Comma Analysis and Processing for Improving Translation Quality of Long Sentences in Rule-based English-Korean Machine Translation

Sung-Dong Kim
School of Computer Engineering, Hansung University, Seoul, Republic of Korea

Keywords: English-Korean Machine Translation, Comma Usage Analysis, Natural Language Processing, Rule-based Machine Translation.

Abstract: Current English-Korean machine translation system cannot provide practical translation quality mainly due to the difficulties in long sentence parsing. Long sentences generally include commas, resulting in lots of different possible sentence structures. It is very difficult to accurately parse the long sentences that have commas. The roles of the commas in constructing sentences have to be identified and then the syntactic analysis should be performed according to the roles of the commas for accurate parsing of the long sentences. This paper presents the analysis results of the comma usages and the comma processing methods for each comma usage. And it also proposes the comma usage classification method using machine learning technique. In experiment, some improved translation results, by identifying comma usage and processing the commas, are also presented.

1 INTRODUCTION

Current English-Korean machine translation system generates relatively correct generation for short sentences. But it is not useful to use the system for the sentences over 20 words, which generally appear, because the system generates unnatural translations which are difficult to understand the meanings. In order to achieve the practical performance, which means the English-Korean machine translation system can be used in daily life, the system should be able to correctly parse the long sentences, which helps the system generate natural and understandable translation results.

The recent advances of neural machine translation (NMT) provide understandable translation in English to Korean translation. But NMT system is not appropriate for specific organizations to use for their own purposes. It is difficult for them to refine and improve the NMT system for their special usages. The rule-based MT system in this paper may contribute to the organizations that need their own MT system. Therefore, the rule-based MT system needs to achieve the performance comparable to NMT system. The solution to the problem in this paper will help improve the translation quality of the rule-based MT system.

Long English sentences generally include commas. Long sentence parsing is very difficult problem in statistical machine translation as well as in rule-based machine translation (Cettolo and Federico, 2006). In (Badr et al., 2008, Kim, 2013), intra-sentence segmentation method is applied to achieve faster and more correct translation. But the long sentences frequently appear with commas, so the analysis the comma usage and the processing with the identified comma useage is essential for correct parsing and translation. In (Jin et al., 2006), they present the segmentation method of Chinese sentences, in which they use the results of the comma classification.

We should identify the role of the comma before parsing in order to generate accurate syntactic structures and translation for the long sentences with commas. We consider the role of the commas as separation and list. The commas for separation split the sentence into sub-sentences such as preposition phrases, subordinate clauses, modifier phrases, and so on. The comma for list enumerates words and phrases for connecting them with coordinates conjunctions. This paper proposes the comma usage classification

Kim, S.
Comma Analysis and Processing for Improving Translation Quality of Long Sentences in Rule-based English-Korean Machine Translation.
DOI: 10.5220/0007310604740479
In *Proceedings of the 11th International Conference on Agents and Artificial Intelligence (ICAART 2019)*, pages 474-479
ISBN: 978-989-758-350-6

methods using support vector machine. Also we present the comma usages and the method with which we should process the commas before parsing.

In section 2, we show the comma usages in constructing sentences. The comma role classification method is explained in section 3. Section 4 describes the comma processing methods and the experimental results are shown in section 5. Section 6 concludes the paper with presenting further works.

2 COMMA USAGES

Commas are frequently used in constructing long sentences, and they help the reader understand the meaning of the sentence. In (Darling), they explain how to use commas in writing sentences in view of linguistics. In (Bayraktar et al., 1998, Srikumar et al., 2008, Arivazhagan et al., 2016), they show various comma usages and emphasize the importance of identifying the roles of the commas for accurately understanding the meaning of the sentences. In Table 1, we show the roles of the commas by analysing the comma usages in the sentences.

In row 6, "parenthetical elements", can be removed without changing the meaning of the sentence and be considered as "added information," and "interrupter" means inserted words, phrases, or clauses that block the logical flow of the sentence.

In view of English-Korean machine translation, we can classify the comma usage in Table 1 as follows: Usage 1, 2, and 7 are related to commas connecting sentence elements and usage 3, 4, 5, 6 and 8 are for the commas that plays the role of separating the sentence elements. The comma for the usage 9 is regarded as doing the role of building special patterns. Therefore, we can consider the problem of comma role identification as the two-class classification problem: connection vs. separation.

Table 1: Comma Usages and Examples.

