# Statistical Machine Translation without a Source-side Parallel Corpus Using Word Lattice and Phrase Extension

**Takanori Kusumoto, Tomoyosi Akiba**

Toyohashi University of Technology
Toyohashi, Aichi, Japan
kusumoto@cl.ics.tut.ac.jp, akiba@cl.ics.tut.ac.jp

### Abstract

Statistical machine translation (SMT) requires a parallel corpus between the source and target languages. Although a pivot-translation approach can be applied to a language pair that does not have a parallel corpus directly between them, it requires both source–pivot and pivot–target parallel corpora. We propose a novel approach to apply SMT to a resource-limited source language that has no parallel corpus but has only a word dictionary for the pivot language. The problems with dictionary-based translations lie in their ambiguity and incompleteness. The proposed method uses a word lattice representation of the pivot-language candidates and word lattice decoding to deal with the ambiguity; the lattice expansion is accomplished by using a pivot–target phrase translation table to compensate for the incompleteness. Our experimental evaluation showed that this approach is promising for applying SMT, even when a source-side parallel corpus is lacking.

**Keywords:** statistical machine translation, resource-scarce language, word lattice

## 1. Introduction

Statistical machine translation (SMT) systems require a large parallel corpus between the source and target languages to produce good translations. However, constructing a parallel corpus of a specific language pair is such an enormous task that it prevents the application of SMT to a broad range of languages. Using a pivot language $P$, one can apply SMT from a new source language $S$ to any target language $T$ that already has a parallel corpus between $T$ and $P$ by using the corpus between $S$ and $P$ instead of constructing a corpus between $S$ and $T$ (Utiyama and Isahara, 2007). A major language, such as English, sometimes can be used as such a pivot language.

However, for resource-limited languages that do not have a parallel corpus to any major language, this pivot-language approach cannot be applied.

The proposed method uses a Vietnamese–English word dictionary instead of a Vietnamese–English parallel corpus to translate from the source language to the pivot language, and then applies an English–Japanese SMT trained by using the parallel corpus between the pivot and the target languages. The problems with dictionary-based translations lie in their ambiguity and incompleteness. The proposed method uses the word lattice representation of multiple English sentences to deal with the ambiguity. Lattice decoding can be applied to translate from the representations of the multiple sentences of the pivot language. It also uses lattice extensions to compensate for incompleteness. The dictionary-based translation is incomplete in word order and it misses words. We use an English–Japanese phrase translation table for examples of possible revisions, and add alternative candidates to the word lattice representation.

The rest of paper is organized as follows. In Section 2, we describe how to construct and extend word lattice. Experimental results is reported in Section 3. The paper is concluded in Section 4.

## 2. Related Work

Researchers have studied the approach for overcoming the lack of parallel corpus. For example, various resources such as comparable document pairs (Utiyama and Isahara, 2003), recordings of interpreter-mediated communication (Paulik and Waibel, 2010) have been utilized to extract parallel corpus. The work described in (Ananthakrishnan et al., 2010) made the best use of small amount of parallel corpus. Pivot language is often utlized to translate the language pair which does not have parallel corpus directly, but have source–pivot and pivot–target parallel corpus (Utiyama and Isahara, 2007). To the best of our knowledge, this work is the first attempt to deal with the scenario where parallel corpus is available only on the target side but is not on the source side.

## 3. Proposed Method

The pivot-translation approach uses two parallel corpora of source-to-pivot and pivot-to-target language pairs; the proposed method also uses the pivot-to-target parallel corpus but, instead of a source-to-pivot parallel corpus, it uses a source-to-pivot word dictionary. Therefore, the proposed method can be applied to the translation scenario where there is no parallel corpus on the source side but there is one on the target side. This approach is often appropriate for translating from a resource-limited language to a moderately resourced language. In this paper, we focus on translating from Vietnamese to Japanese, as one such language pair.

As expected, the dictionary-based translation itself performs poorly. The major problems with the dictionary translation are its ambiguity and incompleteness. Because a dictionary entry has multiple translation candidates, a mechanism to disambiguate them, i.e., select one of them, is required. In addition, the word dictionary can only translate each word of the source language into the corresponding word of the pivot language and so it leaves the word

