Japanese-Hungarian Dictionary Generation using Ontology Resources

VARGA István, YOKOYAMA Shoichi

Yamagata University, Graduate School of Science and Engineering
4-3-16 Jonan, Yonezawa-shi Yamagata,
992-8510 Japan
dyn36150@dip.yz.yamagata-u.ac.jp, yokoyama@yz.yamagata-u.ac.jp

Abstract

In this paper we describe an automatized dictionary generation method that can be applied with most language pairs. As an example, we created a Japanese-Hungarian dictionary. Our approach is a pivot language based method, with English as the intermediate language. We use the Japanese-English and English-Hungarian dictionaries to create a set of translation candidates, for correct translation pair selection we use the English WordNet as a main resource. Our purpose is to achieve a high precision dictionary with a maximized recall. Evaluations showed an improved recall and precision when compared with two conventional methods, showing that ontology based selection can produce better results than dictionaries based methods.

Introduction

Computer-readable bilingual dictionaries are an essential and indispensable tool for numerous fields that deal with natural languages, such as machine translation, language educational CAI software development, and so on. However, in most cases the cost in time and manual labour in such a task is too high; therefore development of automatic methods is needed. In the case of less-frequent language pairs the task is especially difficult because bilingual linguists or reliable paper dictionaries of the chosen languages that can offer guidance during development are not always available.

Many dictionary generating methods were proposed as an alternative to the manual construction of dictionaries. Compared with manual construction, automated methods have the advantage of being able to create a dictionary in considerably less time; and in the same time, both language speaking personnel can also be omitted. However, the recall and precision of current automatically created dictionaries is considerably lower opposed to the manually edited ones. More precisely, many words cannot be translated with current methods and many entries contain noise in form of erroneous translations.

In the case of the Japanese-Hungarian language pair there is no reliable bidirectional paper dictionary and there are no direct digital resources either, although the need for them is increasing. As a result, in this paper we propose a method to automatically create a Japanese-Hungarian machine readable dictionary with the purpose of minimizing cost and provide a method that can be applicable to other language pairs as well.

In this paper first we analyze the problems of the previous methods, after that we present our proposal. We evaluate the dictionary generated with our method and we compare it with two existing ones. Finally we formulate our conclusions and plans for the future.

Conventional Methods and Problems

There are two main principles in automated dictionary generation: corpus based and pivot language based methods. Corpus based methods rely mainly on information extracted from large bilingual corpora, selecting among translation candidates based on *mutual information* (Brown et. al., 1998), *Dice-coefficient* (Kay & Röscheisen, 1993), *correspondence-table* (Brown, 1997), etc.

As an alternative to corpus based methods, a pivot language based method was introduced by Tanaka & Umemura (1994). Pivot language based methods rely on the idea that the lookup of a word in an uncommon language through a third, intermediated language can be automated. These methods presume the existence of at least one intermediate language between the source and the target languages. Based on this principle there are many different approaches, all of them select among the translation candidates based on relations between the source-intermediate and target-intermediate entries.

Tanaka & Umemura's method uses integrated bidirectional source-intermediate and intermediate-target dictionaries ("harmonized dictionaries") to create source-target translation candidates. Correct candidates are mainly selected by means of "inverse consultation", a method that relies on counting the number of intermediate language definitions of the source word, through which the target language definitions can be identified. Based on experiments on an English pivoted French-Japanese dictionary they concluded that their method is useful for revising and supplementing the vocabulary of existing dictionaries.

Shirai and Yamamoto (2001) also used English to design a Korean-Japanese dictionary. They compare the English definitions of Korean and Japanese entries, counting the word overlap between the definitions. They achieved 72% accuracy, but with only 36.5% recall.

Other refinements include the tentative of using multiple pivots (English and Chinese) in the case of Sino-Japanese words in a Korean-Japanese dictionary (Paik & Bond & Shirai, 2001). They achieved a very high precision, but the 12% recall was extremely low.

Sjöbergh (2005) presents an innovative approach for pivot language oriented dictionary generation. When generating his English intermediated Swedish-Japanese dictionary, each Japanese-to-English description is compared with each Swedish-to-English description. Scoring is based on word overlap, weighted with inverse document frequency, the best matches being selected as translation pairs. Sjöbergh reports a dictionary with high recall and good precision.

