Towards a Linguistically Motivated Segmentation for a Simultaneous Interpretation System

Youngeun Koo[1], Jiyoun Kim[1], Jungpyo Hong[1], Munpyo Hong[1*] and Sung-Kwon Choi[2]
[1]Dept. of German Linguistics & Literature, Sungkyunkwan University,
25-2, Sungkyunkwan-ro, Jongno-gu, Seoul, Korea
[2]Language Intelligence Research Section, Electronics and Telecommunications Research Institute(ETRI), 218 Gajeong-ro, Yuseong-gu, Daejeon, Korea
{sarah8835,kite92,jphong2800,skkhmp}@skku.edu;choisk@etri.re.kr

Abstract

For simultaneous interpretation, it is very important to identify appropriate segmentation boundaries so that the source text can be translated accurately and promptly. This paper proposes four different segmentation methods for a simultaneous interpretation system. These methods are designed considering the balance between translation accuracy and translation latency. They employ various linguistic features such as prosodic, part-of-speech (POS), dependency, discourse, and cognitive features. This paper conducts experiments on segmentation in English to Korean and Korean to English simultaneous interpretation. Our finding shows that different segmentation method should be applied depending on the source language.

1 Introduction

Simultaneous interpretation aims to accurately translate what is being said in a source language into a target language quickly. To this aim, a strategy that segments the source text at appropriate points is often used by both human interpreters and simultaneous interpretation systems. Ideally, simultaneous interpretation systems should provide interpretation results as soon as possible while minimizing translation latency. However, there is a trade-off between translation accuracy and latency. The longer the segmentation unit is, the higher the translation accuracy will be, while the latency gets worse. In contrast, if the segmented unit is short, the latency will be better; however, the accuracy tends to be worsened.

In this paper, we investigate various segmentation features to determine the optimal segmentation points. The features were designed through a linguistic investigation into the prosodic, POS, dependency, and discourse-level characteristics in simultaneous interpretation. Also, we tried to find out what the appropriate segmentation length should be. In this paper, we propose four methods to derive optimal segmentation points employing these features. The segmentation features and methods considering both translation accuracy and latency may help to improve the performance of a simultaneous interpretation system.

In section 2, related works are introduced. In section 3 we suggest linguistic features for the segmentation. Section 4 shows our experimental setup with proposed features and methods and analyses the results. Finally, Section 5 concludes this paper and discusses future researches.

2 Related Works

Previous researches showed various approaches for investigating segmentation boundaries. The simplest way is to find a possible sentence unit (Cettolo and Federico, 2006; Sridhar et al., 2013b). Sridhar et al. (2013b) found segmentation boundaries based on predicting possible sentence end. Also, Sridhar et al. (2013b) utilized commas in sentences for segmentation.

[*] Corresponding author

Another approach for segmentation boundary detection is to use POS of the source text (Stolcke and Shriberg, 1996; Sridhar et al., 2013b; Nakabayashi et al., 2019). Stolcke and Shriberg (1996) tested two models for segmentation based on the POS of an input. The first model used POS tags labeled on every token and the second model used both POS and 'segmentation related' information, such as filled pause and discourse markers like 'okay', 'well'. Nakabayashi et al. (2019) found segmentation boundaries by aligning source text with target text made by human interpreters. Based on the analysis of segmentation boundaries, except for punctuation marks, coordinate conjunctions showed the highest rank followed by wh-words, adverbs, prepositions, and subordinate conjunctions.

Some researches focused on pause for segmentation (Kashioka, 2002; Bangalore et al., 2012). Bangalore et al. (2012) tried various lengths as a threshold of a meaningful pause and found that pauses over 100ms are meaningful for segmentation.

Such features mentioned above are derived from the aspect of a translation quality. Meanwhile, some studies focus on the translation latency (Cettolo and Federico, 2006; Rao et al., 2007; Sridhar et al., 2013b; Ma et al., 2019). Cettolo and Federico (2006) established segmentation boundaries every 10, 20, 30, 40, 50, 60, or 70 words and compared the translation quality of each approach. Ma et al. (2019) proposed to train a neural network model based on prefix-to-prefix and start translating source text from *k*-words behind ('*k*-wait'). This allowed the model to predict words at the sentence final position and translate with less latency. Ma et al. (2019) stated that '5-wait', approximately 3 seconds, results in the highest performance.

