Error Analysis of English-Chinese Machine Translation

Fei Fang[1], Shili Ge[1, 2,*], and Rou Song[2]

[1]School of English for International Business, Guangdong University of Foreign Studies
510420 Guangzhou, China
[2]Guangdong Collaborative Innovation Center for Language Research and Service
510420 Guangzhou, China

fangfei562@126.com,geshili@gdufs.edu.cn,songrou@126.com

Abstract. In order to explore a practical way of improving machine translation (MT) quality, the error types and distribution of MT results have to be analyzed first. This paper analyzed English-Chinese MT errors from the perspective of naming-telling clause (NT clause, hereafter). Two types of text were input to get the MT output: one was to input the whole original English sentences into an MT engine; the other was to parse English sentences into English NT clauses, and then input these clauses into the MT engine in order. The errors of MT output are categorized into three classes: incorrect lexical choices, structural errors and component omissions. Structural errors are further divided into SV-structure errors and non-SV-structure errors. The analyzed data shows firstly, the major errors are structural errors, in which non-SV-structural errors account for a larger proportion; secondly, translation errors decrease significantly after English sentences are parsed into NT clauses. This result reveals that non-SV clauses are the main source of MT errors, and suggests that English long sentences should be parsed into NT clauses before they are translated.

Keywords: Machine Translation; Error Analysis; NT clauses; SV clauses; Non-SV clauses.

1 Categorization of Errors in Machine Translation

The past 60 years has witnessed great progress in machine translation. Nowadays, online automatic translation systems play a vital role in rough reading of foreign-language texts, which has become the necessary function of word processing systems and information retrieval systems. However, as for the need of intensive reading, there still exist many problems in automatic machine translation, especially in the translation of long sentences.

In the field of machine translation, there are two widely-used evaluation methods: manual evaluation and objective evaluation (also known as automatic evaluation). According to Koehn [1], the former is a method that evaluates outputs of machine translation systems by subjective judgments; the latter evaluates MT outputs automat-

∗Corresponding author.
adfa, p. 1, 2011.

ically according to a certain mathematical model. Neither of them can show the source of MT errors, not to mention the improvement of long sentence translation.

According to Zhao et al [2], translation errors can be divided into 7 types: 1) incorrect words, 2) missing content words, 3) wrong word order, 4) translation with meaning contrary to the original, 5) errors of named entity, 6) errors of numerals and quantifiers/temporal words and 7) other errors. That paper comes to the conclusion that the first 3 types, especially incorrect words and wrong word order, accounts for the largest proportion. This analysis is correct, but it does not explain the objective phenomenon that MT has a poorer performance in long-sentence translation than in short-sentence translation. Still it fails to reveal the causes of the errors. Thus the analysis provides no direct benefit for the quality improvement of long sentence machine translation.

The paper explores the results of English-Chinese machine translation, categorizing errors in machine translation into 3 types: incorrect lexical choices, structural errors and component omissions. Examples are employed to illustrate the significant influence of the complexity of naming-telling clause (NT clause, hereafter) in English sentences on the performance of machine translation and to suggest the possibility that parsing long English sentences into NT clauses will reduce errors. The paper is expected to offer some inspirations for breaking through the bottleneck of automatic long sentence translation.

As for the categories in machine translation errors, our ideas are as follows:

The aim of translation is to keep the semantic consistency between the source text and the target text. As semantics has its own constructions, so we can examine the correctness of forms of semantic structures, including errors like additions, omissions, mistaken usages, wrong types of semantic structures and wrong word choices in leaf nodes of semantic structures. Giving that additions rarely occur, we can categorize translation errors into 3 types: omissions of semantic structures (component omissions), errors of semantic structures in form or type (structural errors) and errors of word choice (incorrect lexical choices).

The semantic structure mentioned in this paper is in view of a higher and more abstract level, which includes the referential component plus its statement (subject-predicate), the modifier plus its modified (attribute-head or the adverbial-head), the action plus its object (verb-object), the action plus its supplementary instructions (verb-complement), preposition plus its object (preposition-object), conjunctive and its logical arguments and so on. Particularly, if the antecedent and consequent of a certain structure are transposed, it will be classified into errors of structural type, naming the structural error, equivalent to the wrong word order proposed by Zhao et al [2]; but the structural error we refer to includes not only the wrong word order but also other errors. For example, some construction should be translated into attribute-head construction but the MT output is a structure of adverbial-head, or there is confusion of logical arguments on whether they are the referential or the statement on the two sides of coordinate conjunctions, etc. In this paper, component omissions are generally the type of missing content words proposed by Zhao et al [2]. Incorrect lexical choices are mostly incorrect words proposed by Zhao et al [2]. Yet, if prepositions, conjunctions, or any verbs are omitted or mistranslated, causing structural er-

rors, they are classified into errors of structural types. As for the rest 4 types proposed by Zhao et al [2], they can also be classified into the 3 types mentioned above.