The main deficiencies of the conventional methods can be grouped as follows:

1. Translation pairs are selected mainly based on information retrieved from the dictionaries.

We can argue that most of the above mentioned pivot language based methods fail to deliver the desired precision and recall, because dictionaries only do not provide enough information about the languages itself. Whatever language pair we use, the meaning-ranges of most of the corresponding words are not identical, they only overlap at a certain extent, an uncertainty that is doubled by the intermediate language. Moreover, bilingual dictionaries do provide with bidirectional or unidirectional lexical relations between words across the languages, but they do not provide enough semantic information or full description of words. Even if the translation from the source and target languages is correctly transferred to the intermediate language, due to the inconsistent translations from the target and source languages to the intermediate language and lack of proper correspondences between the two definitions, the recall suffers. As an example pointed out in Table 1, the Hungarian *sok* and Japanese *沢山 (takusan)*, or Hungarian *ember* and Japanese *人 (hito)* have approximately identical meanings respectively, but the meaning-range and semantical structure implemented in the two dictionaries are different, therefore correct translation pair extraction is difficult only based on information from the two dictionaries. In this case conventional methods cannot identify the difference between totally different definitions resulted by unrelated concepts and differences in only nuances resulted by lexicographers describing the same concept, but with different words. Some intentions to use semantic information is expressed by Paik, Bond & Shirai (2001), but they only use synonymy information, and only as an auxiliary, not as a main translation pair selecting tool. On the other hand, from this point of view corpus based methods provide the extra information about the meaning ranges of the words at a certain extent, because their semantic behaviour can be observed from the context. However, the first major drawback of corpus based methods is the necessity of a large bilingual corpus, available only in a few language pairs.

Source word	Hungarian to English	Japanese to English
sok = 沢山 *(takusan)*	a good many, a great many, a lot of, a number of, any amount, any number, gob, lots of, many, might, much, numerous, power, scores, several, whacking, whacking-great	many, a lot, much
ember = 人 *(hito)*	bleeder, body, man, men, mortal, number, one, people, person, soul, walla, wallah	man, person, human being, mankind, people, character, personality, true man, man of talent, adult, other people, messenger, visitor

Table 1: Translation differences across different dictionaries

2. Most evaluation methods are not accurate

We believe that there are a number of problems in conventional dictionary evaluation. Traditional calculation of word entry recall in the source language does not reflect the true recall value of the dictionary, especially when comparison with other methods is discussed. It is well-known that the biggest difficulty in dictionary generation lies on the fact that ambiguous words are hard do disambiguate. However, most low frequency or very low frequency words are less ambiguous than entries with high frequency, therefore they are easier to recall and translate. Traditional recall is a simple count of translated entries versus total number of entries, the recall value of a less frequent word being equal with the recall value of a very frequent word. The weight of every word should be proportional with its frequency; therefore a missing word with a high frequency should have a bigger impact on the recall score than a missing word with a low frequency.

Improved Method

Since there is no large bilingual Japanese-Hungarian corpus, we implement a pivot-language based method. We chose English as an intermediate language, since large digital Japanese-English and Hungarian-English dictionaries are freely available. However, as a main difference from the related methods, we do not intend to achieve the correct translation pairs based on the two dictionaries, we mainly use them only to collect some of the data. Instead, we use ontology for the reasons already mentioned above.

Although we expect a higher precision from our method, we do not believe that the resulting dictionary will be error-free, manual labour will still be needed to correct the faulty results. A dictionary with low recall needs a more careful manual revision than a dictionary with high recall, because the human corrector needs to discover the word entries that were not recalled, the translation of them becoming entirely manual. On the other hand, with high recall we can argue that manual work will be mainly a procedure of erasing or replacing the noisy translation pairs. Thus our dictionary is built accordingly: we try to obtain good accuracy with the highest possible recall.

Lexical Resources

To generate the Japanese-Hungarian dictionary we use the following resources:

- edict: Japanese to English unidirectional dictionary created and maintained by Jim Breen (1995) that has a number of 197282 1-to-1 entries after cleaning. It is freely downloadable from the Internet: *http://www.csse.monash.edu.au/~jwb/j_edict.html*
- a Hungarian-English bidirectional dictionary created and maintained by Vonyó Attila that has a number of 189331 1-to-1 entries after cleaning. This dictionary is also freely downloadable from the Internet: *http://almos.vein.hu/~vonyoa/SZOTAR.HTM*
- WordNet 2.1: a large lexical database of English, in which nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms called *synsets*, each describing a distinct concept (Miller et. al., 1990). WordNet is also freely downloadable from the Internet: *http://wordnet.princeton.edu/*

The Hungarian-English dictionary does not contain part-of-speech information. Furthermore, part-of-speech is highly inconsistent between Hungarian and Japanese; therefore although this kind of information is available from the Japanese-English dictionary, we do not use it.

Proposed method

Our method consists of two steps: in step 1 we collect a number of translation candidates that we think will contain most of the desired translation pairs. In step 2 first we perform only a restricted number of lexical analyses on the translation candidates to extract the unambiguous pairs. Also in step 2 we score all translation candidates using semantic information from the ontology and we select the most appropriate candidates based on the scoring. Below is a detailed description of the method.