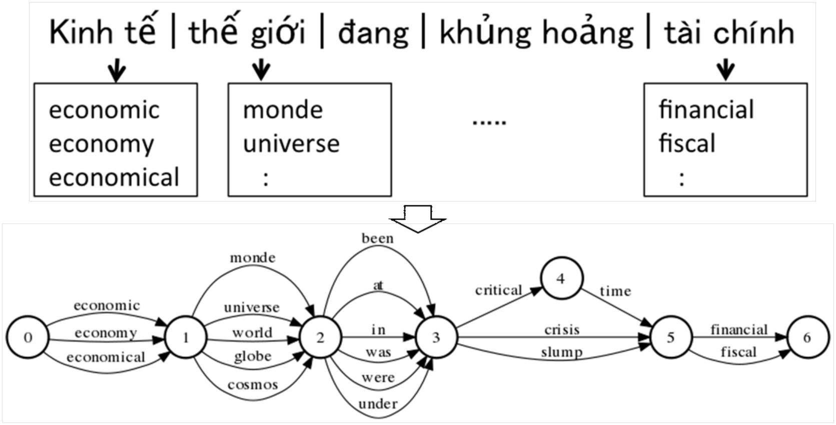

Figure 1: An example of conversion from a Vietnamese sentence to an English word lattice.

sequence the same as that in the original source sentence. Therefore, a mechanism to reorder the translated word sequence into one that matches the word order of the pivot language is also required. Moreover, because certain words often cannot be translated appropriately by the word dictionary, a mechanism to complement such words is needed.

Without parallel corpora, one must: (1) select the appropriate word from the dictionary without translation probabilities; (2) reorder the words so as to form an English sentence that reflects the meaning of the input sentence; and (3) append lacking words as necessary.

We address these problems by introducing the word lattice representation of the translation candidates of the pivot language. For the first problem, all word translation candidates obtained from the word dictionary can be represented by the lattice. For the second and third problems, we extend the lattice by adding new paths that represent the alternative candidates for the word sequence; they are obtained from the pivot-language side of the phrase translation table used for the SMT between the pivot and the target languages. The idea behind the proposed method is that each phrase in the phrase translation table can be seen as an example of a word sequence in correct order and that such phrases are useful for the back-end pivot-to-target SMT, because they actually appear in the phrase translation table of the SMT and they have a high probability of being selected as the translation result.

### 3.1. Selection from Multiple Word Translation Candidates

An intuitive way to choose the most likely combination of translations is to try to translate all combinations and select the most likely one. Using word lattice representation and word lattice decoding, SMT can accomplish this task.

A word lattice can represent multiple word sequences. For each word in the source language, its multiple translation candidates are obtained from the word dictionary and form parallel arcs, or paths when its translation is a phrase, in the lattice representation. The word lattice obtained has a sausage-like style, as shown in Figure 1.

Formally, the word lattice is constructed in the following steps. Here a directed edge from a vertex $u$ to a vertex $v$ with a label $l$ is denoted as a tuple $(u, v, l)$.

**Input:** A segmented input sequence $f = f_1, f_2, \cdots, f_n$.

**Output:** A directed acyclic graph $\Sigma = (V, E)$.

```
1. Initialize $V = \{v_0\}, E = \{\}$.
2. For $i = 1$ to $n$:
    (a) Create a vertex $v_i$ and $V \leftarrow V \cup \{v_i\}$.
    (b) Obtain the set of translation candidates $S$ from $f_i$ by using the word translation dictionary. If $S = \{\}$, use the original token $f_i$ as its translation (i.e., $S = \{f_i\}$).
    (c) For each $s \in S$:
        • If $s$ is a word, create an edge $(v_{i-1}, v_i, s)$ and add it to $E$.
        • If $s$ is a phrase of $k$ words $w_1, w_2, \cdots w_k$, create $k - 1$ new vertices $v'_1 \cdots v'_{k-1}$ and $k$ edges $(v_{i-1}, v'_1, w_1)$, $(v'_1, v'_2, w_2)$, $\cdots$, $(v'_{k-1}, v_i, w_k)$. Add these vertices and edges to $V$ and $E$, respectively.
```

Note that, for the segmentation of the input step, we can use a word segmentation tool of the source language, or we can simply segment the sequence by using white space as the delimiters; we can even use the word translation dictionary itself for the segmentation by finding the longest matching entry and segment it at the position.

### 3.2. Word Reordering

The sentence that can be derived from the word lattice we made in Section 3.1. may not follow the English word order because it still follows the word order of the source language (Vietnamese). To correct this, we append alternative path candidates to the lattice. Each candidate is created from an $n$-gram path in the original lattice by reordering its word sequence.

