Selecting Syntactic, Non-redundant Segments in Active Learning for Machine Translation

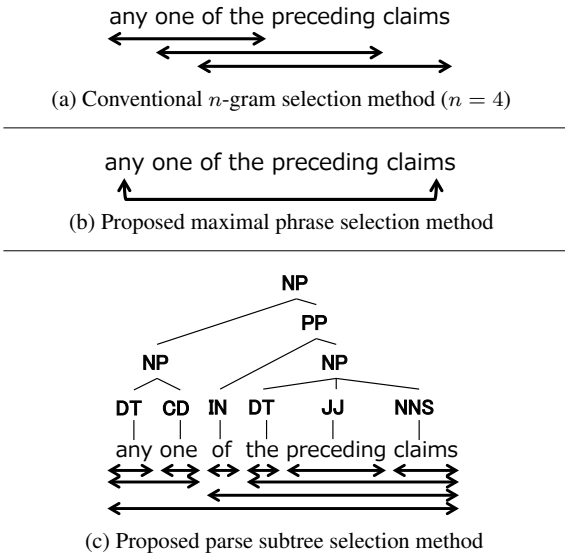
Akiva Miura[†], **Graham Neubig**[†], **Michael Paul**[‡], **Satoshi Nakamura**[†]
[†] Nara Institute of Science and Technology, Japan
[‡] ATR-Trek Co. Ltd., Japan
miura.akiba.lr9@is.naist.jp neubig@is.naist.jp
michael.paul@atr-trek.co.jp s-nakamura@is.naist.jp

Abstract

Active learning is a framework that makes it possible to efficiently train statistical models by selecting informative examples from a pool of unlabeled data. Previous work has found this framework effective for machine translation (MT), making it possible to train better translation models with less effort, particularly when annotators translate short phrases instead of full sentences. However, previous methods for phrase-based active learning in MT fail to consider whether the selected units are coherent and easy for human translators to translate, and also have problems with selecting redundant phrases with similar content. In this paper, we tackle these problems by proposing two new methods for selecting more syntactically coherent and less redundant segments in active learning for MT. Experiments using both simulation and extensive manual translation by professional translators find the proposed method effective, achieving both greater gain of BLEU score for the same number of translated words, and allowing translators to be more confident in their translations[1].

1 Introduction

In statistical machine translation (SMT) (Brown et al., 1993), large quantities of high-quality bilingual data are essential to achieve high translation accuracy. While in many cases large corpora can be collected, for example by crawling the web (Resnik and Smith, 2003), in many domains or language pairs it is still necessarily to create data by hand, either by hiring professionals or crowdsourcing (Zaidan and Callison-Burch, 2011). In these cases, *active learning* (§2), which selects which data to annotate based on their potential benefit to the translation system, has been shown to be effective for improving SMT systems while keeping the required amount of annotation to a minimum (Eck et al., 2005; Turchi et al., 2008; Haffari et al., 2009; Haffari and Sarkar, 2009; Ananthakrishnan et al., 2010; Bloodgood and Callison-Burch, 2010; González-Rubio et al., 2012; Green et al., 2014).



(a) Conventional n-gram selection method ($n = 4$)

(b) Proposed maximal phrase selection method

(c) Proposed parse subtree selection method

Figure 1: Conventional and proposed data selection methods

Most work on active learning for SMT, and natural language tasks in general, has focused on choosing which *sentences* to give to annotators. These

[1]Code to replicate the experiments can be found at https://github.com/akivajp/naacl2016

Proceedings of NAACL-HLT 2016, pages 20–29,
San Diego, California, June 12-17, 2016.

methods generally assign priority to sentences that contain data that is potentially useful to the MT system according to a number of criteria. For example, there are methods to select sentences that contain phrases that are frequent in monolingual data but not in bilingual data (Eck et al., 2005), have low confidence according to the MT system (Haffari et al., 2009), or are predicted to be poor translations by an MT quality estimation system (Ananthakrishnan et al., 2010). However, while the selected sentences may contain useful phrases, they will also generally contain many already covered phrases that nonetheless cost time and money to translate.

To solve the problem of wastefulness in full-sentence annotation for active learning, there have been a number of methods proposed to perform *sub-sentential annotation of short phrases* for natural language tasks (Settles and Craven, 2008; Bloodgood and Callison-Burch, 2010; Tomanek and Hahn, 2009; Sperber et al., 2014). For MT in particular, Bloodgood and Callison-Burch (2010) have proposed a method that selects poorly covered n-grams to show to translators, allowing them to focus directly on poorly covered parts without including unnecessary words (§3). Nevertheless, our experiments identified two major practical problems with this method. First, as shown in Figure 1 (a), many of the selected phrases overlap with each other, causing translation of *redundant* phrases, damaging efficiency. Second, it is common to see *fragments* of complex phrases such as "one of the preceding," which may be difficult for workers to translate into a contiguous phrase in the target language.