Step 1: translation candidate generation

In step 1 we first generate the translation candidates. While doing this, we are only concentrating on obtaining the highest possible recall. We look up every entry in turn from the Japanese-English dictionary, looking up also every English-Hungarian entry in which the English entry matches the Japanese translation definition result from the Japanese-English dictionary. We consider every Japanese-Hungarian word/expression pair as a translation candidate for the next step. For example, according to our Japanese-English dictionary the Japanese word *曖昧 (aimai)* has three translation into English: *vague*, *ambiguous* and *unclear*. The English translations in turn have a total of seven translations into Hungarian: *bizonytalan, halvány, határozatlan, homályos, tétova, félreérthető, kétértelmű*. Thus the Japanese *曖昧* and the seven Hungarian words become seven different translation candidates (Figure 1).

Understandably a large number of obviously erroneous pairs or pairs with too little semantic similarity will be also included this way. For example, the pair *曖昧 – halvány* is not a correct translation pair, although both can be translated into English as *vague*. While the Japanese word is closer to *vague* as *obscure, not clearly understood or expressed*, the Hungarian one has the same meaning with *vague* as *dim, lacking clarity or distinctness*. However, we do not intend to disambiguate the definitions in this step, the identification of homonyms and elimination of noisy translations will be performed in the next step.

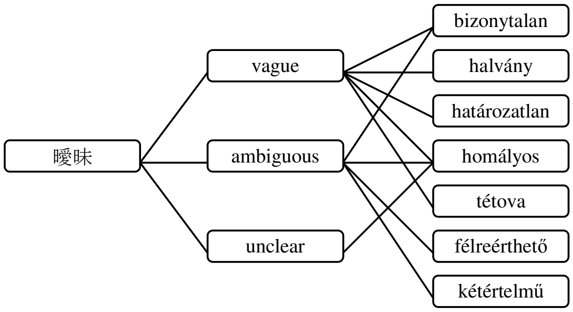


Figure 1: Translation candidate generation example

This way we identified all translation candidates that have at least one common word in their definitions. The direction of this operation is not relevant, it is not necessary to perform the same operation starting from the Hungarian entries too, it would result in exactly the same translation candidates.

With the method described above we accumulated a number of 436966 Japanese-Hungarian translation candidates.

Step 2: lexically unambiguous translation pair extraction and scoring

In step 2 we examine the translation candidates one by one, looking up the Japanese-English and Hungarian-English dictionaries, comparing the English descriptions. The translation candidates are basically scored by the correlation of the two English word sets.

First we perform a strictly lexical match based only on the dictionaries. Finally a thorough semantic analysis is performed based on information retrieved from the ontology.

1. Lexically unambiguous translation pair extraction

Some of the translation candidates have exactly the same English definitions; we consider these pairs as being correct by default. Also among the translation candidates we identified a number of Japanese entries that had only one Hungarian translation; and a number of Hungarian entries that had only one Japanese translation. Being the sole candidates for the given entries, we consider these pairs too as being correct.

37391 translation pairs were retrieved with this method; we call them *type A* translations.

During correct translation pair selection by means of semantic analysis it is our purpose to use as many semantic relations as possible. WordNet provides with information on a number of relations (Table 2).

Semantic information	**Part-of-speech**			
	Noun	Verb	Adjective	Adverb
sense classification	●	●	●	●
synonymy	●	●	●	●
antonymy	●	●	●	●
hypernymy/ hyponymy	●	●	×	×
troponymy	×	●	×	×
holonymy/ meronymy	●	×	×	×

Table 2: Semantic information from WordNet

From the relations described above, we consider four types of semantic information: sense classification, synonymy, antonymy and hypernymy/hyponymy. First we score each translation candidate by these information separately, as described below.

2. Sense classification

WordNet has detailed description on each word regarding its senses. Explanations as well as a number of synonyms are listed with most senses of each word. We use the synonymy information to try to disambiguate the senses in the intermediate English translation.

The scoring method is as follows: for a given Japanese-Hungarian translation candidate *(j, h)* we look up their translations into English from the respective dictionaries ($e_J=e_1(j), e_2(j), \dots e_n(j)$ and $e_H=e_1(h), e_2(h), \dots e_m(h)$, respectively). We select the English definitions that are common in the two definitions ($e_i(h,j)$) and we look up their respective senses ($sense(e_i(j)), sense(e_i(h))$) using WordNet. We identify the word's senses comparing each synonym in the WordNet's synonym description of the word in question with each word from the dictionary definition. As a result, we arrive at a certain set of senses from the Japanese-to-English definitions and a certain set of senses from the Hungarian-to-English definitions. We mark $score_{sense}(j, h)$ the maximum ratio of the identical and total identified sets of the common words. The higher the $score_{sense}(j,h)$, the probable that the candidate *(j,h)* is a valid translation.

$$score_{sense}(j,h) = \max_{e_i(j,h)\in e_J \cap e_H} \frac{count(sense(e_i(j)) \cap sense(e_i(h)))}{count(sense(e_i(j)) \cup sense(e_i(h)))} \quad (1)$$

As an example, there are 44 Hungarian translation candidates for the Japanese word *正解 (seikai: correct, right, correct interpretation)*. Among the 44 translations let's analyze *helyes (correct, right, legitimate, proper, appropriate, etc)* and *becsületes (fair, honest, honorable, honourable, just, right, trusty,* etc*)*. By common sense *helyes* should get a higher score then *becsületes*.