3 Segmentation for Simultaneous Interpretation

In this section, we propose linguistically motivated segmentation features and methods for a simultaneous interpretation system. In section 3.1, we introduce the linguistic features for segmentation. These features are taken into account to find out how suitable the point is to determine the segmentation boundary. In section 3.2, we suggest four segmentation methods. They differ in what they put stress on, when deciding segmentation boundaries.

3.1 Segmentation Features

We propose various linguistic features of segmentation for simultaneous interpretation: prosodic, POS, dependency, discourse, and cognitive information. In our method, the 'segmentation score' is calculated based on these features to decide the segmentation boundaries.

3.1.1 Prosodic Information

Prosodic information such as height and loudness of a sound can give clues to appropriate segmentation boundaries. However, to the best of our knowledge, there is not enough research on the impact of prosodic information on segmentation. Instead, many researches have been made on the impact of prosodic information on Transition Relevance Places (TRPs). TRP is a concept in Conversational Analysis. It denotes an end of Turn Construction Units (TCUs), unit of an utterance (Sacks et al., 1974). In human conversation, we can easily guess when the partner's utterance will end and when we can begin our turn. Ishimoto et al. (2011) investigated relation between prosodic information and TRPs in Japanese conversation.

Based on some similarities between simultaneous interpretation and conversation, Koo et al. (2019) applied the relation between prosodic information and TRPs to the relation between prosodic information and segmentation boundaries. Koo et al. (2019) analyzed pitch and power contours near segmentation boundaries. As a result, Koo et al. (2019) assumed that the fall of both pitch and power leads to segmentation.

In this sense, this paper sets the fall of pitch and power as one of the linguistic features for segmentation. Not only that, pauses in source text hint segmentation boundaries. This paper deals with two types of pause. Pauses marked as 'SENT_STR' by an automatic speech recognition system are relatively short and recognized as a start of a sentence by the system. Whereas pauses marked as 'SENT_END' are relatively long and recognized as an end of a sentence. We included these pauses as a linguistic feature of segmentation.

3.1.2 Part-of-speech (POS) Information

Syntactic structures take different forms depending on the grammatical features of each language. Some

POS information, of both English and Korean, characterizes the phrasal or clausal boundaries.

In English, a conjunction is used to connect two linguistic units (e.g. sentences, clauses). Therefore, we can effectively split two units by segmenting before conjunctions.

Korean is a SOV language and the end of each sentence is marked by the sentence-final ending. It is therefore the most obvious feature that can be used to make a segment between two sentences. While sentence-final ending marks the end position of a sentence, conjunctive ending connects two clauses. Since a clause is a syntactically complete unit, we can segment after conjunctive endings.

So, for most of the languages, POS information is a useful feature to find segmentation boundary.

3.1.3 Dependency Information

In simultaneous interpretation, due to speech situations such as pauses and lapses, a source text may be segmented at inappropriate segmentation points. To solve this problem, Koo et al. (2019) mentioned the need for semantic features that can prevent a semantically cohesive unit from being segmented into two separate units. As these units lose their original meaning when segmented, they should be maintained unsegmented.

In this study, we elaborate on this idea and suggest dependency features. The dependency features that we propose are summarized in Table 1.

Language	Feature name	Value
English	Adjective + Noun	JJ+N*
	Determiner + Noun	DT+N*
	Modal Auxiliary Verb + Verb	MD+VB*
	Auxiliary Verb + Verb	VB*+VB*
	Phrasal Verb(Verb + particle)	VB*+RP
	POS(not Noun) + Preposition	not N+IN
Korean	Adjectivalization Ending + Noun	ETJ+[N* or XPN]
	Adjective + Noun	D+[N* or XPN]
	Case Particle for Adjective + Noun	FM+[N* or XPN]
	Bound Noun	ND
	Auxiliary Verb	VX

Table 1. Dependency Features

3.1.4 Discourse Information

People try to be coherent when they are talking. This coherence is usually achieved by structuralizing the talk. Rhetorical Structure Theory (RST) is a theory that explains the structure of a text using a hierarchy between the sentences inside the text (Mann and Thompson, 1987). Texts in RST are hierarchic, built on partial texts which make a certain relation to each other. If two sentences have distinctive rhetorical characteristics, a linguistic marker appears between these two sentences to show such a transition. We call that Rhetorical Structure Markers (RSMs).