In this work, we propose two methods that aim to solve these two problems and improve the efficiency and reliability of segment-based active learning for SMT (§4). For the problem of overlapping phrases, we note that by merging overlapping phrases, as shown in Figure 1 (b), we can reduce the number of redundant words annotated and improve training efficiency. We adopt the idea of *maximal substrings* (Okanohara and Tsujii, 2009) which both encode this idea of redundancy, and can be calculated to arbitrary length in linear time using enhanced suffix arrays. For the problem of phrase structure fragmentation, we propose a simple heuristic to count only *well-formed syntactic constituents* in a parse tree, as shown in Figure 1 (c).

To investigate the effect of our proposed methods on learning efficiency, we perform experiments on English-French and English-Japanese translation tasks in which we incrementally add new parallel data, update models and evaluate translation accuracy. Results from both simulation experiments (§5) and 120 hours of work by professional translators (§6) demonstrate improved efficiency with respect to the number of words annotated. We also found that human translators took more time, but were more confident in their results on segments selected by the proposed method.

2 Active Learning for Machine Translation

In this section, we first provide an outline of the active learning procedure to select phrases for SMT data. In this paper, we regard a "phrase" as a word sequence with arbitrary length, which indicates that full sentences and single words both qualify as phrases. In Algorithm 1, we show the general procedure of incrementally selecting the next candidate for translation from the source language corpus, requesting and collecting the translation in the target language, and retraining the models.

Algorithm 1 Active learning for MT

1: **Init:**
2: $SrcPool \leftarrow$ source language data including candidates for translation
3: $Translated \leftarrow$ translated parallel data
4: $Oracle \leftarrow$ oracle giving the correct translation for an input phrase
5: **Loop Until** $StopCondition$**:**
6: $TM \leftarrow TrainTranslationModel(Translated)$
7: $NewSrc \leftarrow SelectNextPhrase(SrcPool, Translated, TM)$
8: $NewTrg \leftarrow GetTranslation(Oracle, NewSrc)$
9: $Translated \leftarrow Translated \bigcup \{\langle NewSrc, NewTrg \rangle\}$

In lines 1-4, we define the datasets and initialize them. $SrcPool$ is a set with each sentence in source language corpus as an element. $Translated$ indicates a set with source and target language phrase pairs. $Translated$ may be empty, but in most cases will consist of a seed corpus upon which we would like to improve. $Oracle$ is an oracle (e.g. a human translator), that we can query for a correct translation for an arbitrary input phrase.

In lines 5-9, we train models incrementally. $StopCondition$ in line 5 is an arbitrary timing when to stop the loop, such as when we reach an accuracy goal or when we expend our translation bud-

get. In line 6, we train the translation model using $Translated$, the available parallel data at this point. We evaluate the accuracy after training the translation model for each step in the experiments. In line 7, we select the next candidate for translation using features of $SrcPool$, $Translated$ and TM to make the decision.

In the following sections, we discuss existing methods (§3), and our proposed methods (§4) to implement the selection criterion in line 7.

3 Selection based on n-Gram Frequency

3.1 Sentence Selection using n-Gram Frequency

The first traditional method that we cover is a sentence selection method. Specifically, it selects the sentence including the most frequent uncovered phrase with a length of up to n words in the source language data. This method enables us to effectively cover the most frequent n-gram phrases and improve accuracy with fewer sentences than random selection. Bloodgood and Callison-Burch (2010) demonstrate results of a simulation showing that this method required less than 80% of the data required by randomly selected sentences to obtain the same accuracy.

However, as mentioned in the introduction, the selected full sentences include many phrases already covered in the parallel data. This may cause an additional cost for words in redundant segments, a problem resolved by the phrase selection approach detailed in the following section.

3.2 Phrase Selection using n-Gram Frequency

In the second baseline approach, we directly select and translate n-gram phrases that are the most frequent in the source language data but not yet covered in the translated data (Bloodgood and Callison-Burch, 2010). This method allows for improvement of coverage with fewer additional words than sentence selection, achieving higher efficiency by reducing the amount of data unnecessarily annotated. Bloodgood and Callison-Burch (2010) showed that by translating the phrases selected by this method using a crowdsourcing website, it was possible to achieve a large improvement of BLEU score, outperforming similar sentence-based methods.

However, as mentioned in the introduction, this method has several issues. First, because it uses short phrases, it often selects phrases that are not linguistically well-formed, potentially making them difficult to translate concisely. Second, it also has problems with redundancy, with no device to prevent multiple overlapping phrases being selected and translated. Finally, the previous work limits the maximum phrase length to $n = 4$, precluding the use of longer phrases. However, using a larger limit such as $n = 5$ is not likely to be a fundamental solution, as it increases the number of potentially overlapping phrases, and also computational burden. In the next section we cover our proposed solutions to these problems in detail.