正解 and *helyes* have two common English translations, namely *right* and *correct*. *Right* has 13 senses according to WordNet, among them 4 where identified from the Japanese-to-English definition (#1, #3, #5, #10, all with *right*) and 5 from the Hungarian-to-English definition (#1, #3, #5, #6, #10, with *right* and *proper*). As a result, 4 senses are common, and one is different. Based on equation (1) $score'_{sense}$ *(正解, helyes) =0.8*, when scoring is done through the word *right*. Correct has 4 senses according to WordNet, all of them are recognized by both definitions through *right*, therefore the score through *correct* is $score'_{sense}$ *(正解, helyes) =1*. As a result, the maximized score becomes $score_{sense}$*(正解, helyes)=1*.

正解 and *becsületes* have one common English translation: *right*. As already described above, among the 13 senses 4 are identified from the Japanese-to-English definition (#1, #3, #5, #10, all with *right*). However, only one sense is identified from the Hungarian-to-English definition (#4, with *honorable* and *honourable*). Because there are no common senses were identified, $score_{syns}$*(正解, becsületes)=0* and the translation candidate should not qualify as a translation pair, because it is obvious that the common English definition *right* is used with different senses in the two definitions.

Since we do not use part-of-speech information from the dictionaries, the translation candidates are verified based on all four part-of-speech parts available in WordNet. Scores that pass a global $threshold_{sense}$ are considered as correct translations. Empirically this $threshold_{sense}$ was set to *0.1*; a number of 33971 candidates (*type B* translations) were selected.

3. Synonymy

As mentioned in our introduction, different dictionaries have different lexical and semantic structures, therefore although the definitions describe the same concept, the different selection of words in the descriptions results in a difficult identification based on lexical information only. We try to overcome this problem by expanding the translation candidates' English descriptions with all synonyms of its words and expressions. As a result, the similarity of the two expanded English descriptions gives a better indication on the suitability of the translation candidate.

According to a Leibniz's definition, two expressions are synonymous if the substitution of one for the other never changes the truth value of a sentence in which the substitutions is made. In WordNet a weaker definition is applied: *"two expressions are synonymous in a linguistic context C if the substitution of one for the other in C does not alter the truth value"*, arguing that true synonyms might be extremely rare, if they exist at all (Miller et al., 1990). The loose definition of synonyms allows us to have a wider range of words to score our translation candidates.

The scoring method is as follows: for a given Japanese-Hungarian translation candidate *(j, h)* we look up for their translations into English from the respective dictionaries ($e_J=e_1(j), e_2(j), \dots e_n(j)$ and $e_H=e_1(h), e_2(h), \dots e_m(h)$, respectively). For every Japanese-to-English and Hungarian-to-English translation we look up their synonyms ($syn(e_J)$ and $syn(e_H)$, respectively) using WordNet. $score_{syns}(j, h)$ is the ratio of the common and total number of words from the newly expanded translations. The higher the $score_{syns}(j,h)$, the most likely that the candidate *(j,h)* is a correct translation.

$$score_{syns}(j,h) = \frac{count((e_J \cup syn(e_J)) \cap (e_H \cup syn(e_H)))}{count((e_J \cup syn(e_J)) \cup (e_H \cup syn(e_H)))} \quad (2)$$

Since synonymy information from WordNet is available for nouns, verbs, adjectives and adverbs, four separate scores are calculated for each part-of-speech.

With this selection method we cannot set a global threshold, because the scores based on this relation highly depend on the number of English translations, moreover the quality of the expanded word group is manipulated by the initial descriptions. However, we can create lists of words grouped by Japanese or Hungarian word entries and create a local classification, based on the score. We create four lists for the four part-of-speeches. Under the same word entry most correct translations were among the top scoring pairs, therefore we empirically determined a local threshold to select the top scoring candidates: $threshold_{syns}=max(score_{syns})*9/10$. However, when even the top score fails to go over *0.1*, we chose not to select it, considering that in the case of the entry word in question the synonymy information is not reliable.

Because local orders are created by grouping based on the same word entry, the selection procedure is performed twice: once by grouping the Japanese entries and once by grouping the Hungarian entries. With the introduction of selection based on local thresholds we maintain a high percentage of recall, because in most of the cases at least one translation is selected for every entry.

A total of 196775 candidate pairs were selected, these are called *type C* translations.

4. Antonymy

Another method to expand the entry definition is the usage of antonymy information. An antonym is a word that means the opposite of another word. However, because it is difficult to compare two definition sets that contain words with opposite meanings too, instead of expanding the initial definition with the antonyms, we expand it with the antonyms of the antonyms.