	Comma Usages	Examples
1	Connection of elements listed in a series	My frequent uses of the Internet is sending e-mail, surfing the Web, and using chat rooms.
2	Connection of cause with coordinate conjunctions	The public seems eager for some kind of gun control legislation, but the congress is obviously too timid to enact any truly effective measures.
3	Separation of sentence-initial elements	Honestly, why would you think that?
4	Separation of sentence-final elements (phrase, subordinate clause)	A face-to-face meeting with Mr. Gorbachev should damp such criticism, though it will hardly eliminate it. She ran faster, her breath coming in deep gasps.
5	Separation of additive(nonessential part for the intended meaning) phrases/clauses	A Western Union spokesman, citing adverse developments in the market for high-yield junk bond, declined to say what alternatives are under consideration. Almost all of the shares in the 20-stock Major Market Index, which mimics the industrial average, were sharply higher.
6	Separation of parenthetical elements	(1) appositive - Robert Frost, perhaps America's most beloved poet, died when he was 88. (2) interjection - There are, of course, many points of view that we must consider before voting. (3) interrupter - The new bacteria recipients of the genes began producing pertussis toxin which, because of the mutant virulence gene, was no longer toxic.
7	Connection of adjectives modifying same noun	It was a long, noisy, and nauseating flight.
8	Separation of quoted sentences	"We can't see into the future," said the President, "but we have to prepare for it nonetheless."
9	Geographical names, Date	The wedding date was set for August 5, 2000. The conference was originally set for Geneva, Switzerland.

3 COMMA ROLE CLASSIFICATION METHOD

This section explains the comma role classification method. The target roles of commas are separation (usage 3, 4, 5, 6, 8) and connection (usage 1, 2, 7) as described in section 2. The detailed role classification requires more elaborate features, so we leave the problem as a future work. We choose simple and fast machine learning method, Support Vector Machine, for classifying commas into separation role commas and connection role commas. The problem is two-class classification, the number of features is 12, and

we have small size training data. SVM shows good performance under these circumstances, so we adopt SVM as classification algorithm.

We choose 12 features for classifying comma roles as shown in Table 2. The features are the number of words, part-of-speech (POS), the number of commas, the ordinal number of the comma in the sentence, word forms such as past/present participle, and the existence of coordinate conjunctions. For example, a sentence, "*My frequent uses of the Internet is sending e-mail, surfing the Web, and using chat room*", has two commas. [*My frequent uses of the Internet is sending e-mail*] is the left side of the first comma, [*surfing the Web*] is the right side. Also [*surfing the Web*] is the left side of the second comma, and [*and using chat room*] is the right side. As a result, the features for the first comma are as follows: l_length=9, r_length=3, l_first_POS=ADJ, l_last_POS=NOUN, r_first_POS=VERB, c_count=2, c_ord=1, c_count_ord=21, r_first_coordConj=0, r_first_pastp=0, r_first_three_presp=1, r_coordConj_exist=0. They are represented as integer values as follows: 9, 3, 1, 3, 4, 2, 1, 21, 0, 0, 1, 0. In the same manner, features for the second comma as follows: 3, 4, 4, 3, 8, 2, 2, 22, 1, 0, 1, 1. These are the input for SVM training and test.

Table 2: Features for Classifying Comma Roles.

Feature	Description
l_length	# of words in left side of the comma
r_length	# of words in right side of the comma
l_first_POS	POS of the first word in left-side of the comma
l_last_POS	POS of the last word in left-side of the comma
r_first_POS	POS of the first word in right-side of the comma
c_count	# of commas in the sentence
c_ord	order of the comma in the sentence
c_count_ord	Combination of total comma count and the order of the comma
r_first_coordConj	whether POS of the first word in the right side of the comma is coordinate conjunction or not
r_first_pastp	whether the first word in the right side of the comma is a past participle form or not
r_first_three_presp	whether one of the the first three words in the right side of the comma is a present participle form or not
r_coordConj_exist	Whether one of the words in the right side of the comma is a coordinate conjunction or not

4 COMMA PROCESSING METHOD

Figure 1 shows the translation process of the English-Korean machine translation system in this paper. The system uses sentence segmentation method for efficient analysis. An input sentence is segmented at the comma positions in the *1st segmentation*. Among the resulting segments, long segments (currently longer than 15 words) are again split in the *2nd segmentation* step. In *parsing* step, each segment is parsed and the resulting structures are combine in *parse tree combination* step to generate the final sentence structure. In order to generate accurate translation, *1st segmentation* and *parse tree combination* steps are performed differently according to the identified role of the commas. This section describes the comma processing methods according to the roles of the commas. The *comma role classification* step, explained in section 3, lies between *lexical analysis* and *1st segmentation* as shown in Figure 1.