To illustrate semantic structure more clearly, we offer the following example of semantic structure in the form of a semantic tree.

Original English text: (It) announced new advertising rates for 1990 and said

The MT output: 宣布新的广告费率1990和

The revised translation: 宣布1990年新的广告费率并说

The comparison of semantic structures:

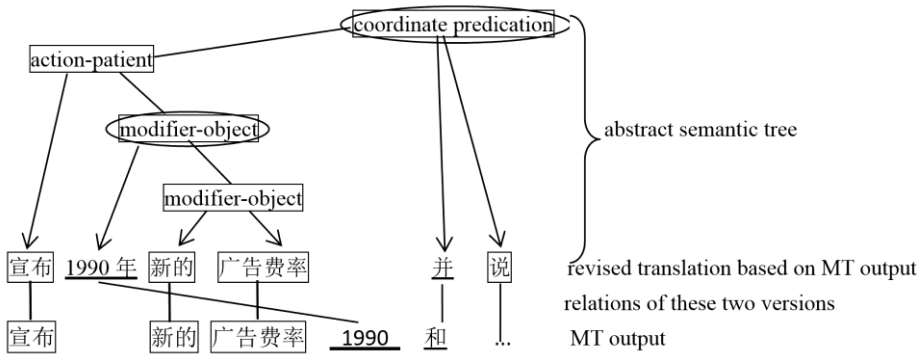

Fig. 1. Error analysis of machine translation from the perspective of semantic structure

In Fig. 1, the last line is Chinese MT output; the line above it is the manually revised translation based on the MT output. From Fig. 1, we can see that "1990" should be translated into "1990 年". So, it is an incorrect lexical choice which is annotated by underlined words. Besides, "1990年" is the modifier of "新的广告费率", but "1990" is placed after "新的广告费率" in the MT output, unable to show its modifier-object structure. It is also a structural error marked with ellipses on nodes of the semantic tree. "and", here showing the coordination of two predicate components, should be translated into "并", not "和" which shows the coordination of referential components. It is not only an incorrect lexical choice but also a structural error, therefore marked with an underline and an ellipse respectively. "said" should be translated into "说", which does not occur in the MT output. This kind of error, i.e., component omission, is marked with dots.

2 NT Clause, SV clause and non-SV clause

Song and Ge [3] define an NT clause as the structure consisting of a naming and a telling. A naming is a referential component and a telling is the description or post-modification component of the naming. In English language there are 8 specific relationships between a naming and its telling: subject and predicate, the referential component and its seven types of telling, including relative clauses, past participial

phrases, present participial phrase, infinitive phrase, adjective phrases, declarative prepositional phrases and the explanatory noun phrases.

Example 1: *Documents filed with the Securities and Exchange Commission on the pending spinoff disclosed that Cray Research Inc. will withdraw the almost $ 100 million in financing it is providing the new firm if Mr. Cray leaves or if the product-design project he heads is scrapped.*

For the sake of visual cognition, we represent the relation between naming and telling in an English sentence with specific method called newline-indented schema by Song [4]. The specific method is: when a naming and its telling are not subject-predicate relation, or they are non-adjacent subject-predicate relation, we place the telling part in a new line and indent it after its naming part.