4 Phrase Selection based on Maximal Phrases and Parse Trees

4.1 Phrase Selection based on Maximal Phrases

To solve both the problem of overlapping phrases and the problem of requiring limits on phrase length for computational reasons, we propose a method using the idea of *maximal substrings* (Okanohara and Tsujii, 2009). Maximal substrings are formally defined as "a substring that is *not* always included in a particular longer substring." For example, if we define p_i as a phrase and $occ(p_i)$ as its occurrence count in a corpus, and have the following data

p_1 = "one of the preceding", $occ(p_1) = 200,000$
p_2 = "one of the preceding claims", $occ(p_2) = 200,000$
p_3 = "any one of the preceding claims", $occ(p_3) = 190,000$

p_1 = "one of the preceding" always co-occurs with the longer p_2 = "one of the preceding claims" and thus is not a maximal substring. On the other hand, p_2 does not always co-occur with p_3, and thus p_2 will be maximal. This relationship can be defined formally with the following semi-order relation:

$$p_1 \preceq p_2 \Leftrightarrow \exists \alpha, \beta : p_1 = \alpha p_2 \beta \wedge occ(p_1) = occ(p_2). \quad (1)$$

Demonstrating this by the previous example, $p_1 = \alpha p_2 \beta$, α = "", β = "claims" hold, meaning p_1 is a sub-sequence of p_2, and p_2 is a sub-sequence of p_3 in a similar manner. Since p_1 is a sub-sequence of p_2 and $occ(p_1) = occ(p_2) = 200,000$, $p_1 \preceq p_2$ holds. However, although p_2 is a sub sequence of p_3,

because $occ(p_2) = 200,000 \neq 190,000 = occ(p_3)$, the relation $p_2 \preceq p_3$ does not hold. Here, we say p has *maximality* if there does not exist any q other than p itself that meets $p \preceq q$, and we call such a phrase a *maximal phrase*.

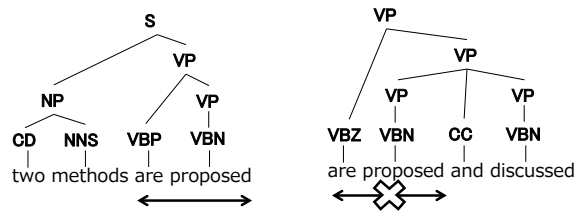
To apply this concept to active learning, our proposed method limits translation data selection to only maximal phrases. This has two advantages. First, it reduces overlapping phrases to only the maximal string, allowing translators to cover multiple high-frequency phrases in the translation of a single segment. Second, maximal phrases and their occurrence counts can be enumerated efficiently by using enhanced suffix arrays (Kasai et al., 2001) in linear time with respect to document length, removing the need to set arbitrary limits on the length of strings such as $n = 4$ used in previous work.

However, it can be easily noticed that while in the previous example p_2 is included in p_3, their occurrence counts are close but not equivalent, and thus both are maximal phrases. In such a case, the naïve implementation of this method can not remove these redundant phrases, despite the fact that it is intuitively preferable that the selection method combines phrases if they have almost the same occurrence count. Thus, we also propose to use the following semi-order relation generalized with parameter λ:

$$p_1 \overset{*}{\preceq} p_2 \Leftrightarrow \exists \alpha, \beta : p_1 = \alpha p_2 \beta \wedge \lambda \cdot occ(p_1) < occ(p_2). \quad (2)$$

where λ takes a real numbered value from 0 to 1, which we set to $\lambda = 0.5$ in this research.

This removes the restriction that the two phrases under comparison be of exactly equal counts, allowing them to have only approximately the same occurrence count. We redefine maximality using this semi-order $\overset{*}{\preceq}$ as *semi-maximality*, and call maximal phrases defined with $\overset{*}{\preceq}$ *semi-maximal phrases* in contrast to normal maximal phrases. By using semi-maximal phrases instead of maximal phrases, we can remove a large number of phrases that are included in a particular longer phrase more than half the time, indicating that it might be preferable to translate the longer phrase instead.



(a) "are proposed" is counted (b) "are proposed" is not counted

Figure 2: Phrase counting based on parse trees

4.2 Phrase Selection based on Parse Trees

In this section, we propose a second phrase selection method based on the results from the syntactic analysis of source language data. This method first processes all the source language data with a phrase structure parser, traverses and counts up all the subtrees of parse trees as shown in Figure 2, and finally selects phrases corresponding to a subtree in frequency order.[2] We propose this method because we expect the selected phrases to have syntactically coherent meaning, potentially making human translation easier than other methods that do not use syntactic information.

It should be noted that because this method counts all subtrees, it is capable of selecting overlapping phrases like the methods based on n-grams. Therefore we also experiment with a method using together both subtrees and the semi-maximal phrases proposed in Section 4.1 to select both syntactic and non-redundant segments.

5 Simulation Experiment

5.1 Experimental Set-Up

To investigate the effects of the phrase selection methods proposed in Section 4, we first performed a simulation experiment in which we incrementally retrain translation models and evaluate the accuracy after each step of data selection. In this experiment, we chose English as a source language and French and Japanese as target languages. To simulate a realistic active learning scenario, we started from given parallel data in the general domain and sequentially added additional source language data in a specific target domain. For the English-French translation task, we adopted the Europarl corpus

[2]The method does not distinguish between equivalent word sequences even if they have different tree structures

Lang Pair	Domain	Dataset	Amount
En-Fr	General (Base)	Train	1.89M Sent. En: 47.6M Words Fr: 49.4M Words
	Medical (Target)	Train	15.5M Sent. En: 393M Words Fr: 418M Words
		Test	1000 Sent.
		Dev	500 Sent.
En-Ja	General (Base)	Train	414k Sent. En: 6.72M Words Ja: 9.69M Words
	Scientific (Target)	Train	1.87M Sent. En: 46.4M Words Ja: 57.6M Words
		Test	1790 Sent.
		Dev	1790 Sent.

