

# Multilingual Translation via Annotated Hub Language

KANAYAMA Hiroshi      WATANABE Hideo

Tokyo Research Laboratory, IBM Japan, Ltd.  
1623-14 Shimo-tsuruma, Yamato-shi, Kanagawa 242-8502, Japan  
{kanayama,watanabe}@trl.ibm.com

## Abstract

This paper describes a framework for multilingual translation using existing translation engines. Our method allows translation between non-English languages through English as a “hub language”. This hub language method has two major problems: “information loss” and “error accumulation”. In order to address these problems, we represent the hub language using the Linguistic Annotation Language (LAL), which contains English syntactic information and source language information. We show the effectiveness of the annotation approach with a series of experiments.

## 1 Introduction

Due to the worldwide expansion of the Internet, multilingual machine translation systems are more in demand than ever before, but what have been intensively developed are only translation engines which translate English into another language or another language into English. Developing all translation engines including such as Spanish-to-Chinese or Japanese-to-Italian is extremely hard work since  $(n^2 - n)$  translation engines would have to be prepared to cover all pairs among  $n$  languages.

In this paper we describe an any-to-any translation system using annotated English as the “hub language”. The key feature of our method is to annotate the English sentences in order to solve the problems in the hub language approach. The annotation is represented by using the Linguistic Annotation Language (LAL) (Watanabe et al., 2002).

The hub language approach allows translation between non-English languages by making use of the

existing English-related translation engines as illustrated in Figure 1. This approach requires much less labor than designing and implementing all of the translation engines independently. Another advantageous point is that any enhancement of a translation engine can be shared by all of the translation systems which use the same engine. However, the naïve hub model illustrated in Figure 2 has two fundamental problems, “information loss” and “error accumulation”.

Figure 3 illustrates the concept of information loss. Suppose both of two expressions X1 and X2 in Language X are translated into E1 in English. In this case, the English-to-Y translation engine can produce only the translation Y1 for both X1 and X2, so the distinction in Language X is lost due to the lack of appropriate expressive power in English.

The most typical cases of information loss are caused by polysemous words in English. In Example (1), two Japanese sentences J1 and J2 are translated into the same English sentence E1. E1 is translated into not F2 but F1, thus the translation of J2 into French fails because of information loss in the process of Japanese-to-English translation.

- (1) J1 *Kare ha ginkou ni itta.*  
          ‘He went to the (financial) bank.’  
      J2 *Kare ha teibou ni itta.*  
          ‘He went to the (river) bank.’  
      E1 He went to the bank.  
  
      F1 Il est allé à la banque.  
          ‘He went to the (financial) bank.’  
      F2 Il est allé à la digue.  
          ‘He went to the (river) bank.’

Another type of information loss is caused by grammatical forms which English do not have. In Example (2), both German sentences G1 and G2 are translated into E1, which should be translated into

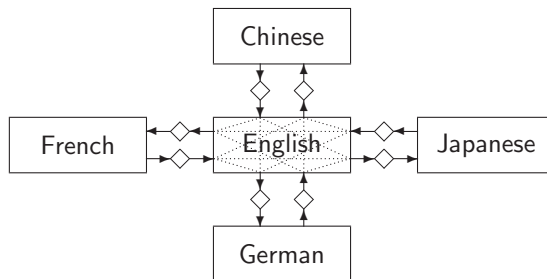


Figure 1: The hub language model for multilingual translation. ◇ denotes an existing translation engine.

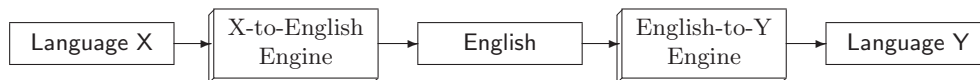


Figure 2: A translation system from Language X to Language Y by the naïve hub model using English as the hub language.

Japanese as J1 and J2, respectively. This problem occurs because English does not have the polite form.

- (2) G1 Wie geht es dir?
- G2 Wie geht es Ihnen?
- E1 How are you?
- J1 *Ogenki desu ka.*
- J2 *Genki kai.*

The second problem, error accumulation is illustrated in Figure 4. Errors in the engines such as parsing errors can occur in both the X-to-English engine and the English-to-Y engine, thus the translation precision of the whole system becomes lower than that of each engine.