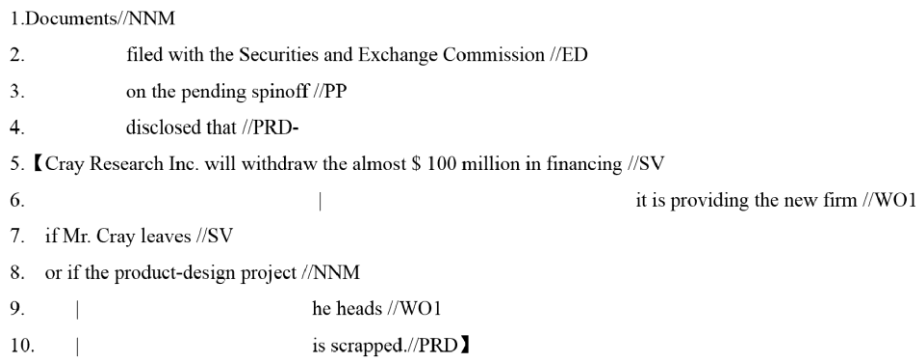

Fig. 2. Newline-indented schema of example 1

In Fig. 2, the first line and the eighth line except the conjunction “or” and “if” are referential components, i.e. naming, annotated with NNM; the fifth and seventh lines are subject-predicate NT clauses, annotated with SV; in the remaining six lines, the second line is past participial phrases, typed ED; the third line is a prepositional phrase, typed PP; in the fourth line, the finite verb acts as predicate of the first line, but lacking the object, so typed PRD-; “Documents” in the first line acts as the naming of these three tellings, so they are indented to its right end in the new lines. The sixth line is a relative clause, and its antecedent is “the almost $100 million in financing” in the fifth line and is the direct object of the relative clause, which is suggested by the WO1 in the sixth line. The sixth line is indented to the right end of its antecedent and a vertical bar is used to mark the left end of the antecedent, signaling the beginning of the naming. The types of the ninth line and the tenth line can be inferred from the above explanation. A pair of black square brackets is employed to mark the object clause ranging from the fifth line to the tenth line, which belongs to the fourth line.

In Fig. 2, each naming is placed on the upper-left side of its telling, thus sequences of NT clauses of this English sentence can be constructed mechanically as follows. These NT clauses are numbered according to the line number of their tellings.

2. Documents+filed with the Securities and Exchange Commission //ED

3. Documents+on the pending spinoff //PP
4. Documents+disclosed that //PRD-
5. Cray Research Inc. will withdraw the almost $ 100 million in financing //SV
6. the almost $ 100 million in financing +it is providing the new firm //WO1
7. if Mr. Cray leaves //SV
9. the product-design project+ he heads //WO
10. or if the product-design project+ is scrapped.//PRD

NT clauses can be used for the basic structure of categorizing translation errors. NT clause is classified into two types. One is adjacent subject-predicate structures, called SV clauses, while the other one is non-adjacent subject-predicate structures or non-subject-predicate structures, called non-SV clauses. Therefore, the structural error can be classified into SV-structure error and non-SV-structure error. The reason that we distinguish these two types is the different difficulty levels they occur in machine translation. The former structure (subject-predicate NT clauses) accords with syntax of clauses and the subject and predicate are adjacent, which can be handled effectively by either rule method or statistical method; however, the latter does not accord with the syntax of normal English clauses, and non-adjacent subject-predicate structures or non-subject-predicate structures as they are, the above two traditional methods may be confronted with great difficulty.

By using the newline-indented schema, we have annotated NT-clause structure for several thousands of English sentences from the Wall Street Journal in Penn Treebank and have a more detailed test on 253 English sentences among them. Two ways are adopted for machine translation in the paper. One is that we input these whole English sentences into *Baidu Translate*, a popular machine translation engine in China, and amend the MT outputs manually. Then, through comparing the MT outputs with the manually amended texts, we classify and tag the errors in MT outputs. The other way is very similar to the former one, only with a difference that the whole English sentences are parsed into English NT clauses in advance, and these English NT clauses are then translated separately. The errors are tagged according to three types we mentioned above, and the structural errors among them are divided into SV-structure errors and non-SV-structure errors. Examples will be given in section 3 to illustrate the specific procedure of these two ways. In section 4, we categorize and compare errors occurring in MT outputs which are obtained through these two different ways. Section 5 will list conclusions drawn from the comparison.

3 Analysis of Errors in English-Chinese Machine Translation

Example 2: *Newsweek, trying to keep pace with rival Time magazine, announced new advertising rates for 1990 and said it will introduce a new incentive plan for advertisers.*

The alignment of the translated Chinese word sequences and the original English text are as follows:

Fig. 3. Alignment of English and Chinese in words of example 2

MT output and revised translation are as follows:

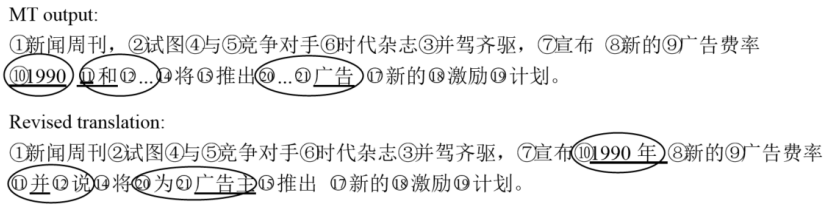

Fig. 4. MT output and revised translation of example 2

In Fig. 4, in order to save printing space, words where structural errors occur are directly enclosed with ellipses, which is different with the practice in Fig. 1 of the semantic tree that ellipses are placed on nodes of the semantic tree; incorrect lexical choices are still underlined under the words; component omissions are marked with dots on corresponding MT output texts.