In Table 1, we classify comma uses into 3 types: connection, separation, and special pattern. Table 3, 4 and 5 present the comma processing methods for each type respectively.

Table 3: Comma Processing Method 1.

Comma Usage	Processing Methods and Examples
Connection of sentence elements	Rewrite commas into "and"
	(usage1) My frequent uses of the Internet is sending e-mail ***and*** surfing the Web ***and*** using chat rooms. (usage7) It was a long ***and*** noisy ***and*** nauseating flight.
	split into independent translation units -> translate each translation unit -> combine the translation results
	(usage2) ① The public seems eager for some kind of gun control legislation. ② ***but*** the congress is obviously too timid to enact any truly effective measures.

Commas with connection role (usage 1, 7) can be rewritten to "and". As a result, the elements separated by commas are analysed together instead of being treated independently. The elements now are not segmented in *1st segmentation* step. For the usage 2, the input sentence is split into two translation units. The translation process (from *1st segmentation* to *parse tree combination* steps) performs on each translation unit. In this case two translation process

generate two translations for each translation unit. The resulting translations are simply merged into the final translation.

Sentences with commas with separation role (usage 3, 4, 5, 6) can be analysed step-by-step as shown in Figure 1. In case of *interrupter* in usage 6, the interrupter should be relocated by rewriting. This paper classifies the roles of comma into connection and separation, thus the sub-case identification after the role classification is not addressed in this paper. For the usage 8 separating quoted sentences, the input sentence is split into independent translation units by commas and translation process performs on each translation unit.

Special patterns (date, geographical names, …) including commas (usage 9) are recognized as one unit in the pre-processing step (Kim, 2011).

Table 4: Comma Processing Method 2.

Comma Usage	Processing Methods and Examples
Separation of sentence elements	Process as shown in Figure 1.
	(usage3) [Honestly] , [why would you think that?] (usage4) [A face-to-face meeting with Mr. Gorbachev should damp such criticism] , [though it will hardly eliminate it.] (usage5) [Almost all of the shares in the 20-stock Major Market Index] , [which mimics the industrial average] , [were sharply higher.] (usage6) appositives - [Robert Frost] , [perhaps America's most beloved poet] , [died when he was 88.] (usage6) interjections - [There are] , [of course] , [many points of view that we must consider before voting.]
	Rewrite sentence with repositioning the interrupter.
	(usage6) interrupter - [The new bacteria recipients of the genes began producing pertussis toxin which was no longer toxic] , [because of the mutant virulence gene]
	Process as shown in Figure 1.
	(usage8) ① We can't see into the future. ② said the President. ③ but we have to prepare for it nonetheless.

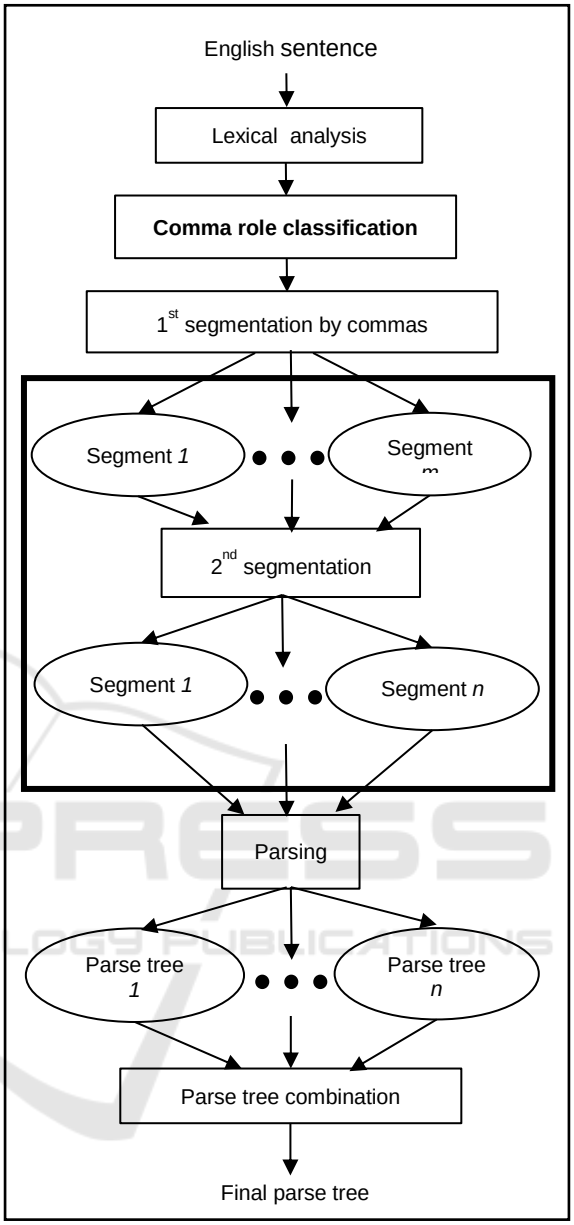

Figure 1: Analysis Steps in English-Korean Machine Translation System.

