICAN?” The other meaning is that, “Can you book flights on behalf of AMERICAN?” Their trees are as follows:

Left tree: Right tree:



The probability of left tree is:
S→Aux NP VP [.15]
NP→Pro [.40]
VP→V NP [.40]
NP→Nom [.05]
Non→Pnoun Nom [.05]
Nom→Noum [.75]
Aux→can [.40]
NP→Pro [.40]
Pro→you [.40]
verb→book [.30]
Pnoun→AMERICAN [.40]
Noun→flights [.50]
The probability of right tree is:
S→Aux NP VP [.15]
NP→Pro [.40]
VP→V NP NP [.05]
NP→Nom [.05]
NP→Pnoun [.35]
Nom→Noun [.75]
Aux→can [.40]
NP→Pro [.40]
Pro→you [.40]
Verb→book [.30]
Pnoun→AMERICAN [.40]
Noun→flights [.50]
The probability $P(T_1)$ of left tree is:

$$P(T_1)=.15*.40*.40*.05*.05*.75*.40*.40*.40*.30*.40*.50=1.7X10^{-6}$$

The probability $P(T_1)$ of right tree is:

$$P(T_1)=.15*.40*.05*.05*.35*.75*.40*.40*.40*.30*.40*.50=1.5X10^{-6}$$

It thus can be seen that, the probability of left tree is greater than that of right tree. So, the left tree can be judged to be the final correct result.

Ambiguity resolution algorithm chooses the best tree (we call it T) from the parse trees (we call it t(S)) of sentence S as the correct analysis result. Formally, if $T \in t(S)$, the tree T(S) with the largest probability will be equal to argmax P(T). Thus, T(S)=argmax P(T)

Lexicalized PCFG

PCFG has the problem of structure dependence and vocabulary dependence.

Problem of structure dependence

CFG supposes that, when non-terminal symbol at the left side of rules is rewritten, it is independent of other non-terminal symbols. So, each rule is independent in PCFG. In this way, the probability of rules can multiply. However in English, transcription of rules on the node is related to the location of the node in the tree diagram. For instance, the subject of English sentence tends to apply the pronoun, because the subject generally represents the theme or old information. During citing the old information, the pronoun is generally used. Other nouns are generally used to introduce new information.

Problem of vocabulary dependence

PP adhesion

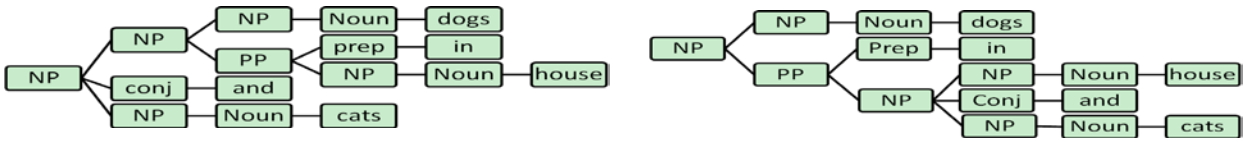
In English sentences, the prepositional phrase PP can serve as the adverbial modifier of central verb phrase VP or the modifier of noun phrase, which is related to the vocabulary. For instance, “Washington sent more than 10,000 soldiers into Afghanistan”, PP “into Afghanistan” is attached to NP “more than 10,000 soldiers”, or attached to the verb “sent”.

In PCFG, such adhesion judgment should be chosen from the following rules: NP→NP PP(NP adhesion) and VP→VP PP(VP adhesion). In the two rules, NP adhesion is in the preponderant position. But in the above sentence, PP “into Afghanistan” should be attached to the verb “sent”, because the verb “sent” requires a PP to express the direction, while the PP “into Afghanistan” just meets this requirement. Obviously, PCFG cannot handle such problem of vocabulary dependence.

Ambiguity of parallel structure

The sentence “dogs in houses and cars” has structural ambiguity.

Left tree: Right tree:



Although we instinctively consider the left tree is correct, their probability is same because both trees use the identical rules. In such case, PCFG will assign the same probability to the two trees. In other words, PCFG cannot judge the ambiguity of this sentence. Thus, vocabulary information may be introduced in PCFG to solve these problems.

PCFG and lexicalized PCFG carry out in-depth exploration of combining rule method and statistical approach and have gained significant results. They have achieved very strong ambiguity resolution function in English translation, greatly reduce translation error rate and improve English translation quality and efficiency. Translation experts have applied them in machine translation system and also gained great achievements.

Conclusion

In conclusion, the application of probabilistic syntax in English translation has gained significant achievements. However, PCFG hypothesis including structure context-free and vocabulary context-free must be deeply studied. At present, we have gained some results in structure context correlation. Seeing from international mainstream syntactic analysis method, lexicalization of syntactic analyzer and penetration of vocabulary information into each analysis level can improve the

property in essence. In one word, we must fully mine and utilize existing resources and achieve their maximum value in order to improve our English translation level furthest.

References

[1] Zeng Longhua, Zhou Kun, Comparison of English and Chinese Syntactic Structure and Their Translation. *Journal of Hubei Radio& Television University*, 2013 (6): 95-96.

[2] He Hong, Differences of English and Chinese Syntax and Translation Flexibility. *Journal of Zhengzhou Institute of Aeronautical Industry Management (Social Sciences)*. 2010 (6): 97-100.

[3] Wu Xiaojun, Comparison of English and Chinese Syntax and Structural Transfer in English-Chinese Translation. *Journal of Hunan University of Science and Engineering*. 2006 (8): 179-182.

[4] Xu Hui, Analysis on English Translation Methods and Skills. *Campus English*. 2014 (27): 186-187.

[5] Li Na, English Translation Methodology. *Cultural Geography*. 2014 (10X): 245-246.