We utilize the English phrase table used in the following lattice decoding process for examples of English $n$-grams that have correct word order. We take every $n$-gram path that conveys an $n$-word sequence from the word lattice and

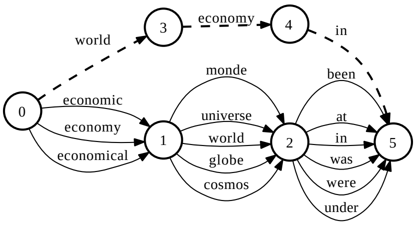

Figure 2: An example of lattice extension. The English phrase "world economy in" is appended to the lattice to change the word order.

compare it with phrases of the same length on the English side of the phrase table. If some permutation of the $n$-word sequence is identical to an English phrase in the table, a new path corresponding to the phrase is created and inserted into the same place in the lattice.
Formally, this process is described as follows.

```
Input: A word lattice $\Sigma = (V, E)$ and a phrase table $T$.

Output: An extended word lattice $\Sigma = (V', E')$.

  1. Initialize $V' = V, E' = E$.

  2. For all $v_i \in V$:

     (a) Collect set $P$ of all the $n$-gram paths $p = ((v_i, v_{i+1}, w_1), (v_{i+1}, v_{i+2}, w_2), \cdots, (v_{i_{+n-1}}, v_{i_{+n}}, w_n))$ starting from $v_i$.
     (b) For each $p \in P$, if one of the permutations of $w_1, w_2, \cdots, w_n$ is identical to a phrase $x_1, x_2, \cdots, x_n$ in $T$:
         i. Create new vertices $V'' = \{v'_1, v'_2 \cdots, v'_{n-1}\}$ and edges $E'' = \{(v_i, v'_1, x_1)$, $(v'_1, v'_2, x_2)$, $\cdots$, $(v'_{n-1}, v_{i_{+n}}, x_n)\}$.
         ii. $V' \leftarrow V' \cup V''$
             $E' \leftarrow E' \cup E''$.
```

Figure 2 shows an example of word lattice extension. Note that the new $n$-gram paths introduced using the process described above are promising candidates for selection as parts of the target sentence in the final translation, because each of them is an English phrase included in the phrase table used by the following decoding process.

### 3.3. Word Completion

Every word in the word lattice described in Section 3.1. is produced from an existing word in an input sentence. However, some English words in the desired translation may have no relation to any word in the input sentence and therefore they cannot be produced. For example, the English article "the" often cannot be derived from any Vietnamese words in the input sentence. To consider these words, we add snippets that contain such "spontaneous" words to the lattice. To deal with this problem, we also apply another lattice extension method to complement these additional words.
Again, we use English phrases in the phrase table as the reference for English. Word lattice extension for word completion is done by the following steps.

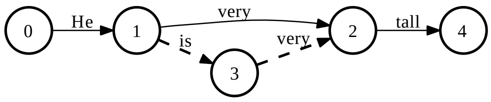

Figure 3: An example of lattice extension. An English fragment "He is very" is appended to the lattice to insert the missing word "is".

```
Input: A word lattice $\Sigma = (V, E)$, a phrase table $T$ and a list of spontaneous words $L$.

Output: An extended word lattice $\Sigma = (V', E')$.

  1. Initialize $V' = V, E' = E$.

  2. For all $v_i \in V$:

     (a) Collect the set $P$ of all the $n$-gram paths $p = ((v_i, v_{i+1}, w_1), (v_{i+1}, v_{i+2}, w_2), \cdots, (v_{i_{+n-1}}, v_{i_{+n}}, w_n))$ starting from $v_i$.
     (b) For each $p \in P$, if a word sequence $w_1, w_2, \cdots w_j, y, w_{j+1}, \cdots w_n$ is identical to a phrase $x_1, x_2, \cdots x_{n+1}$ in $T$ for some $y \in L$ and $j(1 \leq j \leq m)$:
         i. Create new vertices $V'' = \{v'_1, v'_2, \cdots v'_n\}$ and edges $E'' = \{(v_i, v'_1, x_1)$, $(v'_1, v'_2, x_2)$, $\cdots$, $(v'_{n-1}, v'_n, x_n)$, $(v'_n, v_{i+n}, x_{n+1})\}$.
         ii. $V' \leftarrow V' \cup V''$
             $E' \leftarrow E' \cup E''$.
```

An example of word completion is shown in Figure 3.

## 4. Experiment

The proposed Vietnamese–Japanese translation system requires a Vietnamese–English dictionary and an English–Japanese word lattice decoder (Dyer et al., 2008).
In the experiment, all the English words in the Vietnamese–English dictionary and the English–Japanese parallel corpus were lemmatized in the preprocessing.