Table 5: Comma Processing Method 3.

Comma Usage	Processing Methods and Examples
Special patterns	Combine into one unit
	(usage9) The wedding date was set for ***August 5, 2000***. (usage9) The conference was originally set for ***Geneva, Switzerland.***

5 EXPERIMENTS

In this section, we present the statistics of the commas in the sentences extracted from "*Daily Joongang English News.*" And, we show the comma role classification performance using the SVM method. Table 6 shows the statistics for commas shown in sentences of the news articles. We collect statistics from 1,686 sentences in economy part, 1,751 sentences in science/technology part, and 1,871 sentences in industry part. The number of words in total is 111,304. Among 5,308 (*total # of sentences*), 3,422 sentences (about 65%, *total # of sentences including commas*) include commas. About 92% sentences among them (sentences including commas) include up to 3 commas. Therefore, we collect training data from those sentences for classifying comma roles. Among 5,308 sentences, 5,048 sentences (about 95%) have 0 ~ 3 commas, so translation accuracy for practical sentences (generally frequently appears) can be improved apparently through identifying comma roles in those sentences.

Table 6: The Distribution of the Number of Sentences according to the Number of the Commas.

	E	I	S-T	Total	Distribution (%)	
					Dist_1	Dist_2
0	589	645	652	1,886	35.5	-
1	550	583	486	1,619	31.9	47.3
2	356	417	318	1,091	20.6	31.7
3	134	154	164	452	8.5	13.2
4	50	49	71	170	3.2	5
5+	18	34	38	90	1.7	2.6

In Table 6, 'E', 'I', and 'S-T' mean economy, industry, and science-technology domain, respectively. '5+' means 5 or more commas. And 'Dist_1' and 'Dist_2' are calculated as follows:

$$\text{Dist}_1 = \frac{\#\ of\ sentences}{total\ \#\ of\ sentences}$$

$$\text{Dist}_2 = \frac{\#\ of\ sentences}{total\ \#\ of\ sentences\ including\ commas}$$

Table 7 shows the comma usage distribution in 313 sentences from Table 6. The separation usage (3, 4, 5, 6, and 8) takes about 85%. This can be the baseline accuracy of the SVM classification below.

We construct training and test data for SVM for classifying comma roles from the sentences in "*Daily Joongang English News.*" The target sentences include 1, 2 or 3 commas. The training data consists of 900 sentences in which 477 sentences include 1 comma, 306 sentences include 2 commas, and 117 sentences include 3 commas. This accords to the Dist_2 in Table 6. Also we have 90 test sentences (48, 31, and 11 sentences including 1, 2 and 3 commas, respectively). We assume there is no ill-used commas in the training and test data. So commas in the data fall in one of the categories (separation and connection) Table 8 shows the SVM classification accuracies using several kernels function such as linear, polynomial, RBF, and sigmoid.

Table 7: Comma Usage Distribution.

	Usage	Count/ Dist. (%)
1	Connection of elements list in a series	63 (10)
2	Connection of clause with coordinate conjunctions	19 (3)
3	Separation of sentence-initial elements	**121 (19.2)**
4	Separation of sentence-final elements	**242 (38.4)**
5	Separation of additive phrases/clauses	**54 (8.6)**
6	Separation of parenthetical elements	**86 (13.7)**
7	Connection of adjectives modifying same noun	5 (0.8)
8	Separation of quoted sentences	**36 (5.7)**
9	Geographical names, Date, ...	4 (0.6)

Table 8: The Comma Role Classification Accuracy of SVM.

Kernel function	Training Time (sec)	Training accuracy (%)	Test accuracy (%)
Linear	**0.031**	91	**94**
Polynomial	655.78	94	**94**
RBF	0.047	93	92
Sigmoid	0.047	88	90

Table 8 shows that simple linear kernel function suffices for the comma role classification problem with very fast training time and relatively good test accuracy.