Table 1: Details of parallel data

from WMT2014[3] as a base parallel data source and EMEA (Tiedemann, 2009), PatTR (Wäschle and Riezler, 2012), and Wikipedia titles, used in the medical translation task, as the target domain data. For the English-Japanese translation task, we adopted the broad-coverage example sentence corpus provided with the Eijiro dictionary[4] as general domain data, and the ASPEC[5] scientific paper abstract corpus as the target domain data. For preprocessing, we tokenized Japanese corpora using the KyTea word segmenter (Neubig et al., 2011) and filtered out the lines of length over 60 from all the training parallel data to ensure accuracy of parsing and alignment. We show the details of the parallel dataset after pre-processing in Table 1.

For the machine translation framework, we used phrase-based SMT (Koehn et al., 2003) with the Moses toolkit (Koehn et al., 2007) as a decoder. To efficiently re-train the models with new data, we adopted inc-giza-pp,[6] a specialized version of GIZA++ word aligner (Och and Ney, 2003) supporting incremental training, and the memory-mapped dynamic suffix array phrase tables (MMSAPT) feature of Moses (Germann, 2014) for on-memory construction of phrase tables. We train 5-gram models over the target side of all the general domain and target domain data using KenLM (Heafield, 2011).

[3]http://statmt.org/wmt14/

[4]http://eijiro.jp

[5]http://lotus.kuee.kyoto-u.ac.jp/ASPEC/

[6]https://github.com/akivajp/inc-giza-pp

For the tuning of decoding parameters, since it is not realistic to run MERT (Och, 2003) at each retraining step, we tuned the parameters to maximize the BLEU score (Papineni et al., 2002) for the baseline system, and re-used the parameters thereafter. We compare the following 8 segment selection methods, including 2 random selection methods, 2 conventional methods and 4 proposed methods:

sent-rand: Select sentences randomly.

4gram-rand: Select n-gram strings of length of up to 4 in random order.

sent-by-4gram-freq: Select the sentence including the most frequent uncovered phrase with length of up to 4 words (baseline 1, §3.1).

4gram-freq: Select the most frequent uncovered phrase with length of up to 4 words (baseline 2, §3.2).

maxsubst-freq: Select the most frequent uncovered maximal phrase (proposed, §4.1)

reduced-maxsubst-freq: Select the most frequent uncovered semi-maximal phrase (proposed, §4.1)

struct-freq: Select the most frequent uncovered phrase extracted from the subtrees (proposed, §4.2).

reduced-struct-freq: Select the most frequent uncovered semi-maximal phrase extracted from the subtrees (proposed, §4.1 and §4.2).

To generate oracle translations, we used an SMT system trained on all of the data in both the general and target-domain corpora. To generate parse trees, we used the Ckylark parser (Oda et al., 2015).

5.2 Results and Discussion

Comparison of efficiency: In Figure 3, we show the evaluation score results by the number of additional source words up to 100k and 1M words. We can see that in English-French translation, the accuracy of the selection methods using parse trees grows more rapidly than other methods and was significantly better even at the point of 1M additional words. In the case of English-Japanese translation, the gains over 4-gram frequency are much smaller, but the proposed methods still consistently perform as well or better than the other methods. Besides, in all the graphs we can see the improvement of reduced-maxsubst-freq and reduced-struct-freq over maxsubst-freq and struct-freq respectively, demonstrating that avoiding selecting redundant segments is helpful in improving efficiency.