Figure 5 illustrates the proposed method which uses annotation. The X-to-English translation engine attaches the source language information and structural information to the English sentence by annotation. The English-to-Y translation engine interprets them. The source language information makes it possible for the English-to-Y engine to consult the X-Y lexical dictionary, and the structural information such as parsing results, sentence segmentations, and parts-of-speech data prevents parsing errors in the English-to-Y engine.

Section 2 overviews LAL. Section 3 shows the design of our multilingual translation system using LAL, and the effectiveness is evaluated in Section 4. In Section 5, we discuss several approaches for multilingual translation.

## 2 Linguistic Annotation Language

Linguistic Annotation Language (Watanabe et al., 2002) is an XML-compliant tag set, and its XML namespace prefix is `lal`. Originally, LAL was designed for manual annotations which would assist several natural language processing applications by addressing several types of ambiguities. Our multilingual translation method adopts LAL because it has simplicity and generality. Mainly we use two tags, `<lal:s>` and `<lal:w>`, from the LAL specification are mainly used.

The tag `<lal:s>` delimits a sentence. This tag avoids the confusion of sentence boundaries which often happen when a word has a punctuation mark which doesn't signify the end of the sentence. In Example (3), `<lal:s>` prevents the sentence from being divided after 'Prof.'

- (3) `<lal:s>It is Prof. Smith who taught us English.</lal:s>`

The tag `<lal:w>` delimits a word and it can have several attributes. The part-of-speech of a word is specified by the attribute `pos`, and the dependency structure between two words is represented by the attributes `id` and `mod`. For instance, the annotated sentences (4a) and (4b) represent two possible parsing results of 'She saw a man with a telescope.'<sup>1</sup>

<sup>1</sup> In the examples of LAL-annotation in this paper, sentences are partially annotated for simplicity.

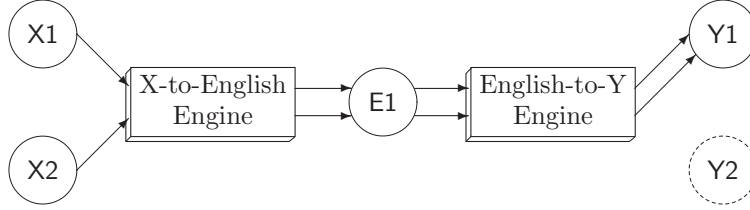


Figure 3: The concept of information loss. When two expressions in Language X are not distinguished in English, one of them may be translated into a wrong expression in Language Y.



Figure 4: The concept of error accumulation. The X-to-English translation engine may produce English sentences with some errors. Regardless of the existence of errors in the intermediate English, also the English-to-Y engine causes some errors, thus the quality of Language Y is worse than that of English.

- (4) a. `<lal:s>She`  
`<lal:w id="2">saw</lal:w> a man`  
`<lal:w id="5" mod="2">with</lal:w>`  
`a telescope.</lal:s>`  
 b. `<lal:s>She saw a`  
`<lal:w id="4">man</lal:w>`  
`<lal:w id="5" mod="4">with</lal:w>`  
`a telescope.</lal:s>`

The utilization of LAL and its extension in the proposed method is described in Section 3.

### 3 Multilingual Translation Using a Hub Language

This section describes the usage of LAL for multilingual translation. We add new attributes of `<lal:w>` to address the problem of information loss.

#### 3.1 Recovery from Information Loss

As described in Section 1, different expressions in the source language may be translated into a single expression in English as the hub language. Our method solves this problem of information loss by using LAL-annotation.

Information loss is often caused by English polysemous words that are differentiated in both the source language and the target language. To retrieve the lost information, we attach the lexicon in the source language to the corresponding English word as Example (5). Two attributes of `<lal:w>` are added here: one is `orig_lang` which denotes the source language. The other is `orig_lex` whose value is the lexicon in the source language.