The scoring method is similar with the one used with synonymy information: for a given Japanese-Hungarian translation candidate *(j, h)* we look up for their translations into English from the respective dictionaries (e_J*=*e_1*(j),* e_2*(j), … *e_n*(j)* and e_H*=*e_1*(h),* e_2*(h), … *e_m*(h)*, respectively). For every Japanese-to-English and Hungarian-to-English translation we look up the antonyms of the antonyms (*ant(ant(*e_J*))* and *ant(ant(*e_H*))*, respectively). The resulting $score_{ants}(j, h)$ is the ratio of the common and total number of words from the newly expanded definitions. The higher the $score_{ants}(j,h)$, the most likely that the candidate *(j,h)* is a correct translation.

$$score_{ants}(j,h) = \frac{count((e_J \cup ant(ant(e_J))) \cap (e_H \cup ant(ant(e_H))))}{count((e_J \cup ant(ant(e_J))) \cup (e_H \cup ant(ant(e_H))))} \quad (3)$$

Like in the case of synonymy, every translation candidate is verified based on all four part-of-speech parts available in WordNet; all four antonymy based scores are separately handled during selection. Also we cannot use a global threshold; word entry and part-of-speech governed local lists are created based on the $score_{ants}$. The empirically determined threshold is set to $threshold_{antn}$*=max(*$score_{antn}$*)*9/10*. Similarly with synonymy relation based selection, top scores that fail to pass the *0.1* value are not selected, antonymy relation being considered as unreliable in that case.

With the double directional selection method introduced with the synonymy relation, 99614 translation candidates were selected (*type D* translations).

5. Hypernymy/hyponymy

Unlike synonymy and antonymy, which are lexical relations between word forms, hypernymy/hyponymy is a semantic relation between word meanings (Miller et. al., 1990). Since hyponymy is transitive and asymmetrical, it generates a hierarchical semantic structure, where the hyponym inherits all features of the more generic concept. This convention provides the central organizing principle for the nouns in WordNet. In the case of verbs hypernymy/hyponymy is a more delicate question, but to a certain extent WordNet provides with a hierarchical structure. As a result, nouns and verbs are grouped into semantic categories; the relatedness or similarity in meaning of the words in the same category can contribute to the selection of new translation pairs. It is rational to think that correct translation pairs share a high percentage of semantic categories, with effect in their respective translations to English by means of a high number of common semantic categories.

The scoring method is as follows: for a given Japanese-Hungarian translation candidate *(j, h)* we look up the translations into English from the respective dictionaries (e_J*=*e_1*(j),* e_2*(j), … *e_n*(j)* and e_H*=*e_1*(h),* e_2*(h), … *e_m*(h)* respectively). For all English words from the Japanese-to-English translation we collect all semantic categories in which they belong (*cat(e(J))*). This is repeated for the Hungarian-to-English side too (*cat(e(H))*). As a result, we have two sets of semantic categories; the score is calculated based on the number of common categories and the number of total categories (4). The higher the $score_{hype}(j,h)$, the probable that the candidate *(j,h)* is a good translation.

$$score_{hype}(j,h) = \frac{count(\sum cat(e_J) \cap \sum cat(e_H))}{count(\sum cat(e_J) \cup \sum cat(e_H))} \quad (4)$$

Every translation candidate is verified based on the noun part and verb part of WordNet. $score_{hype}$ also highly depends on the word entry, therefore local threshold is used in this selection method too, with a threshold empirically set to $threshold_{hype}$*= max(*$score_{hype}$*)*4/5*. Scores that pass this threshold but are less than *0.1* are not selected, the hypernymy/hyponymy relation being considered as unreliable in the case of the word entry in question.

Also with double directional selection method 195480 pairs were selected as translation candidates (*type E* translations).

Type A (lexically unambiguous) and *type B* (sense classification based selection) translations are selected pairs that should provide with a good accuracy, but only a limited number of word entries qualify for this type of selections. Lexical limitations are obvious for *type A*. Lexical and semantic structural limitations apply with *type B* too, because only with a finite number of words sense categorization can be performed with synonyms; those synonyms might not even be recognized due to the structural differences of the dictionaries. On the other hand, *type C* (synonymy), *type D* (antonymy) and *type E* (hypernymy/hyponymy) methods create an order among the candidates of a given dictionary entry. Because of the nature of the selections and the translation candidates, the vast majority of entries qualify for these three selections.

As a pre-evaluation of our dictionary, we randomly selected 200 1-to-1 entries for each selection method. We scored the translation pairs as *good* (○), *undecided* (△) or *erroneous* (×). We marked as *good* the translation pairs that convey the same meaning; or the meanings are slightly different, but in certain textual contexts the translation is possible. *Undecided* is the category for pairs that are similar, but a translation based on them would be considered faulty. The *erroneous* category counts the translation pairs that conveyed a different meaning.