Therefore, RSM is an effective segmentation feature that can capture the general rhetorical relation of the text. In this study, we collected RSMs for each language: 160 for English, 140 for Korean. Table 2 is the example of RSMs.

Type	English	Korean
Addition	additionally, also, likewise	게다가(gedaga), 또한(ttohan), 유사하게(yusahage)
Contrast	although, conversely, in contrast	~에도 불구하고 (edo bulguhago), 반대로(bandaero), 대조적으로(daejojeogeuro)
Emphasis	in particular, specifically, without a doubt	특히(teukhi), 구체적으로(guchejeogeuro), 의심의 여지없이 (uisimui yeojieopsi)

Table 2. Examples of Rhetorical Structure Markers

3.1.5 Cognitive Information

While the features mentioned above are the features that guarantee the translation quality, the length of the segmentation unit is a feature for maintaining an appropriate translation latency. In order to do that, it is necessary to set an appropriate length of segmentation units. Based on the results of the previous studies and the analysis of the simultaneous interpretation data, the optimal length of the simultaneous interpretation unit was set to 4.5 seconds in this study.

First, the optimal length of the segmentation unit is based on the previous study in the field of simultaneous interpretation. Ear-Voice Span (EVS) refers to the time it takes for an interpreter to hear the words spoken in the source language and then interpret them to the target language. In other words, EVS refers to the time it takes for a simultaneous interpreter to hear the source utterance and then obtain all the information needed to understand and interpret it.

Lederer (1978) showed that the average EVS occurring in the simultaneous interpretation of the English to French was measured between 3 and 6 seconds. Ono et al. (2008) analyzed the EVS in simultaneous interpretation between Japanese to English and English to Japanese. They found that the average EVS times were 4.532 seconds and 2.446 seconds, respectively. Also, according to Lee (2002), the English to Korean EVS averaged 3 seconds. Through the studies mentioned, it was found out that the segmentation unit for accurate translation was 3 seconds or more.

The length of the segmentation unit proposed in this study was set considering also the psychological state of the audience who listened to the interpretation in the target language. Sridhar et al. (2013a) found that the listeners feel psychologically tired when the lapse of more than 4 to 5 seconds occurs during the simultaneous interpretation. That is, in order to be a good simultaneous interpreter, the lapse does not occur for more than 5 seconds when interpreting the source text to target text.

The segmentation unit length of 4.5 seconds was also derived from the dataset, constructed by ‘Electronics and Telecommunications Research Institute (ETRI)’. As a result of examining the 'SENT_END' tag in 19 English files and 10 Korean files, the source text length between 'SENT_END' and the next 'SENT_END' averages 4.55 seconds in English and 3.57 seconds in Korean. Refer to section 4.1 for more detailed information about the data.

3.2 Segmentation Methods

In this section, we propose four different segmentation methods using segmentation features, mentioned in section 3.1, to detect segmentation boundaries. We conducted experiments using these segmentation methods and compared the results in section 4.

3.2.1 Method 1 (Koo et al., 2019)

Segmentation method 1 was proposed in Koo et al. (2019). Method 1 segments a sentence when the ‘priority feature of segmentation’ appears. When it does not appear until ‘optimal length of segmentation unit’, then segments at the point, within 3.5~5.5 seconds, that has the highest segmentation score. Here, segmentation score is calculated by the sum of the values of segmentation feature. In detail, while the ‘priority feature of segmentation’ in English to Korean (En→Ko) simultaneous interpretation is RSMs, the ‘priority feature of segmentation’ in Korean to English (Ko→En) is RSMs and final endings.

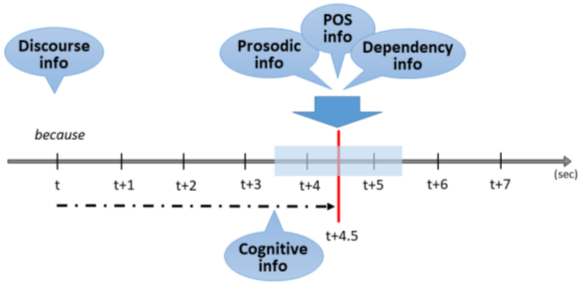

Figure 1. Flow of Segmentation Method 1

3.2.2 Method 2

Method 1 tends to segment only near 4.5 seconds, the range of 3.5~5.5 seconds, even if a better segmentation boundary is positioned right after that. Method 2 is designed to solve this limitation.