- (5) `<lal:w orig_lang="ja"`  
`orig_lex="teibou">bank</lal:w>`

The annotation in Example (5) means the word ‘bank’ is derived from the Japanese word ‘*teibou*’ (river bank). When this word is translated into another language, this information in original language is referred to. In this case, English ‘bank’ can be translated into the French ‘digue’ using the Japanese-French bilingual dictionary, while ‘bank’ may be translated into French ‘banque’ without the annotation. Note that the Japanese-French dictionary need not have full coverage. Only words which can not be correctly translated via English are sufficient.

Our method does not use the pseudowords such as ‘bank1’ and ‘bank2’, but annotates the lexicon in the source language. This is because the translation engines should be developed independently except for a minimum set of specifications about the annotations.

The annotation of the source word can improve Chinese-to-Japanese (or reverse) translation drastically, even if direct bilingual dictionaries are not used. When proper nouns in Chinese or Japanese are translated into English as transliterations to alphabetic representations, most of the translated English words are regarded as unknown words by the English parser. Back transliteration into Chinese or Japanese is very difficult due to the ambiguity. But if the source word can be obtained from the attributes of `<lal:w>`, often we can get the correct translation because the ideographic characters are often used in common between Chinese and Japanese for proper nouns.

Also some grammatical information is lost in X-to-English translation. A typical example is the polite form: English has the only second-person pronoun ‘you’, while most of the other European languages

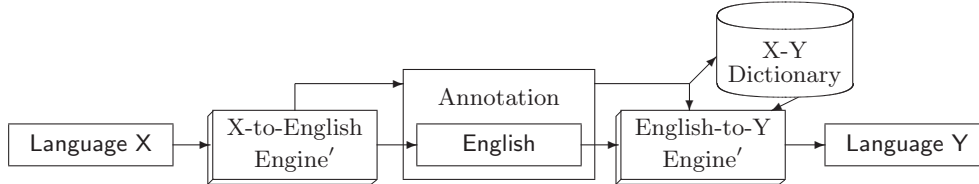


Figure 5: Hub model via annotated English. Here the X-to-English engine produces English sentences with annotation, and the English-to-Y engine interprets the annotation. The X-Y bilingual dictionary are available.

	Japanese	English	French	Correct French	LAL improves?
1	<i>karai</i>	hot	chaud	pimenté	Yes
2	<i>Chiyoda-ku</i>	Chiyoda-ku	ku, Chiyoda	Chiyoda-ku	Yes
3	<i>heya</i>	room	pièce	salle	No

Table 1: Examples of mistranslated words in French. The words in the column ‘Japanese’ were translated into the words in the column ‘English’. The column ‘French’ shows the translation results from the English words without LAL. They should be translated as the column ‘Correct French’. The rightmost column tells whether the annotation removes the error.

have informal and polite forms of ‘you’. In Japanese and Korean, the end of a sentence varies according to the politeness of the sentence. In our approach, an attribute `polite` is added to `<lal:w>` or `<lal:s>`. For example, German ‘Wie geht es Ihnen?’ is translated into an annotated English sentence in Example (6).

(6) `<lal:s orig_lang="de" polite="yes">`  
 How are you?`</lal:s>`

The politeness of the sentence helps the appropriate generation of other languages. The above sentences are translated into Japanese ‘*O genki desu ka?*’, and Spanish ‘¿Cómo está usted?’. When the value of `polite` is “no”, they are translated into ‘*Genki kai?*’ and ‘¿Cómo estás?’, respectively.

Such attributes should not be overused because too many attributes make the specification of LAL complex. We adopt the attribute `polite` because the differentiation of politeness is important and the lack of politeness is one of the major features of English.

### 3.2 Reduction of Error Accumulation

In addition to recovery from information loss, LAL-annotation works effectively to reduce parsing errors in the stage of English-to-Y translation. The X-to-English translation engine generates the English sentence with the syntactic information represented by `id`, `mod` and `pos` attributes of `<lal:w>`, and the parser of the English-to-Y engine follows the parsing structure. Thus our method does not suffer from accumulation of parsing errors while there are two parsing processes.