There are errors in 3 phrases of the MT output:

(1) “1900” should be translated into “1990 年(year)”. Numerals can be functioned as dates or years in English without any addition, but in Chinese, these characters “日(date), 月(month) and 年(year)” should be added after the numerals. So it belongs to incorrect lexical choices. Besides, “1990年(for 1990)” is the modifier of “新的广告费率(new advertising rates)”, but “1990” is placed after “新的广告费率” in the MT output, which is ungrammatical Chinese structure.

(2) “and said” should be translated into “并说”, but the machine translates it into “和 (and)”. Even though “and” has two corresponding meanings of “和” and “并” in Chinese, “和” is used for the coordination of the referential components while “并” signifies for the coordination of statements. Here the context contains two statements, thus “and” should be translated into “并”. This point belongs to incorrect lexical choices. Besides, it also belongs to errors of structural type. Furthermore, the meaning of “said” is omitted in the Chinese text, so this belongs to component omission.

(3) “for advertisers” should be translated into “为广告商”, but the machine translates it into “广告”, which signals the omission of beneficiary argument “为(for)”. “广告商(advertisers)” is translated into “广告(advertisement)”, which belongs to incorrect lexical choices. Besides “为广告商(for advertiser)” is an adverbial and should be placed before its modified predicate “推出新的激励计划 (announced a

new incentive plan)". The post-position of the adverbial "广告" makes the translated text hard to understand. So this also belongs to structural errors.

To sum up, we can make a calculation that there are 3 structural errors, 3 incorrect lexical choices and 2 component omissions.

In order to analyze the causes for these errors, we parse the original sentence into NT clauses by showing with a newline-indented schema as follows, and original words of errors in the MT output are underlined, where the different line types will be explained soon:

Newsweek ,//NNM
    trying to keep pace with rival Time magazine ,//ING
    announced new advertising rates for 1990 //PRD
    and said//PRD-
        it will introduce a new incentive plan for advertisers .//SV

Fig. 5. Newline-indented schema of example 2

In Fig. 5, the third line and the fourth line where errors occur are both predicates of the first line. These two lines are also tellings of the first line. The difficulty of translation increases because of the two pairs of subjects and predicates being non-adjacent. Besides, the conjunction "and" before the predicate of the fourth line also increase the difficulty of machine translation. The fifth line is a NT clause of subject-predicate type and is not long, but is placed at the end of the whole sentence without any punctuation before it, so translation errors before it would be propagated into it. Therefore, there is no surprise that errors occur in this simple clause.

The above analysis of reasons for translation errors enlightens us to think that if we parse long English sentences into NT-clause sequences, errors in machine translation may decline. In order to verify the assumption, we parse this long English sentence into 4 NT clauses and make simple mechanical changes to let every NT clauses be grammatically correct clauses with a subject-predicate structure. Then we input these NT clauses into machine translation system, and the results are as follows:

(1)

The NT clauses: ①Newsweek②(trying|tries) to③keep pace④with⑤rival⑥Time magazine

Note: "(trying|tries)" means shifting the present participle *trying* into finite verb *tries*.

The MT output: ①新闻周刊杂志②试图④与⑤对手⑥... ③并驾齐驱

The revised translation: ①新闻周刊 ②企图④与⑤对手⑥时代杂志③并驾齐驱

Analysis of errors: "⑥时代杂志" is omitted in the sentence.

(2)

The NT clauses: ①Newsweek⑦announced⑧new⑨advertising rates⑩for 1990

The MT output: ①新闻周刊 ⑦宣布 ⑧新的 ⑨广告费率 ⑩为1990

The revised translation: ①新闻周刊 ⑦宣布 ⑩1990年⑧新的 ⑨广告费率

Analysis of errors: there still occurs an incorrect lexical choice and a structural error "for 1990".

(3)

The NT clauses: ①Newsweek (and) ⑫said

Note: (and) means temporarily deleting the conjunction *and* between the naming and the telling before translation.

The MT output: ①新闻周刊⑫说

No error.

(4)

The NT clauses: ⑬it⑭will⑮introduce⑯a⑰new⑱incentive⑲plan⑳for㉑advertisers

The MT output: ⑬它⑭将⑳为㉑广告主⑮引入⑯一个⑰新的⑱激励⑲计划

No error.

Total errors: 1 structural error, 1 incorrect lexical choice and 1 component omissions.

Compared with errors in the MT output of the original whole sentence, the errors in MT output of NT clauses are with 2 structural errors, 2 incorrect lexical choices and 1 component omission less.