Like method 1, method 2 segments when ‘priority feature of segmentation’ appears. When it does not appear until ‘optimal length of segmentation unit’, then segmentation occurs at the point, after 4.5 seconds, where the segmentation score exceeds the threshold.

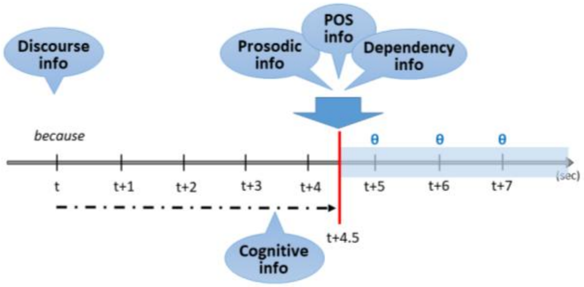

Figure 2. Flow of Segmentation Method 2

3.2.3 Method 3

Method 3 is similar to method 2, in that it segments when the segmentation score exceeds the threshold. However, method 3 gradually drops the threshold as time passes. This is inspired by human simultaneous interpreters. As time passes and the latency increases, they tend to accept less suitable points as segmentation boundaries, due to the pressure to give the audience a quick translation. Through this diminishing threshold, we expect less translation latency and guarantee translation quality.

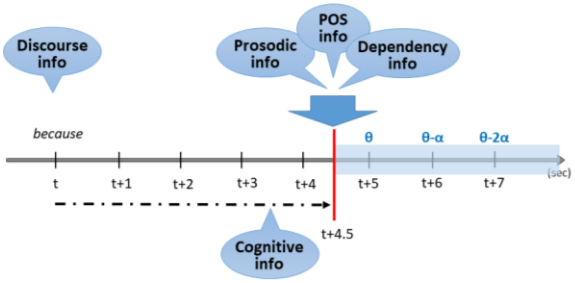

Figure 3. Flow of Segmentation Method 3

3.2.4 Method 4

Method 4 is similar to method 3, in that it considers both factors of simultaneous interpretation, translation quality and translation latency, when searching segmentation boundaries. However, method 4 directly utilizes latency as a variable for calculating the segmentation score.

To be specific, at the points before 4.5 seconds, method 4 focuses on guaranteeing only the translation quality and therefore uses linguistic features, mentioned in 3.1, for segmentation. On the other hand, at the points after 4.5 seconds, method 4 takes both the translation quality and translation latency into account for segmentation. Thus, methods 4 quantifies the latency and includes it as a feature for segmentation.

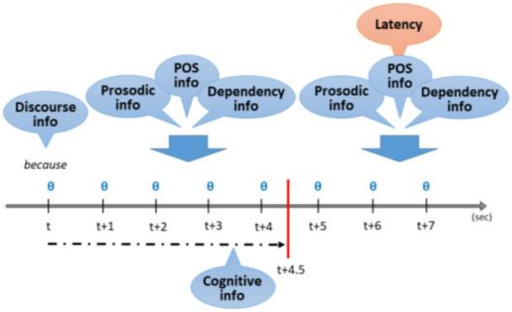

Figure 4. Flow of Segmentation Method 4

4 Experiments

In this section, we verify the linguistic features that we suggest and evaluate the best segmentation method for simultaneous interpretation systems. Section 4.1 explains about data used for the experiments. The evaluation results are shown in section 4.2. In section 4.3, we discuss and compare the results of the experiment.

4.1 Data

The experiment was conducted using the ETRI data, which are transcriptions of lectures. English data consist of 4 complete TED talks whose topics are artificial intelligence. The average length of videos is about 12 minutes long and the data include 6,711 tokens in total. Korean data are also composed of 4 lectures which are 'Sebasi' or 'K-MOOC' lectures with 5,193 tokens. Each video is about 12 minutes long on average.

As mentioned above, the transcription data include not only pause information but also POS tags. Additionally, we attached feature values for each token and determined whether the point should be segmented or not depending on each segmentation method.