Moreover, the segmentation by the tags helps

English-to-Y translation. As shown in Example (3), `<lal:s>` avoids the wrong segmentation of sentences. When a multiple word such as ‘New Orleans’ is delimited by `<lal:w>`, it is securely regarded as a single word.

## 4 Evaluation

To evaluate how the annotation contributes to improve the multilingual translation via a hub language, we conducted experiments using the Japanese-to-English and English-to-French translation engines. As the test set, we used 214 Japanese sentences which the Japanese-to-English engine can translate into English correctly. The average length of the sentences is 9.4 words, and the test set covers several linguistic phenomena.

### 4.1 Replacement of Words by Japanese-to-French Dictionary

In the French translation results by the naïve hub method via English, we found 23 problematic word selections. A total of 17 out of the 23 cases are translated better if the corresponding Japanese word is attached to the English and a Japanese-French dictionary is available.

See the first example in Table 1. The Japanese word ‘*karai*’ (spicy) was translated into the French word ‘chaud’ via the English word ‘hot’, when LAL was not used. But the source language information can disambiguate ‘hot’ because Japanese ‘*karai*’ corresponds to not French ‘chaud’ but ‘pimenté’<sup>2</sup>.

<sup>2</sup> This problem does not occur if the Japanese-to-English engine outputs ‘spicy’ as the translation of

J	<i>Kare ha hashitte, pan wo taberu.</i> “He runs, and eats bread.”
E	He runs and eats bread.
F	Il court le pain et le mange. “He runs the bread and eats it.”
EL	<code>&lt;lal:w id="1" mod="3"&gt;He&lt;/lal:w&gt; &lt;lal:w id="2" mod="3"&gt;runs&lt;/lal:w&gt; &lt;lal:w id="3" mod="0"&gt;and&lt;/lal:w&gt; &lt;lal:w id="4" mod="3"&gt;eats&lt;/lal:w&gt; &lt;lal:w id="5" mod="4"&gt;bread&lt;/lal:w&gt;.</code>
F2	Il court et mange le pain. “He runs and eats the bread.”
J	<i>Watashi ha kare no tame no hon wo katta.</i> “I bought a book which is for him.”
E	I bought the book for him.
F	Je lui ai acheté le livre. “I bought him the book.”
EL	<code>I bought the &lt;lal:w id="4"&gt;book&lt;/lal:w&gt; &lt;lal:w id="5" mod="4"&gt;for&lt;/lal:w&gt; him.</code>
F2	J'ai acheté le livre pour lui. “I bought the book for him.”

Table 2: Example of reduction of parsing error by LAL. E is the translation result of J by the Japanese-to-English engine, and F is the translation result of E without annotation. EL contains LAL-annotation on the structural information of E generated by the Japanese-to-English engine, and F2 is the result when the annotated information is used in the English-to-French engine.

The second example is a proper noun which should be translated as it is<sup>3</sup>. Most of Japanese names of place, humans, and organizations required annotations, because multiple words were not recognized as single proper nouns by the English parser. The third example cannot be solved easily by the annotation, because the Japanese ‘*heya*’ (room) can mean both French ‘*pièce*’ and ‘*salle*’.

The third case is not an example of information loss, so the LAL annotation can avoid most of the information loss problems caused by word selection.

#### 4.2 English Parsing with Annotation

The test set was translated into French via English with LAL-annotation about parsing structure, and they were compared without LAL. As a result, 35 sentences were translated better due to the correct specification of attachments or part-of-speech information. A total of 14 sentences became worse, but they were caused by the mismatches between the output LAL of the Japanese-to-English engine and the expected input structures of the English-to-French engine, so the bad effects can be avoided by simple

<sup>3</sup>‘*karai*’, however, we do not want to develop X-to-English dictionaries with taking English-to-Y translation into consideration.

<sup>3</sup> English ‘Chiyoda Ward’ and French ‘Arrondissement de Chiyoda’ are also correct translations, but ‘Chiyoda-ku’ (Japanese transcription) is better for addressing.

modification of annotation functions in the engines.