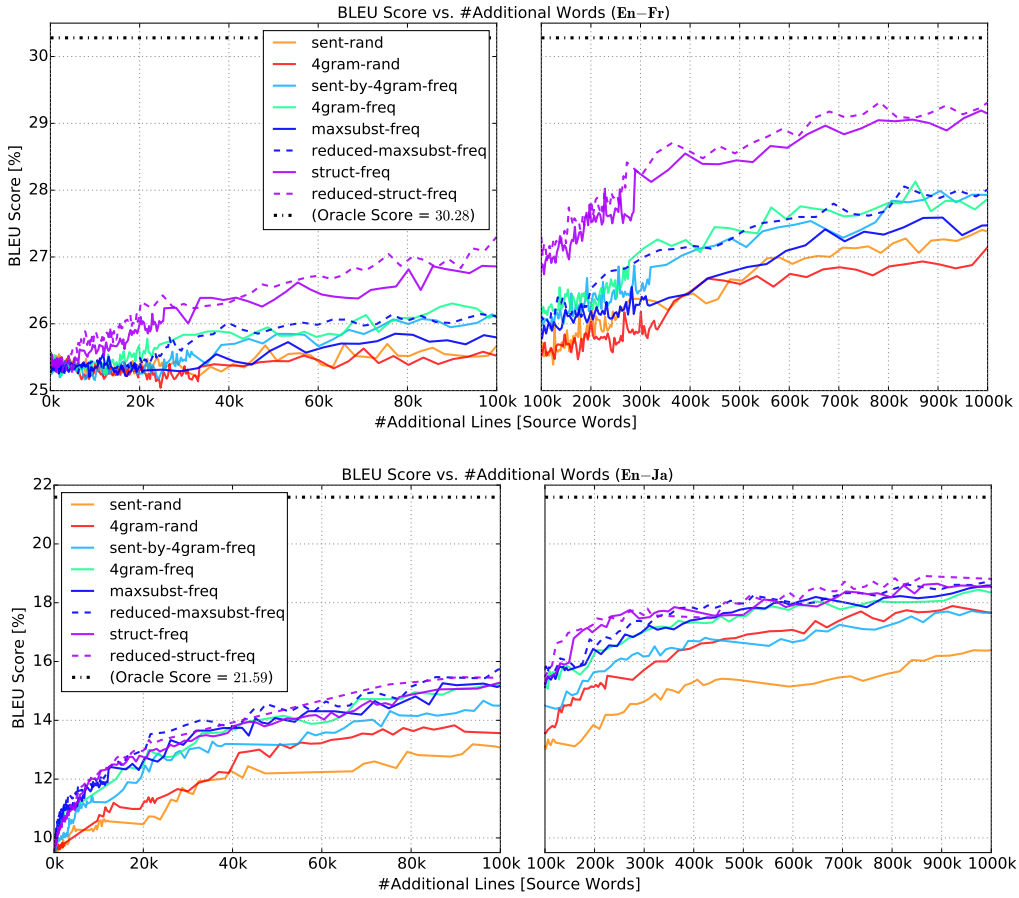


Figure 3: BLEU score vs. number of additional source words in each method

Length of selected phrases: Due to the different criteria used by each method, there are also significant differences in the features of the selected phrases. In Table 2, we show the details of the number of all selected phrases, words and average phrase length until the stop condition, and at the point of 10k additional source words. Here we see the tendency that the selection methods based on parse trees select shorter phrases than other methods. This is caused by the fact that longer phrases are only counted if they cover a syntactically defined phrases, and thus longer substrings that do not form syntactic phrases are removed from consideration.

Phrase coverage: This difference in the features of the selected phrases also affects how well they can cover new incoming test data. To demonstrate this, in Table 3 we show the 1-gram and 4-gram coverage of the test dataset after 10k, 100k and 1M words have been selected. From the results, we can see that the reduced-struct-freq method attains the highest 1-gram coverage, efficiently covering unknown words. On the other hand, it is clear that methods selecting longer phrases have an advantage for 4-gram coverage, and we see the highest 4-gram coverage in the sent-by-4gram-freq method.

6 Manual Translation Experiment

6.1 Experimental Set-Up

To confirm that the results from the simulation in the previous section carry over to actual translators, we further performed experiments in which professional translators translated the selected segments. This also allowed us to examine the actual amount of time required to perform translation, and how confident the translators were in their translations.

We designed a web user interface as shown in Figure 4, and outsourced to an external organization

Lang Pair	Selection Method	All Selected Phrases			First 10k Words	
		#Phrases	#Words	Average Phrase Length	#Phrases	Average Phrase Length
En-Fr	sent-by-4gram-freq	10.6M	269M	25.4	310	32.1
	4gram-freq	40.1M	134M	3.34	3.62k	2.76
	maxsubst-freq	62.4M	331M	5.30	2.39k	4.17
	reduced-maxsubst-freq	45.9M	246M	5.36	2.95k	3.39
	struct-freq	14.1M	94.2M	6.68	4.01k	2.49
	reduced-struct-freq	7.33M	41.3M	5.63	4.55k	2.20
En-Ja	sent-by-4gram-freq	1.28M	33.6M	26.3	560	17.8
	4gram-freq	8.48M	26.0M	3.07	4.70k	2.13
	maxsubst-freq	7.29M	25.8M	3.54	4.51k	2.22
	reduced-maxsubst-freq	6.06M	21.7M	3.58	4.76k	2.10
	struct-freq	1.45M	4.85M	3.34	6.64k	1.51
	reduced-struct-freq	1.10M	3.33M	3.03	6.73k	1.49