4.2 Result

To evaluate the accuracy of the proposed methods, the acceptability of each segmentation point was calculated. As there is no absolutely correct answer for the segmentation points, only the acceptability of the points was taken into account. Table 3 shows the criteria that we set.

Grade	Criteria
Correct	The segmentation result contains all the syntactic and semantic pieces of the sentence, which are necessary for interpretation.
Acceptable	Some parts of information are missing, but still enough for interpretation.
Incorrect	Too much information is absent for interpretation.

Table 3. Criteria for Evaluation

Based on the criteria, three annotators who are native Korean speakers and possess a good command of English judged the appropriateness of segmentation points. Each annotator evaluated the accuracy of segmentation points, thus three results of accuracy evaluation were derived. The agreement rate among three annotators is 76.6%. Then we took the average of these three as the final accuracy.

We analyzed two measurements: strict and loose accuracy. Strict accuracy only considers 'correct' segmentation points, while loose accuracy includes 'correct' and 'acceptable' points.

		Methods			
		1	2	3	4
Accuracy (%)	**strict**	68.6	70.5	78.3	**80**
	loose	76.3	80.7	88.7	**88.4**
Duration(sec)		4.6	7.5	7	**3.6**

Table 4. Evaluation of Segmented Units (English)

As Table 4 shows the results of evaluating segmented units, which are split depending on each method. We designated 0.33 as a threshold for English data and 0.01 as time weight. The threshold was calculated from the average of the feature values of each segmentation point in other data. These segmentation points are marked by professional human translators.

Though method 3 in Table 4 shows the highest accuracy, its average duration takes about 7 seconds per each segmented unit. However, method 4 represents slightly lower but relatively similar accuracy to method 3. Also, the average duration of segmented units of method 4 is about 3.6 seconds, which is the lowest latency. Considering the trade-off between accuracy and latency, it implies that method 4 is the most proper method for English simultaneous interpretation.

Table 5 compares an original text with texts segmented by using the method 4. Compared to the original text, the segmented text shows that the text is properly segmented without hurting the original meaning and showing lower latency. As for the first segment, pause information played the crucial role in segmentation. The second segmentation was geared by both pause and POS information. The third and last segmentation were caused by RSM, prosodic and pause features.

	Segmented Unit	**Time**
Original Text	In such a brutal environment entrepreneurs learned to grow very rapidly they learned to make their products better at lightning speed and they learned to hone their business models until they're impregnable.	14.52
Segmented Text (Method 4)	in such a brutal environment entrepreneurs learned to grow very rapidly	6.14
	they learned to make their products better at lightning speed	3.94
	and they learned to hone their business models	2.41
	until they're impregnable	2.01

Table 5. Examples of Segmented Units – Original Text vs. Method 4

Table 6 shows the comparison of the results of each segmentation method for Korean. We specified 0.26 as a threshold for Korean data and 0.01 as time weight. The threshold was assigned in the same way as for English data.

		Methods			
		1	2	3	4
Accuracy (%)	**strict**	76	85.6	**88.2**	80.1
	loose	78.9	88.4	**92.6**	87.9
Duration(sec)		3.7	5.5	**5.1**	3.4

Table 6. Evaluation of Segmented Units (Korean)

The evaluation results indicate that the method 3 seems to be the most powerful method to segment Korean data with the topmost loose/strict accuracy with lower latency. In contrast to English data, method 4 results in noticeably lower accuracy compared to the method 3.

Compared to Koo et al. (2019) we added and elaborated dependency features. Table 7 shows the effects of them that induce better segmentation points. With this example, we can confirm that segmentation between adjective and noun is prevented by dependency features.

	Segmented Unit	Time
Without dependency features	파이썬 이라는 단어는요 원래 저기 뱀 이 큰 (The word Python is actually a huge) (paisseon iraneun daneoneunyo wonrae jeogi baem i keun)	4.78
	보아뱀이라고 하나요 (snake so-called Boa) (boabaemirago hanayo)	0.90
With dependency features	파이썬 이라는 단어는요 원래 저기 뱀 이 큰 보아뱀이라고 하나요 (The word ‘Python’ is actually a huge snake so-called Boa) (paisseon iraneun daneoneunyo wonrae jeogi baem i keun boabaemirago hanayo)	5.68

Table 7. Examples of Segmented Units – Effect of Dependency features

4.3 Discussion

As mentioned in Koo et al. (2019), method 1 tends to segment only near 4.5 seconds even if a better segmentation boundary is positioned right after that. Koo et al. (2019) expected that segmentation accuracy will increase if the system waits a little longer for a better segmentation boundary. As expected, method 2 showed better segmentation accuracy. Along with that, however, the average length of the segmented unit increased. This implies that method 2 caused more translation latency. Table 8 compares the segmentation result of method 1 and 2.