In order to show the comparison between the two results, in the above newline-indented schema, namely, in Fig. 4, we underline words and phrases where errors occur in MT output of the original whole sentence with bold underlines while wave lines are used to mark words and phrases where errors occur in MT output of NT clauses, and bold wave lines to mark words and phrases where errors occur in MT outputs of both ways.

Example 3: *About 160 workers at a factory that made paper for the Kent filters were exposed to asbestos in the 1950s .*

The alignment of the translated Chinese word sequences and the original English text as follows:

①About ②160 ③workers ④at ⑤a ⑥factory ⑦that ⑧made paper ⑨for ⑩the Kent
①大约 ②160 名 ③工人 ④在 ⑤一家 ⑥工厂里 ⑦ ⑧造纸 ⑨为 ⑩肯特

⑪filters ⑫were exposed to ⑬asbestos ⑭ in ⑮the 1950s.
⑪过滤嘴 ⑫接触过 ⑬石棉 ⑭在 ⑮20 世纪 50 年代。

Fig. 6. Alignment of English and Chinese in words of example 3

MT output and revised translation are as follows:

MT output:
①大约②160 名③工人④在⑤一家⑥工厂⑧制造的造纸厂⑯的⑨..⑩肯特⑪过滤器⑭在⑮20 世纪 50 年代⑫暴露于⑬石棉。

Revised translation:
①大约②160 名③工人④在⑨为⑩肯特⑪过滤器⑧造纸⑯的⑤一家⑥工厂⑭在⑮20 世纪 50 年代⑫暴露于⑬石棉。

Fig. 7. MT output and revised translation of example 3

Errors in MT output are as follows:

All errors occur in the relative clause “that made paper for the Kent filter”. The verb-object structure “made paper” is translated into a modifier-head structure “制造的造纸厂的” in Chinese. So the above error belongs to structural errors. “for the Kent filter” is translated into “肯特过滤器”. First of all, the preposition “for” is omitted; then, the prepositional phrase “为肯特过滤器(for the Kent filter)” functioning as adverbial should be placed before the verb phrase “造纸(made paper)”. This error belongs to structural errors.

To sum up, there are 2 structural errors and 1 component omission in MT output.

The newline-indented schema of example 3 are as follows:

About 160 workers at a factory//NNM
| that made paper for the Kent filters//WS
were exposed to asbestos in the 1950s.//PRD

Fig. 8. Newline-indented schema of example 3

In Fig. 8, errors occur in the non-subject-predicate NT clause in the second line, because the relative clause of “a factory” is inserted into the middle of the subject and the predicate. Position adjustment is needed because there is a prepositional phrase functioning as adverbial in the relative clause, which increase the difficulty of machine translation. We parse the whole sentence into 2 NT clauses, make simple mechanical changes that make them into normal subject-predicate clause and input them into the machine translation system separately. Results are as follows:

(1)

The NT clause:

①About②160③workers④at⑤a⑥factory⑫were exposed to⑬asbestos⑭in⑮the 1950s

The MT output:

⑭在⑮20世纪50年代⑤一家⑥工厂①大约②160名③工人⑫暴露在⑬石棉中

(2)

The NT clause: ⑤the ⑥factory (that) ⑧made paper⑨for⑩the Kent ⑪filters

Note: (that) means temporally deleting the relative pronoun *that* between the naming and the telling.

The MT output: ⑤这家⑥工厂⑨为⑩肯特⑪过滤器⑧制造了纸

Both of these two clauses have no errors.

Compared with the MT output of the original sentence, according to Fig. 7, we can see there is a reduction of 2 structural errors and 1 component omission.

Example 4: *The survival of spinoff Cray Computer Corp. as a fledgling in the supercomputer business appears to depend heavily on the creativity – and longevity – of its chairman and chief designer.*

The alignment of the translated Chinese word sequences and the original English text is as follows:

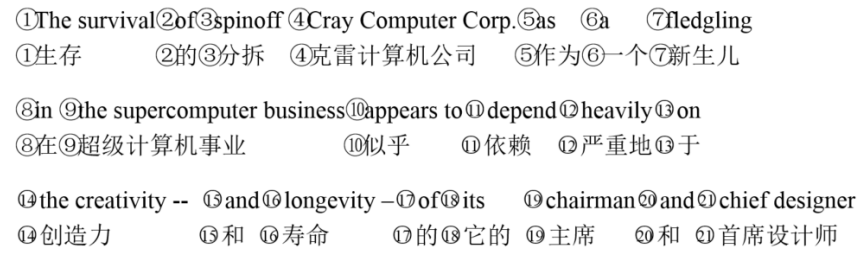

Fig. 9. Alignment of English and Chinese in words of example 4

MT output and revised translation are as follows:

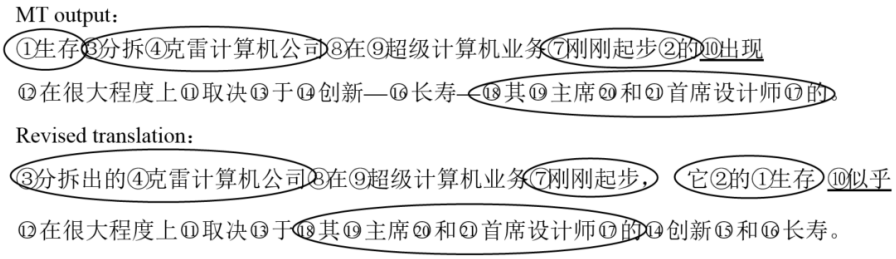

Fig. 10. MT output and revised translation of example 4

Errors in MT output are as follows:

(1) “spinoff Cray Computer Corp.” should be translated into a modifier-head structure “分拆出的克雷计算机公司” in Chinese, but here it is translated into a verb-object structure “分拆克雷计算机公司”, which belongs to structural errors.

(2) it is correct to translate “ the survival” into “生存”, but as the head noun, in Chinese it should be placed after its attribute “分拆出的克雷计算机公司的”. It is placed before its attribute in the MT output, so it is a structural error.

(3) “as a fledgling in the supercomputer business” is the post-modifier of “spinoff Cray Computer Corp.”. It is acceptable to translate “as a fledgling in the supercomputer business” into an adverbial-verb structure “在超级计算机事业中刚刚起步” as a statement of “spinoff Cray Computer Corp.”, but in MT output, the addition of “的” in Chinese makes the adverbial-verb structure shift into a modifier-head structure. It is a structural error.

(4) “of its chairman and chief designer” is the attribute of the head noun “the creativity -- and longevity –”, so when translated into Chinese, it should be placed before its head noun. It is a structural error.

(5) “appear” has two meanings: “出现” and “似乎” in Chinese, but according to the context, it should be translated into “似乎”, while “出现” in MT output, so it is an incorrect lexical choice.

From the above analysis, there are 4 structural errors and 1 incorrect lexical choice in the MT output.

The newline-indented schema of example 4 are as follows:

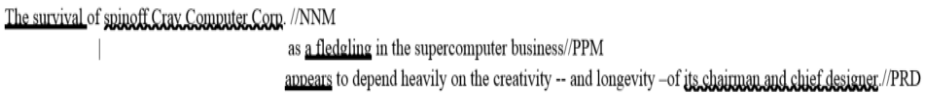

Fig. 11. Newline-indented schema of example 4

In Fig. 11, errors occur in the naming of the first line and the telling of the second and the third lines. The results are as follow after the sentence is parsed into NT clauses and inputted into machine translation system.

(1)

The NT clause: ③spinoff④Cray Computer Corp. [is]⑤as⑥a⑦fledgling⑧in⑨the supercomputer business

Note: [is] is added to make the sentence accord with grammatical rules.

The MT output: ③分拆 ④克雷计算机公司⑧在⑨超级计算机业务⑦刚刚起步

The revised translation:③分拆出的④克雷计算机公司⑧在⑨超级计算机业务中⑦刚刚起步

Analysis of errors: there is still 1 structural error.

(2)

The NT clause:

①The survival②of③spinoff④Cray Computer Corp⑩appears to⑪depend⑫heavily⑬on⑭the creativity --⑮and⑯longevity –⑰of⑱its⑲chairman⑳and㉑chief designer

The MT output:

③分拆④克雷计算机公司②的①生存⑩似乎⑫很大程度上⑪取决⑬于⑭创造力⑮和⑯寿命，⑱其⑲主席⑳和㉑首席设计师⑰...

The revised translation：

③分拆出的④克雷计算机公司②的①生存⑩似乎⑫很大程度上⑪取决⑬于⑱其⑲主席⑳和㉑首席设计师⑰的⑭创造力⑮和⑯寿命

Analysis of errors: there is still 1 structural error and component omission.

Compared with the original result of MT output, there is a reduction of 2 structural errors and 1 incorrect lexical choice. The compared results show that making the subject and the predicate adjacent can strengthen the bondage of their meanings, thus the parsing of NT clauses also can help eliminate incorrect lexical choices.