Some translation results, with the help of the proposed comma processing method, are given with the old results in Table 9. In the comparisons of translation results, phrases with italicized and underlined are the improved part with the help of the comma role classification and the corresponding comma processing methods. In the first sentence in Table 9, meaningless target word "그런데" is removed to be more natural translation. In the second sentence, the translation of the phrase "because of the mutant virulence" is placed on the head of the translation, which is better in understanding the meaning. Also, the translation of the clause "when he

was 88" locates on the proper position in translation.

Table 9: Comparisons of Some Translation Results.

	Source Sentence
	Old Translation Result
	New Translation Result
1	"We can't see into the future," said the President, "but we have to prepare for it nonetheless."
	"우리는 그 미래를 조사할 수 없습니다," 대통령이 말했습니다. *그런데* "그러나 우리는 그것을 그럼에도 불구하고 준비해야 합니다."
	"우리는 그 미래를 조사할 수 없습니다," 대통령이 말했습니다. "그러나 우리는 그것을 그럼에도 불구하고 준비해야 합니다."
2	The new bacteria recipients of the genes began producing pertussis toxin which, because of the mutant virulence gene, was no longer toxic.
	그 유전자의 새로운 박테리아 수용체는 *그 돌연변이한 독성 유전자 때문에* 더 이상 독이 있지 않은 백일해균 독소를 생산하기 시작했습니다.
	그 돌연변이한 독성 유전자 때문에, 그 유전자의 새로운 박테리아 수용체는 더 이상 독이 있지 않은 백일해균 독소를 생산하기 시작했습니다.
3	Robert Frost, *perhaps America's most beloved poet*, died when he was 88.
	그가 88 이었을 때 Robert Frost, 아마 가장 사랑하는 미국의 시인은 죽었습니다.
	Robert Frost 는, 아마 가장 사랑하는 미국의 시인, *그가88 이었을 때* 죽었습니다.

6 CONCLUSIONS

We say that the proper comma processing is required for improving translation quality in rule-based English-Korean machine translation. For the purpose, we analysis the comma usages and the roles of the commas, and then propose the classification method for comma roles and the comma processing methods according to the identified comma roles.

We show 9 types of comma usages and classify the roles into connection, separation and forming special patterns. The SVM classification method shows better accuracy than the baseline accuracy predicting all commas as a separation role.

We need to develop the analysis algorithm according to the identified comma roles as explained in Table 3, 4 and 5. The sophisticated classification method for identifying the exact comma usage rather than connection and separation will be the future work. The study will contribute to accurate analysis of long sentences including commas, thus translation quality for rule-based English-Korean machine translation can reach the practical usage.

ACKNOWLEDGEMENTS

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (No. NRF- 2017R 1D 1A 1B03030878).

REFERENCES

Cettolo, M., Federico, M., 2006. Text Segmentation Criteria for Statistical Machine Translation. T. Salakoski et al. (Eds.): FinTAL 2006. LNAI 4139. pp. 664-673.

Badr, I., Zbib, R., and Glass, J., 2008. Segmentation for English-to-Arabic Statistical Machine Translation. In Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics. pp. 153-156.

Kim, S.-D., 2013. Intra-sentence Segmentation using Finite Automata for Efficient English Syntactic Analysis. Journal of KIISE: Computing Practices and Letter. Vol. 19, No. 4. pp. 186-193.

Jin, M., Kim, M.-Y., and Lee, J.-H., 2006. Segmentation of Long Chinese Sentences using Comma Classification. Journal of KISS(B): Software and Applications. Vol. 33, No. 5

Darling, C., Rules for Comma Usage. http://grammar.ccc.commnet.edu/grammar/commas.htm

Bayraktar, M., Say, B. and Akman, V., 1998. An Analysis of English Punctuation: The Special Case of Comma. International Journal of Corpus Linguistics. Vol. 3, No. 1, pp. 33-57.

Srikumar, V., Reichart, R., Sammons, M., Rappoport, A., and Roth, D., 2008. Extraction of Entailed Semantic Relations Through Syntax-based Comma Resolution. In Proceedings of ACL-08, pp. 1030-1038.

Arivazhagan, N., Christodoulopoulos, C., and Roth, D., 2016. Labeling the Semantic Roles of Commas. In Proceedings of the 13th AAAI Conference on Artificial Intelligence (AAAI-16). pp. 2885-2891.

Kim, S.-D., 2011. Pre-Processing Tasks for Rule-Based English-Korean Machine Translation System. In Proceedings of the 3rd International Conference on Agents and Artificial Intelligence. Vol. 1, pp. 257-262.