	Segmented Unit	Time
Method 1	now imagine an AI is helping a hiring manager find the next tech leader	5.15
	in the company	1.24
Method 2	now imagine an AI is helping a hiring manager find the next tech leader in the company	6.39

Table 8. Examples of Segmented Units – Method 1 vs. Method 2

We intended method 3 to alleviate translation latency by gradually dropping the segmentation threshold. As mentioned earlier, we expected less translation latency and guaranteed translation quality, when using segmentation method 3. As a result, it did keep high translation accuracy, but could not fully solve translation latency occurred in method 2. Refer to Table 9 for segmentation accuracy and an average length of segmented unit.

	Segmented Unit	Time
Method 2	Think about a pregnant woman in the Democratic Republic of Congo who has to walk seventeen hours to her nearest rural prenatal clinic to get a checkup what if she could get diagnosis on her phone instead	14.30
Method 3	Think about a pregnant woman in the Democratic Republic of Congo who has to walk seventeen hours to her nearest rural prenatal clinic to get a checkup	9.90

Table 9. Examples of Segmented Units – Method 2 vs. Method 3

The most critical problem of method 3 is that the segmented units are relatively long, which raises translation latency. To overcome this problem, method 4 brings in translation latency as a feature for segmentation. In addition, unlike method 3, method 4 checks every point whether it is an appropriate segmentation boundary. Consequently, method 4 resulted in a shorter average length of segmented unit, while maintaining high translation accuracy. Table 10 and 11 illustrates detailed information about the length of the segmented unit per methods for En→Ko and Ko→En.

Duration	Methods			
	1	2	3	4
Average	4.58	7.50	7.04	3.60
Minimum	0.15	0.10	0.10	0.10
Maximum	20.44	33.82	20.44	12.07

Table 10. Length of Segmented Units (English)

Duration	Methods			
	1	2	3	4
Average	3.73	5.35	5.10	3.36
Minimum	0.25	0.18	0.18	0.14
Maximum	9.03	21.21	14.55	8.95

Table 11. Length of Segmented Units (Korean)

Up to now, we looked through the segmentation results of each of four segmentation methods. We saw that each of them has different strengths and weaknesses. But not only that, they showed a different segmentation performance, depending on the source language.

According to the Table 10 and 11, regardless of the source language, segmentation accuracy is higher and average length of segmented unit is shorter, in the order of method 1, 2, and 3. Nevertheless, when comparing method 3 and 4, the result differs with regard to the source language. When segmenting English source text, method 3 and 4 led to similar segmentation accuracy, while method 4 produced considerably shorter segmented units. This indicates that method 4 can perform better for English when considering the trade-off between translation quality and translation latency.

Segmentation for Korean source text shows different aspects from that of English source text. When segmenting Korean source text, method 4 produced shorter segmented units, which implies less translation latency. However, method 4 caused relatively great decrease in segmentation accuracy when it was applied to Korean. This means that segmentation method 3 seems to work well for Korean source text.

This can be attributed to the typological difference between English and Korean. English is a head-initial language, so that a verb is located mostly in the front of a sentence. On the other hand, Korean is a head-final language and its verb appears in the back of a sentence. Since the latency is used as a feature for segmentation, method 4 results in more frequent segmentations after 4.5 seconds, the optimal length of segmentation unit. In this regard, when method 4 is applied to the Korean source text, it is likely that segmentation boundary occurs in between the verb phrase and leads to inappropriate segmentation. Therefore, different segmentation methods should be applied depending on the source language.

5 Conclusion and Future Works

In this paper, we proposed linguistically motivated segmentation features and methods to investigate segmentation units for simultaneous interpretation. Various features such as prosodic, POS, dependency, discourse and cognitive information were set for proper segmentation. Also, to prevent the length of the segment unit from being excessively long, we considered latency as a feature. Based on these features, four segmentation methods were proposed. The highest accuracy was achieved in method 4 (80%) for En→Ko and method 3 (88.2%) for Ko→En.