4 Statistics and Analysis of Errors in English-Chinese Machine Translation

4.1 Analysis of Error Types

We use the method described above to carry on manual evaluation in English-Chinese machine translation. So far, 243 English sentences have been manually evaluated, including 6232 English word tokens in 682 NT clauses. Total amount of errors in MT

outputs is 606 in whole sentence translation. According to our categorization, the proportion of error types in MT output among the 243 English sentences is presented in the pie chart below.

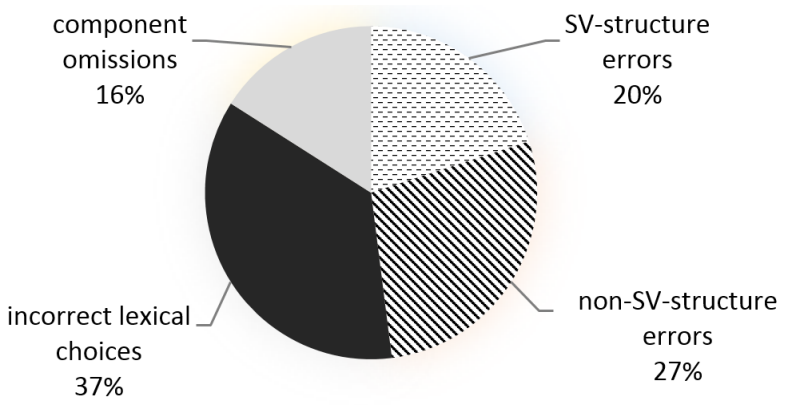

Fig. 12. Distribution of error types

According to the data in Fig. 12, we can see the specific proportion of each error type. The structural error accounts for 47%, in which the proportions of SV-structure errors and non-SV-structure errors are respectively 20% and 27%. Incorrect lexical choices take up 37%, and component omissions 16%. Then we can see that structural errors nearly takes up about a half of all errors, in which non-SV-structure errors are more than SV-structure errors obviously.

Statistic data in table 1 show the distribution of SV-structure errors and non-SV-structure errors among SV clauses and non-SV clauses. According to data in table 1, non-SV-structure errors are 1.39 times as many as SV-structure errors. The causes for the majority of non-SV-structure errors can be attribute to these two aspects: firstly, the ratio of the number of non-SV clauses and SV clauses is 1.02: 1 which is not a very large number; secondly, a more important cause is that structural errors in each non-SV clause are greatly more than those in each SV clause, with the ratio of 1.33: 1. Therefore, we can conclude that non-SV clauses are main source of errors.

Table 1. The distribution of structural errors

Type of clause	Number of clauses	Number of structural errors	Average number of structural errors in one clause
SV clause	337	119	0.36
Non-SV clause	345	166	0.48
Non-SV/SV	1.02	1.39	1.33

4.2 Comparison between Whole Sentence Translation and NT Clause Translation

When parsing the English sentences which have errors in the MT output of whole sentence translation, into NT clauses and then inputting them into Baidu Translate, we obtain the result of MT output of these NT clauses and the distribution of errors. The table 2 is a comparative result of errors before and after parsing.

Table 2. Comparison of the error amount between two ways of translation

Error type / Error number	Structural errors			Incorrect lexical choices	Component omissions	Total
	SV-structure errors	non-SV-structure errors	Total			
Before parsing	119	166	285	222	99	606
After parsing	72	71	143	199	60	402
The reduction	47	95	144	23	39	204
The reduced percentage	39.5%	57.2%	50.5%	10.4%	49.4%	33.7%

According to the data in table 2, structural errors and component omission decline by about one half, and incorrect lexical choices decline by 10%. The total reduction is one third. Non-SV-structure errors decline much more than SV-structure errors. We maintain that there are two causes for the reduction. Firstly, when English sentences (especially long ones) are parsed into NT clause, their naming and telling are linked together, which can strengthen syntactic and semantic constraint, thus largely eliminating ambiguity of phrases; secondly, that sentences are shortened after parsing also helps reduce incorrect lexical choices and omissions.

4.3 Correlation between the NT Clause Number in a Sentence and the Error Number in Its Whole-Sentence Translation

Based on preliminary observation of machine translated results, we predict that machine translation has a better performance in short sentences than in long sentences. In order to verify this predication, we analyze data of the number of NT clauses in English sentences and the distribution of errors among them respectively. The following line chart shows the relationship between these two variables.