Examples of sentences whose French translations were improved by LAL annotation are shown in Table 2. E is the translation result of J by the Japanese-to-English engine, and F is the translation result of E without annotation. EL contains LAL-annotation on the structural information of E generated by the Japanese-to-English engine, and F2 is the result when the annotated information is used in the English-to-French engine. The first example in Table 2 shows an erroneous French translation which can be easily improved by using annotated English. The second example is more interesting: E can be interpreted in two ways by high or low attachment of ‘from’, while original J means only the low-attachment interpretation in E. Thus the meaning of F is different from that of J, while F is syntactically correct. In this case the annotation disambiguates the English expression. This example can be regarded as another instance of information loss problem.

Note that the ratio of improvement does not indicate the frequency of parsing errors, because even if the parsing structure is not the correct one, translation can succeed as in Example (7). Regardless of the modifier of ‘with’ of the English sentence E, its French translation will be F. Therefore the experiment here directly clarified the effectiveness of annotation.

	Quality	Coverage	Cost
Direct Engines	Very High	High	Very High
Interlingua	High	Low	High
Naïve Hub	Low	High	Very Low
Annotated Hub	High	High	Low

Table 3: Features of approaches for multilingual translation. Note that it is better when ‘Cost’ is lower.

- (7) E She saw a man with a telescope.  
 F Elle a vu un homme avec un télescope.

## 5 Discussion

We compared four approaches for multilingual translation, from the viewpoint of quality, coverage and cost. Table 3 shows these features of the approaches.

As described in Section 1, developing direct translation engines is extremely costly. Development of an engine is too hard, moreover, the number of engines to be developed increases with the square of the number of languages to be covered.

Several methods of semantic representation for multilingual translation have been studied such as entity-oriented semantics (Tomita and Carbonell, 1986), KANT-interlingua (Lonsdale et al., 1994) or the Universal Networking Language (UNL) (Uchida and Zhu, 2001). In these interlingua approaches, however, it is difficult to obtain high coverage for handling real-world sentences, because the interlingua must be designed to represent the meanings of all languages.

The hub language method (also known as “pivot language” from long ago: originally by Leon Dostert (Reifler, 1954)) is much more robust than interlingua approaches because the intermediate data structure can be interpreted as natural language. However, the translation quality by the naïve hub method is low because of the problems of information loss and error accumulation.

The annotated hub language method proposed in this paper has the same coverage as the naïve hub method, and its quality is higher than the naïve hub method because the annotation solves the two problems. Though the quality cannot be at the same level as fully tuned direct translation engines, a system which can be realized at a low cost is very practical, because the human ability for translation between non-English languages tends to be limited, so the machine translation systems are accepted even though very high precision is not achieved.

## 6 Conclusion and Future Work

We designed a multilingual translation method using an annotated hub language. The annotation solves the two problems of information loss and error accumulation which are obstacles in the naïve combination of two translation engines. This method allows us to achieve multilingual translation at a very low cost, making use of existing translation engines.

The method of selecting words has room to improve. Our method assumes that the English-to-Y engine deterministically relies on the X-Y dictionary. Ideally, the knowledge of the English-to-Y engine and the X-Y dictionary should be integrated, so a method to do this without losing the independence of developing each engine is required.

## References

- [Lonsdale et al.1994] Deryle W. Lonsdale, Alexander M. Franz, and John R. R. Leavitt. 1994. Large-scale machine translation: An interlingua approach. In *Seventh International Conference on Industrial and Engineering Applications of Artificial Intelligence and Expert Systems*, pages 525–530.
- [Reifler1954] Erwin Reifler. 1954. The first conference on mechanical translation. *Mechanical Translation*, 1(2):23–32.
- [Tomita and Carbonell1986] Masaru Tomita and Jaime G. Carbonell. 1986. Another stride towards knowledge-based machine translation. In *Proc. of 11th COLING*, pages 633–638.
- [Uchida and Zhu2001] Hiroshi Uchida and Meiyong Zhu. 2001. The Universal Networking Language beyond machine translation. In *International Symposium on Language in Cyberspace*.
- [Watanabe et al.2002] Hideo Watanabe, Katashi Nagao, Michael McCord, and Arendse Bernth. 2002. An annotation system for enhancing quality of natural language processing. In *Proc. of 19th COLING*, pages 1303–1307.