### 4.1. Word Dictionary

The Vietnamese–English dictionary and English–Vietnamese dictionary were taken from *Ho Ngoc Duc's Free Vietnamese Dictionary Project* (`http://www.informatik.uni-leipzig.de/~duc/Dict/index.html`). We reversed the Vietnamese–English dictionary to make another English–Vietnamese dictionary and incorporated it into the original one to improve the vocabulary coverage.
The resulting dictionary has 147k indexing terms (Vietnamese) and a total of 239k translation terms (English).

### 4.2. SMT System

We employed the SMT system Moses (Koehn et al., 2007) as our word lattice decoder. The English–Japanese parallel corpus used in this work was extracted from Japanese newspaper articles (Utiyama and Isahara, 2003). We divided the parallel corpus into training, development, and

| lattice | nodes | edges | BLEU (lattice) | BLEU (1-best) |
|---|---|---|---|---|
| **RANDOM** | 2693 | 6052 | 0.0360 | 0.0360 |
| **VANILLA** | 5393 | 20384 | 0.0623 | 0.0583 |
| **REORDER** | 6641 | 22638 | 0.0625 | 0.0580 |
| **COMP** | 8130 | 28711 | 0.0637 | 0.0586 |
| **BOTH** | 9375 | 31030 | 0.0662 | 0.0582 |

Table 1: Statistics for the lattice and BLEU scores.

test sets, with respective sizes of 150k, 1000, and 98 sentences. The Japanese monolingual corpus used for training the language model consisted of 4000k sentences taken from Japanese newspapers.

We trained the translation model (the phrase table) using the Moses training script with default parameters. The 5-gram language model was trained using the IRSTLM toolkit (Federico and Cettolo, 2007).

### 4.3. Results

We made the word lattice from the test set by using the algorithm described in Section 3.1. (**VANILLA**).

It was then extended by using the word reordering algorithm described in Section 3.2. (**REORDER**), the word completion algorithm in Section 3.3. (**COMP**), and both of them at the same time (**BOTH**). In addition, we selected words at random from the word dictionary and arranged them in the original order (**RANDOM**) to evaluate the ability of disambiguation of the word lattice.

These five kinds of lattice were fed into the same lattice decoder, and translation results were obtained for each. We also took 1-best paths from these lattices according to the language model of the pivot language and decoded them with the same decoder. We compared these results by using the BLEU score. Table 4.2. shows the results and lattice size (total nodes and edges).

These results indicate that it is possible to apply SMT even when there is no source-side parallel corpus. As seen from the comparison of **RANDOM** and **VANILLA**, the lattice representation can improve translation quality significantly. The lattice representation also outperforms the single candidate (1-best). Furthermore, extensions of alternative paths such as word reordering and word complements can select more appropriate word from dictionary and improve the BLEU score. We obtained the maximum improvement on the BLEU score when both methods were applied to a lattice; this indicates that the improvement induced by the lattice extensions can be complementary.

## 5. Conclusion

In this work, we have conducted Vietnamese–Japanese SMT in the scenario where there is no parallel corpus on the source side but there is one on the target side. We used technique of lattice decoding to select appropriate translation terms from word dictionary.

We also employed English phrases used in English-Japanese SMT as an example of word sequences to correct the word order and to complement words in the dictionary-based translation.

Currently, the constructed word lattice is sizable, so that it takes considerable times to decode it and seems to prevent the decoder to select good path from it. We are planning to reduce the size by pruning the unlikely edges from the lattice in order to improve the decoding performance.

## 6. References

Sankaranarayanan Ananthakrishnan, Rohit Prasad, and Prem Natarajan. 2010. Phrase alignment confidence for statistical machine translation. In *INTERSPEECH*, pages 2878–2881. ISCA.

Christopher Dyer, Smaranda Muresan, and Philip Resnik. 2008. Generalizing word lattice translation. In *Proceedings of ACL-08: HLT*, pages 1012–1020, Columbus, Ohio, June. Association for Computational Linguistics.

Marcello Federico and Mauro Cettolo. 2007. Efficient handling of n-gram language models for statistical machine translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 88–95, Prague, Czech Republic, June. Association for Computational Linguistics.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, ACL '07, pages 177–180, Stroudsburg, PA, USA. Association for Computational Linguistics.

Matthias Paulik and Alex Waibel. 2010. Rapid development of speech translation using consecutive interpretation. In *INTERSPEECH*, pages 2534–2537.

Masao Utiyama and Hitoshi Isahara. 2003. Reliable measures for aligning japanese-english news articles and sentences. In *Proceedings of the 41st Annual Meeting of the ACL*, pages 72–79, Sapporo, Japan, July. Association for Computational Linguistics.

Masao Utiyama and Hitoshi Isahara. 2007. A comparison of pivot methods for phrase-based statistical machine translation. In *HLT-NAACL*, pages 484–491.