The results showed that *type A* and *type B* selections scored higher than all order-based selections, with *type C*, *type D* and *type E* selections failing to deliver the desired accuracy (Table 3). Experiments showed that synonymy, antonymy and hypernymy/hyponymy based methods all create a slightly different order among translation candidates for a given entry, but most of the correct translations usually are among the top scoring candidates. Consequently, we decided to create a single selection method based on the combined results of synonymy, antonymy and hypernymy/hyponymy relations.

6. Combined semantic information

The three separate lists of synonymy, antonymy and hypernymy/hyponymy based selection methods resulted in relatively different translation pair selections in the case of most entries, proving that they cannot be used as standalone selection methods.

Because of the multiple part-of-speech labelling of numerous words in WordNet, many translation pairs can be selected up to four times based on separate part-of-speech information, all within a single semantic information governed method of the three discussed in this section. Since we use a double directional selection method, we can expect that a pair to be selected several times during the opposite direction too. However, this does not happen in many cases. On the other hand, experiments showed that translation pairs that were selected during both directions with the double directional selection method are indeed correct translations in most cases. In other words, translation pairs whose Hungarian translation was selected as a good translation for the Japanese entry; and whose Japanese translation was also selected as a good translation for the Hungarian entry should be awarded with a higher score. In the same way, entries selected only during one direction should receive a penalty.

The scoring method is based on this idea. For every translation candidate we select the maximum *$score_{rel}(j,h)$* from the several part-of-speech (noun, verb, adjective and adverb for synonymy and antonymy relations; noun and verb for hypernymy/hyponymy relations) based scores, multiplied by a multiplication factor (*$fact_{rel}(j,h)$*). The three separate results calculated on separate semantic relation ($rel \in \{syns, ants, hype\}$) based scores are multiplied in turn.

$$score_{comb}(j,h) = \prod_{rel \in \{syns,ants,hype\}} \begin{matrix}[(c_1 + \max(score_{rel}(j,h)))\cdot \\ (c_2 + c_3 \cdot fact_{rel}(j,h))]\end{matrix} \qquad (5)$$

c_1, c_2 and c_3 are constants used to refine the $score_{comb}$. Empirically these are set to *1*, *0.5* and *0.8*, respectively.

The multiplication factor varies between *0* and *1*, awarding the candidates that were selected based on the same part-of-speech two times during the double directional selection; and punishing when selection was made only in a single direction. For example, if a synonymy relation governed method selects a certain translation candidate two times based on adjectival and adverbial information in the Japanese-to-Hungarian direction, but doesn't selected it during the Hungarian-to-Japanese direction, the translation candidate receives a multiplication factor of *0*. However, if it was selected three times during one direction, and two times during the other direction, it receives the score of *0.66*. Translation candidates that weren't selected at all receive a multiplication factor of *0.5*.

Every translation candidate is verified based on this combined score ($score_{comb}$). $score_{comb}$ also highly depends on the word entry, therefore local threshold is used in this selection method too, with a threshold empirically set to $threshold_{comb} = max(score_{comb})*85/100$. Scores that pass this threshold are selected as translation candidates, regardless of their value.

161202 translation pairs were retrieved with this method; we named them *type F* translations.

We already mentioned that during pre-evaluation *type A* and *type B* translations received a score of above 75%, while *type C*, *type D* and *type E* failed to fulfil the expectations. However, *type F* translations scored close to 80%, therefore from the six translation methods presented above we chose only three (*type A*, *B* and *F*) to construct our dictionary, while the remaining three methods (*type C*, *D* and *E*) are used only indirectly for *type F* selection (Table 3). With the described selection methods a Japanese-Hungarian dictionary with 187761 translation pairs was generated.

selection method	**selection type**	**nr of entries**	**precision**		
			○	△	×
lexically unambiguous	A	37391	75.5%	6.5%	18%
sense classification	B	33971	83%	7%	10%
synonymy	*C*	*196775*	*68%*	*5.5%*	*26.5%*
antonymy	*D*	*99614*	*60%*	*9%*	*31%*
hypernymy/ hyponymy	*E*	*195480*	*71%*	*5.5%*	*23.5%*
combined	F	161202	79%	5%	16%

Table 3: Selection type evaluation

Evaluation and Discussion

To properly evaluate the dictionary, we performed three types of evaluation: recall evaluation based on a frequency dictionary, and two types of precision analysis: single entry and multiple entry evaluation. Single entry evaluation consists of 1-to-1 entries examination, while in multiple entry evaluation all translations are grouped under their single source entry and handled as a single unit.

To be able to compare our proposal with currently the best widely applicable conventional algorithms, we created two another Japanese-Hungarian dictionaries, re-implementing the methods proposed by Sjöbergh (2005) and Tanaka & Umemura (1994), using the same source dictionaries as with our method. However, while Sjöbergh reports a good precision for translation pairs that exceed the *score* of *0.9*, in the case of our Japanese-Hungarian dictionary the number of suitable pairs was too small at 25218, obviously at the recall's expense. To achieve at least similar recall with our method, we decided to modify the threshold to retrieve a similar amount of translation pairs with our method's dictionary. We managed to generate 187610 translation pairs setting the threshold to *0.283*.