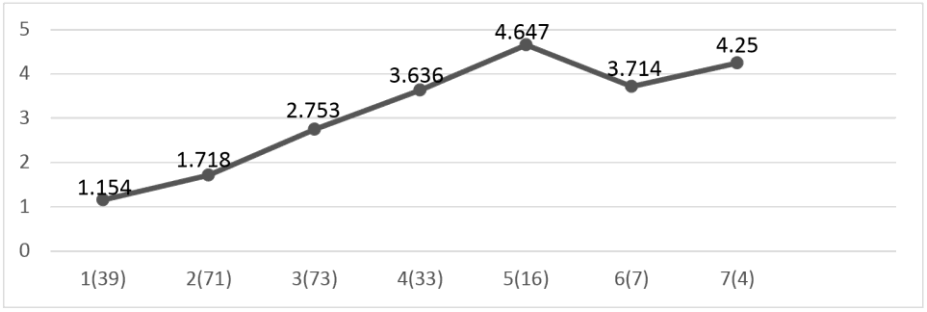

Fig. 13. Correlation between the NT clause number in a sentence and the error number in its whole-sentence translation

In Fig. 13, the horizontal axis represents the numbers of NT clauses in sentences and numbers in brackets represent the total amount of sentences which contain each number of NT clauses in our investigation. The vertical axis represents the average number of errors. So the line in the figure represents the correlation between the number of NT clauses and the average number of errors in sentences. The line shows a positive correlation between these two variables when the number of NT clauses ranges from 1 to 5 in sentences, which is consistent with our predication.

The line in Fig. 13 shows that errors occurring in sentences with 6-7 NT clauses are less than errors in sentences with 4-5 NT clauses, which apparently contradicts with our previous predication. As to causes for this phenomena, we think that compared with sentences containing 1-5 NT clauses, the number of sentences with 6-7 NT clauses is too small to be used as statistics data. Deeper causes for this phenomena need further studies.

5 Discussion

This paper applied the theory of NT clauses to the analysis of results of English-Chinese machine translation, and conclusions are made as follows.

Firstly, the majority of errors in machine translation is structural errors. These structural errors mainly exist in non-SV clauses including non-adjacent subject-predicate structures and adjacent or non-adjacent non-subject-predicate structures.

Secondly, as non-SV structures exist largely in long English sentences, so the above conclusion reveals the cause for the bottleneck of English long-sentence translation. Furthermore, the conclusion above suggests that English long sentences should be parsed into NT clauses before being translated. The suggested model has been proposed by Song & Ge [3], and examples in section 3 and data in section 4 of this paper support the suggestion. Definitely parsing long sentences of English will bring loss, and assembling NT clauses of Chinese into Chinese text after translation will also bring loss. These losses are not discussed in the paper. However, firstly parsing and assembling are tasks made in monolingual category, so their difficulties are apparently lower than tasks made in bilingual category; secondly, as NT clauses are

generally simple and short and if units in training corpus are all NT clauses, the quality of clause translation is expected to be further improved. Therefore, this model is worthy of further exploration.

Thirdly, the relationship between naming and telling, based on cognitive structures, is common to human language. All languages may be parsed into sequences of NT clauses according to relationships of naming and telling. On this point, the method adopted in this paper is language-independent. However, naming and telling are represented by different patterns in different languages. Eight naming-telling relationships have been summarized in this paper, which are based on features of English syntax. Therefore, this classification is language-dependent. The authors will give detailed discussion of this point in the future.

Lastly, translation of NT clauses are generally (not absolutely) independent of each other, because NT clauses are complete cognitive structures with self-sufficient meaning. Natural language is context dependent. For present computer processing power, it is difficult to analyze and translate long English sentences correctly in one shot. Long sentences can usually be parsed into several segments and the parsing method based on structures of NT clauses advocated in this paper is appropriate. Detailed information can be found in Song & Ge [3].

Acknowledgements. This research is supported by the 2016 Key Project of the National Languages Committee (ZDI135-30), Innovative School Project in Higher Education of Guangdong, China (GWTP-LH-2015-10), the Science and Technology Project of Guangdong Province, China (2016A040403113), National Natural Science Foundation of China (61171129) and the fund of Center for Translation Studies, Guangdong University of Foreign Studies (CTS2014-13).

References

1. Koehn, P.: Statistic Machine Translation. Trans. by Zong Chengqing, Zhang Xiaojun. Publishing House of Electronic Industry, Beijing (2012).
2. Zhao Hongmei, Xie Jun, Lv Yajuan, Yu Hui, Zhang Haoliang, Liu Qun: Common Error Analysis of Machine Translation Output. The Ninth China Workshop on Machine Translation, Kunming (2013).
3. Song Rou, Ge Shili: English-Chinese Translation Unit and Translation Model for Discourse-Based Machine Translation. Journal of Chinese Information Processing, pp. 125-135 (2015).
4. Song Rou: Stream Model of Generalized Topic Structure in Chinese Text. Chinese Language, pp. 483-494 (2013).