Table 2: Number of phrases and average words/phrase in each method

Lang Pair	Selection Method	1-gram / 4-gram Coverage [%]			
		No Addition	10k Words	100k Words	1M Words
En-Fr	sent-rand	92.72 / 10.60	92.93 / 10.60	93.73 / 10.71	95.94 / 11.30
	4gram-rand		92.95 / 10.60	93.99 / 10.60	96.42 / 10.64
	sent-by-4gram-freq		92.95 / 10.60	93.96 / **10.72**	96.25 / **11.55**
	4gram-freq		92.92 / 10.60	94.46 / 10.66	96.60 / 11.16
	maxsubst-freq		92.79 / 10.60	93.61 / 10.62	95.99 / 10.92
	reduced-maxsubst-freq		92.92 / 10.60	94.38 / 10.66	96.55 / 11.13
	struct-freq		93.63 / 10.60	96.15 / 10.65	97.84 / 11.28
	reduced-struct-freq		**94.02** / 10.60	**96.38** / 10.69	**98.00** / 11.38
En-Ja	sent-rand	94.36 / 5.38	94.81 / 5.63	95.99 / 6.59	97.54 / 10.06
	4gram-rand		94.80 / 5.38	96.10 / 5.46	97.67 / 5.98
	sent-by-4gram-freq		95.10 / 5.84	96.28 / **7.23**	97.64 / **11.39**
	4gram-freq		95.64 / 5.97	96.87 / 7.14	97.97 / 10.43
	maxsubst-freq		95.59 / 5.96	96.83 / 7.07	97.91 / 10.20
	reduced-maxsubst-freq		95.73 / **6.00**	96.97 / 7.19	98.00/10.57
	struct-freq		96.60 / 5.44	97.80 / 5.79	98.58 / 7.02
	reduced-struct-freq		**96.64** / 5.44	**97.84** / 5.80	**98.61** / 7.14

Table 3: Effect on coverage in each selection method (rounded off to the second decimal place). Bold face indicates the highest coverage for each number of additional words.

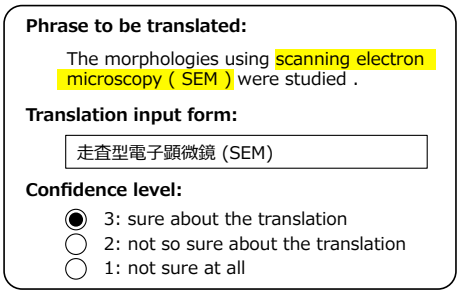


Figure 4: Example of the human translation interface

that had three professional translators translate the shown phrases. As is standard when hiring translators, we paid a fixed price per word translated for all of the methods. Because showing only the candidate phrase out of context could cause difficulty in translation, we follow Bloodgood and Callison-Burch (2010) in showing a sentence including the selected phrase,[7] highlighting the phrase, and requesting that the translator translate the highlighted part. We also requested that every worker select from 3 levels indicating how confident they were of their translation. In the background, the time required to complete the translation is measured from when the new phrase is shown until when the translation is submitted.

The methods selected for comparative evaluation are sentence selection based on 4-gram frequency (sent-by-4gram-freq) and phrase selection based on 4-gram frequency (4gram-freq) as baseline methods, and the phrase selection based on both parse trees and semi-maximality (reduced-struct-freq) as

[7] Specifically, we selected the shortest sentence including the phrase in the source corpus.

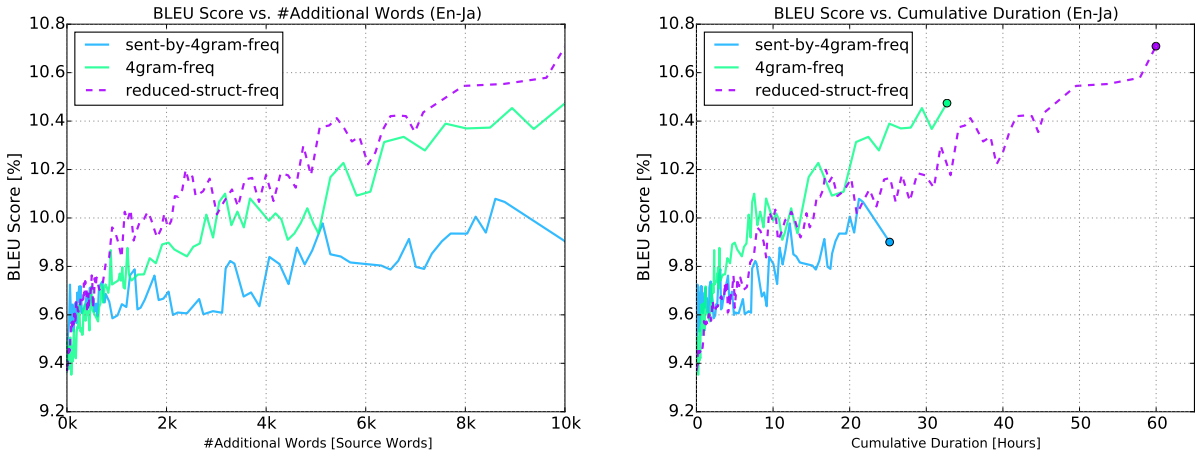


Figure 5: Transition of BLEU score vs. additional source words (left) and vs. cumulative working duration (right)

the proposed method. For each method we collected translations of 10k source words, alternating between segments selected by each method to prevent bias.

We used the same dataset as the English-Japanese translation task and the same tools in the simulation experiment (Section 5). However, for training target language models, we interpolated one trained with the base data and a second trained with collected data by using SRILM (Stolcke, 2002) because the hand-made data set was too small to train a full language model using only this data. We tuned the interpolation coefficient such that it maximizes the perplexity for the tuning dataset.