With the Tanaka-Umemura method we generated a dictionary with 105632 entries.

Recall Evaluation

In the introduction we argued that current recall evaluation methods do not properly reflect the true value of the dictionaries. One possible solution is to weight each word based on its frequency in use. Since there is no frequency dictionary in either Japanese or Hungarian, we created a Japanese frequency dictionary.

Japanese Frequency Dictionary

The EDR corpus (Isahara, 2007) is a large, annotated corpus with 207360 sentences taken from newspaper articles. It has 124071 different words grouped into 12 part-of-speech categories with an average frequency of 39.6. Among these, 51.12% had only 1 occurrence in the entire corpus, therefore we opted for not to consider them, because either they are field-specific words too rare to be part of any regular dictionary, or they are mismatches in the corpus.

We do not affirm that the EDR corpus is a strict reflection of the language in use, but we do believe that there is a correlation between the daily language and the language used in the corpus. We believe that high frequency words in the corpus are high frequency words in the daily spoken language too, although variances in the frequency level might exist. We also do not affirm that a single recall score based on this frequency dictionary can tell about a dictionary whether it is good or not, but we believe that the score will be useful in comparing dictionaries, the higher score pointing out the dictionary with the better recall.

During recall evaluation all w_i words from the frequency dictionary *($dict_{freq}$)* with a frequency value *(freq(w))* of more than 1 are verified whether they are translated or not in the three separately constructed dictionaries *($dict_{jp\text{-}hu}$)*.

$$recall = \frac{\sum_{w_i \in dict_{jp-hu}} freq(w_i)}{\sum_{w_i \in dict_{freq}} freq(w_i)} \qquad (6)$$

As a result, we obtained a recall of 37.03% for the Sjöbergh dictionary, 30.76% for the Tanaka-Umemura dictionary and 51.68% for our dictionary. As another comparison, the initial translation candidates from step 1 had a recall of also 51.68%. While the Sjöbergh and Tanaka-Umemura methods lost between 0.14 and 0.20 from the initial recall value of the translation candidates, our method managed to maintain a seemingly perfect recall.

However, the manually created Japanese-English dictionary that we used for translation candidate generation had a 73.23% recall, significantly higher than our dictionary's value.

Single Entry Evaluation

With single entry evaluation we randomly extracted 1000 1-to-1 translation pairs from each of the three dictionaries. We scored the translation pairs as *good* (○), *undecided* (△) or *erroneous* (×) the same way as with selection type evaluation. Our method managed to outperform every other method, scoring 79.0% against the Sjöbergh method's 54.0% and the Tanaka-Umemura method's 62.5% (Table 5).

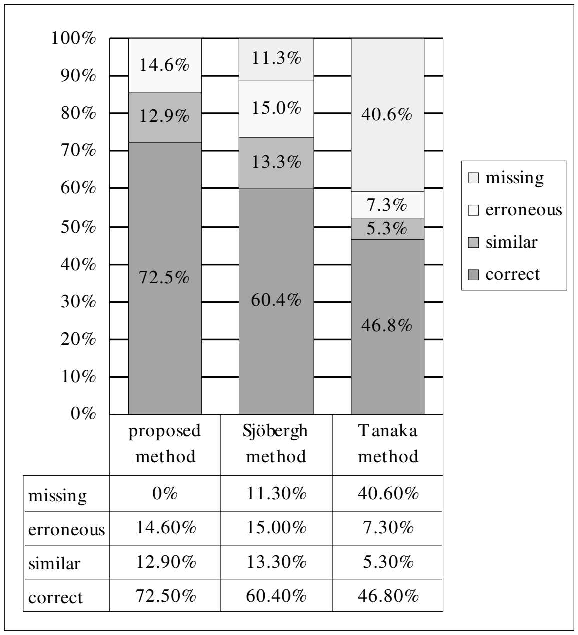
dictionary generation method	**nr of entries**	**single entry precision**		
		○	△	×
proposed method	187761	79.0%	6.3%	14.7%
Sjöbergh method	187610	54.0%	9.9%	36.1%
Tanaka-Umemura method	105632	62.5%	7.9%	29.6%

Table 5: Single entry evaluation results

Multiple Entry Evaluation

For multiple entry evaluation 1000 randomly selected Japanese word entries were selected. Parallel Japanese-to-Hungarian evaluation of the three dictionaries was performed on the selected entries. The entries were scored as *correct*, *similar*, *erroneous* or *missing*. The entry is considered to be *correct* if all Hungarian translations are correct. The entry is *similar* if the Hungarian translations are predominantly correct. If the number of wrong translations exceeds 2, the entry is *erroneous*. The *missing* option refers to the Japanese entries that had no Hungarian translation.