In the future study, the method of evaluating the segmented units should be further revised. In this study, when evaluating the segmented units, we judged only whether information in the segmented unit is sufficient to interpret. However, if the segmented unit contains other segmentation points inside itself, which should have been segmented, this unit should be penalized in the future.

Furthermore, we will check whether the interpretation result which is assisted by segmentation shows a significant performance difference compared to the interpretation result without segmentation. To this end, we plan to develop a suitable evaluation method for simultaneous interpretation that takes into account the differences between machine translation and simultaneous interpretation.

Acknowledgments

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (R7119-16-1001, Core technology development of the real-time simultaneous speech translation based on knowledge enhancement)

References

Akiko Nakabayashi and Tsuneaki Kato. 2019. Simulating Segmentation by Simultaneous Interpreters for Simultaneous Machine Translation. In Proceedings of the 33rd Pacific Asia Conference on Language, Information and Computation (PACLIC), 165-173.

Andreas Stolcke and Elizabeth Shriberg. 1996. Automatic linguistic segmentation of conversational speech. In Proceedings of the 4th International Conference on Spoken Language Processing (ICSLP), 2: 1005-1008.

Harvey Sacks, Emanuel A. Schegloff, and Gail Jefferson. 1974. A Simplest Systematics for the Organization of Turn-taking for Conversation, Language, 50: 696–735.

Hideki Kashioka. 2002. Translation Unit Concerning Timing of Simultaneous Translation. In Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC), 142-146.

Marianne Lederer. 1978. Simultaneous Interpretation – Units of meaning and Other Features. Language interpretation and communication. Springer, 323-332.

Mauro Cettolo and Marcello Federico. 2006. Text segmentation criteria for statistical machine translation. In Proceedings of the International Conference on Natural Language Processing, 664-673.

Mingbo Ma, Liang Huang, Hao Xiong, Renjie Zheng, Kaibo Liu, Baigong Zheng, Chuanqiang Zhang, Zhongjun He, Hairong Liu, Xing Li, Hua Wu, and Haifeng Wang. 2019. STACL: Simultaneous Translation with Implicit Anticipation and Controllable Latency using Prefix-to-Prefix Framework. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL), 3025-3036.

Sharath Rao, Ian Lane, and Tanja Schultz. 2007. Optimizing Sentence Segmentation for Speech Translation. In Proceedings of Interspeech2007, 2845–2848.

Srinivas Bangalore, Vivek K. R. Sridhar, Prakash Kolan, LadanGolipour, and Aura Jimenez. 2012. Real-time incremental speech-to-speech translation of dialogs. In Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT), 437-445.

Tae-Hyung Lee. 2002. Ear Voice Span in English into Korean Simultaneous Interpretation. Meta, 47(4):596-606.

Takahiro Ono, HitomiTohyama, and Shigeki Matsubara. 2008. Construction and Analysis of Word-level Time-aligned Simultaneous Interpretation Corpus. In Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC), 3383-3387.

Vivek K. R. Sridhar, John Chen, and Srinivas Bangalore. 2013a. Corpus analysis of simultaneous interpretation data for improving real time speech translation. INTERSPEECH, 3468-3472.

Vivek K. R. Sridhar, John Chen, Srinivas Bangalore, Andrej Ljolje, and Rathinavelu Chengalvarayan. 2013b. Segmentation strategies for streaming speech translation. In Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT), 230-238.

William C. Mann, and Sandra A. Thompson. 1988. Rhetorical structure theory: Toward a functional theory of text organization. Text-Interdisciplinary Journal for the Study of Discourse, 8(3):243-281.

Youngeun Koo, Jiyoun Kim, Jungpyo Hong, Munpyo Hong and Sung-Kwon Choi. 2019. A Study on Segmentation Unit for the Real-time Simultaneous Interpretation System. In Proceedings of the 31st Annual Conference on Human & Cognitive Language Technology (HCLT), 229-235.

Yuichi Ishimoto, Mika Enomoto, and Hitoshi Iida. 2011. Projectability of Transition-Relevance Places Using Prosodic Features in Japanese Spontaneous Conversation. In Proceedings of Interspeech2011, 2061–2064.