6.2 Results and Discussion

Efficiency results: Figure 5 shows the evaluation scores of SMT systems trained using varying amounts of collected phrases. In the left graph, we see the proposed method based on parse trees and phrase semi-maximality rapidly improves BLEU score, and requires fewer additional words than the conventional methods. Because the cost paid for translation often is decided by the number of words, this indicates that the proposed method has better cost performance in these situations. The right graph shows improvement by the amount of translation time. These results here are different, showing the 4-gram-freq baseline slightly superior. As discussed in Table 3, the methods based on parse trees select more uncovered 1-grams, namely unknown words, and specifically the proposed method selected more technical terms that took a longer time to translate.

Selection Methods	Total Working Time [Hours]	Average Confidence Level (3 Levels)
sent-by-4gram-freq	25.22	2.689
4gram-freq	32.70	2.601
reduced-struct-freq	59.97	**2.771**

Table 4: Total working time and average confidence level

Working time and confidence: We show the total time to collect the translations of 10k source words and average confidence level for each method in Table 4. The total working time for the proposed method is nearly double that of other methods, as seen in the right graph of Figure 5. On the other hand, the segments selected by the proposed method were given the highest confidence level, receiving the maximum value of 3 for about 79% of phrase pairs, indicating that the generated parallel data is of high quality. To some extent, this corroborates our hypothesis that the more syntactic phrases selected by the proposed method are easier to translate.

We can also examine the tendency of working time for segments of different lengths in Table 5. Interestingly, single words consistently have a longer average translation time than phrases of length 2-4, likely because they tend to be technical terms that require looking up in a dictionary. We show the average confidence levels corresponding to phrase length in Table 6. The confidence level of single words in the proposed method is lower than in the baseline method, likely because the baseline selected a smaller amount of single words, and those se-

Selection Method	Average Working Time [Seconds]				
	1 Word	2 Word Phrase	3 Word Phrase	4 Word Phrase	5+ Word Phrase
sent-by-4gram-freq	-	-	-	-	160.64
4gram-freq	30.14	24.76	21.77	21.12	-
reduced-struct-freq	35.61	25.23	21.72	28.13	22.82

Table 5: Average working time of manual translation corresponding to phrase length

Selection Method	Average Confidence Level (3 Levels)				
	1 Word	2 Word Phrase	3 Word Phrase	4 Word Phrase	5+ Word Phrase
sent-by-4gram-freq	-	-	-	-	2.689
4gram-freq	2.885	2.585	2.422	2.300	-
reduced-struct-freq	2.802	2.796	2.778	2.708	2.737

Table 6: Average confidence level of manual translation corresponding to phrase length

lected were less likely to be technical terms. On the other hand, we can confirm that the confidence level for longer phrases in the baseline method decreases drastically, while it is stably high in our method, confirming the effectiveness of selecting syntactically coherent phrases.

Translation accuracy by confidence level: Finally, we show the accuracy of the SMT system trained by all the collected data in each method in Table 7. To utilize the confidence level annotation, we tested SMT systems trained by phrase pairs with confidence levels higher than 2 or 3. From the results, the accuracy of every method is improved when phrases pairs with confidence level 1 were filtered out. In contrast, the accuracy is conversely degraded if we use only phrase pairs with confidence level 3. The translation accuracy of 9.37% BLEU with the base SMT system without additional data became 10.72% after adding phrase pairs having confidence level 2 or higher, allowing for a relatively large gain of 1.35 BLEU points.

7 Conclusion and Future Work

In this paper, we proposed a new method for active learning in machine translation that selects syntactic, non-redundant phrases using parse trees and semi-maximal phrases. We first performed simulation experiments and obtained improvements in translation accuracy with fewer additional words. Further manual translation experiments also demonstrated that our method allows for greater improvements in accuracy and translator confidence.

Selection Methods	BLEU Score [%]		
	Confidence 1+ (All)	Confidence 2+	Confidence 3
sent-by-4gram-freq	9.88	9.92	9.85
4gram-freq	10.48	10.54	10.36
reduced-struct-freq	10.70	**10.72**	10.67

Table 7: BLEU score when training on phrases with a certain confidence level

However, there are still a number of avenues for improvement. Particularly, as the proposed method selected segments that took more time to translate due to technical terms, the combination with methods to harvest unknown words (Daumé III and Jagarlamudi, 2011) or optimize the selected segments based on the time required (Sperber et al., 2014) is potentially useful. In addition, softer syntactic constraints that allow annotation of phrases with variables (Chiang, 2007) such as "one of the preceding X" are another interesting avenue of future work.

Acknowledgments

The authors thank anonymous reviewers for helpful suggestions. This research was supported by ATR-Trek Co. Ltd. The manual translation work was supported by BAOBAB Inc.

References

Sankaranarayanan Ananthakrishnan, Rohit Prasad, David Stallard, and Prem Natarajan. 2010. A Semi-Supervised Batch-Mode Active Learning Strategy for Improved Statistical Machine Translation. In *Proc. CoNLL*, pages 126–134, July.

Michael Bloodgood and Chris Callison-Burch. 2010. Bucking the Trend: Large-Scale Cost-Focused Active Learning for Statistical Machine Translation. In *Proc. ACL*, pages 854–864, July.

Peter F. Brown, Vincent J.Della Pietra, Stephen A. Della Pietra, and Robert L. Mercer. 1993. The Mathematics of Statistical Machine Translation: Parameter Estimation. *Computational Linguistics*, 19:263–312.

David Chiang. 2007. Hierarchical phrase-based translation. 33(2):201–228.

Hal Daumé III and Jagadeesh Jagarlamudi. 2011. Domain adaptation for machine translation by mining unseen words. In *Proc. ACL*, pages 407–412.

Matthias Eck, Stephan Vogel, and Alex Waibel. 2005. Low Cost Portability for Statistical Machine Translation based in N-gram Frequency and TF-IDF. In *Proc. IWSLT*, pages 61–67.

Ulrich Germann. 2014. Dynamic phrase tables for machine translation in an interactive post-editing scenario. In *Proc. AMTA 2014 Workshop on Interactive and Adaptive Machine Translation*, pages 20–31.

Jesús González-Rubio, Daniel Ortiz-Martínez, and Francisco Casacuberta. 2012. Active learning for interactive machine translation. In *Proc. EACL*, pages 245–254, April.

Spence Green, Sida I. Wang, Jason Chuang, Jeffrey Heer, Sebastian Schuster, and Christopher D. Manning. 2014. Human Effort and Machine Learnability in Computer Aided Translation. In *Proc. EMNLP*, pages 1225–1236, October.

Gholamreza Haffari and Anoop Sarkar. 2009. Active Learning for Multilingual Statistical Machine Translation. In *Proc. ACL*, pages 181–189, August.

Gholamreza Haffari, Maxim Roy, and Anoop Sarkar. 2009. Active Learning for Statistical Phrase-based Machine Translation. In *Proc. ACL*, pages 415–423, June.

Kenneth Heafield. 2011. KenLM: Faster and Smaller Language Model Queries. In *Proc, WMT*, July.

Toru Kasai, Gunho Lee, Hiroki Arimura, Setsuo Arikawa, and Kunsoo Park. 2001. Linear-Time Longest-Common-Prefix Computation in Suffix Arrays and Its Applications. In *Proc. CPM*, pages 181–192.

Phillip Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical Phrase-Based Translation. In *Proc. NAACL*, pages 48–54.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. pages 177–180.

Graham Neubig, Yosuke Nakata, and Shinsuke Mori. 2011. Pointwise Prediction for Robust, Adaptable Japanese Morphological Analysis. In *Proc. ACL*, pages 529–533.

Franz Josef Och and Hermann Ney. 2003. A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*, 29(1):19–51.

Franz Josef Och. 2003. Minimum Error Rate Training in Statistical Machine Translation. In *Proc. ACL*, pages 160–167.

Yusuke Oda, Graham Neubig, Sakriani Sakti, Tomoki Toda, and Satoshi Nakamura. 2015. Ckylark: A More Robust PCFG-LA Parser. In *Proc. NAACL*, pages 41–45, June.

Daisuke Okanohara and Jun'ichi Tsujii. 2009. Text Categorization with All Substring Features. In *Proc. SDM*, pages 838–846.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. In *Proc. ACL*, pages 311–318, July.

Philip Resnik and Noah A Smith. 2003. The web as a parallel corpus. *Computational Linguistics*, 29(3):349–380.

Burr Settles and Mark Craven. 2008. An Analysis of Active Learning Strategies for Sequence Labeling Tasks. In *Proc. EMNLP*, pages 1070–1079, October.

Matthias Sperber, Mirjam Simantzik, Graham Neubig, Satoshi Nakamura, and Alex Waibel. 2014. Segmentation for Efficient Supervised Language Annotation with an Explicit Cost-Utility Tradeoff. *TACL*, 2:169–180.

Andreas Stolcke. 2002. SRILM - an extensible language modeling toolkit. In *Proc. ICSLP*, pages 901–904.

Jörg Tiedemann. 2009. News from OPUS-A collection of multilingual parallel corpora with tools and interfaces. In *Proc. RANLP*, volume 5, pages 237–248.

Katrin Tomanek and Udo Hahn. 2009. Semi-Supervised Active Learning for Sequence Labeling. In *Proc. ACL*, pages 1039–1047, August.

Marco Turchi, Tijl De Bie, and Nello Cristianini. 2008. Learning performance of a machine translation system: a statistical and computational analysis. In *Proc. WMT*, pages 35–43, June.

Katharina Wäschle and Stefan Riezler. 2012. Analyzing Parallelism and Domain Similarities in the MAREC Patent Corpus. *Multidisciplinary Information Retrieval*, pages 12–27.

Omar F Zaidan and Chris Callison-Burch. 2011. Crowdsourcing translation: Professional quality from non-professionals. In *Proc. ACL*, pages 1220–1229.