Our method scored a 72.5% multiple entry value, outperforming the Sjöbergh method's 60.4% and the Tanaka-Umemura method's 46.8%. The latter dictionaries suffered mainly because of the missing Hungarian translations (Figure 2).



	proposed method	Sjöbergh method	Tanaka method
missing	0%	11.30%	40.60%
erroneous	14.60%	15.00%	7.30%
similar	12.90%	13.30%	5.30%
correct	72.50%	60.40%	46.80%

Figure 2: Multiple entry evaluation

Discussions

Based on the recall evaluations, the traditional methods showed their major weakness by losing substantially from the initial recall values, scored by the initial translation candidates. Our method maintained the same value with the translation candidates, but we cannot say that the recall is perfect. When compared with a manually created dictionary, our method also lost significantly.

Precision evaluation also showed an improvement compared with the traditional methods, our method outscoring the other two methods in both single entry and multiple entry evaluations.

Discussing the weaknesses of our system, we have to divide the problems into two categories: recall problems deal with the difficulty in connecting the Japanese and Hungarian entries with the intermediate language, while precision problems discuss the reasons why erroneous pairs are produced.

1. Recall problems and possible solutions

The main reason in failing to connect possible translation pairs is that certain words do not exist in the target or intermediate languages, therefore translations are explanations instead of translation equivalents. Japanese language is particularly rich from this point of view. Also certain part-of-speeches (particle, auxiliary verb, etc) don't have direct equivalents in the target language, or their translation is too ambiguous. At this point we cannot think of any automated method to solve these problems.

Another recall problem is that a number of words that are in the intermediate language can not be retrieved from the ontology. Ontology improvement is needed to correct this problem.

2. Precision problems and possible solutions

We identified two types of precision problems. The most obvious reasons for erroneous translations are the polysemous nature of words and the meaning-range differences across languages. With words whose senses are clear and mostly preserved even through the intermediate English language, most of the correct senses were identified and correctly translated. Nouns, adjectives and adverbs had a relatively high degree of accuracy. However, verbs proved to be the most difficult part-of-speech to handle. Because they are more flexible in meaning than other part-of-speeches, and the meaning range is also highly flexible across languages, the correct translation is increasingly difficult. For this reason, the number of faulty translations and the number of meanings that are not translated is relatively high.

One other source of erroneous translations is the quality of the initial dictionaries. Even the supposable unambiguous *type A* translations fail to produce the desired accuracy, although they are the unique candidate for a given word entry. The number one reason for this is the deficiency of the initial dictionaries, which contain a great number of irrelevant or low usage translations, shadowing the main, important senses of some words. The secondary reason for this phenomenon is the meaning shift that occurs between Japanese, English and Hungarian words.

Surprisingly *type A* scores are bettered by the rest of our selected translation methods, proving that shifting the selection method from the dictionaries to the ontology is an efficient method for automatized dictionary generation.

Other ontology resources for the intermediate language should improve the accuracy of this method. More accurate source dictionaries also might raise the quality of the generated dictionary, but even so we believe that most of the corrections will have to be performed manually.

Conclusions and Future Plans

We proposed a new pivot language based method to create bilingual dictionaries. Opposed to conventional methods that use dictionaries as a main resource, our method uses an ontology of the intermediate language to select the suitable translation pairs. As a result, we eliminate most of the weaknesses caused by the structural differences of dictionaries, while profiting from the semantic relations provided by the ontology. We believe that because of the nature of our method it can be re-implemented with most language pairs.

We concentrated on achieving a high recall in order to minimize the work of manual labour during human correction. We generated a mid-large sized dictionary with relatively good recall and promising precision.

Our future plans include improving our method by means of enhanced scoring and possibly adaptation of other semantic resources. We also plan to verify the efficiency of the dictionary with our future Japanese-Hungarian machine translation system.

We also plan to manually correct the errors of our dictionary.

Bibliographical References

Breen, J.W. (1995): "Building an Electric Japanese-English Dictionary", Japanese Studies Association of Australia Conference, Brisbane, Queensland, Australia.

Brown, P., Cocke, J., Della Pietra, S., Della Pietra, V., Jelinek, F., Mercer, R., Roossin, P. (1998): "A Statistical Approach to Language Translation", In COLING-88, 1, 71-76.

Brown, R.D. (1997): "Automated Dictionary Extraction for Knowledge-Free Example-Based Translation", In Proceedings of the 7th International Conference on Theoretical and Methodological Issues in Machine Translation, Santa Fe, 111-118.

Isahara, H. (2007). "EDR Electronic Dictionary – present status" (EDR 電子化辞書の現状), NICT-EDR symposium, pp 1-14. *(in Japanese)*

Kay, M., Röscheisen, M. (1993): "Text-Translation Alignment", Computational Linguistics, 19(1), 121-142.

Miller G.A., Beckwith R., Fellbaum C., Gross D., Miller K.J. (1990): "Introduction to WordNet: An On-line Lexical Database", Int J Lexicography 3(4), pp 235-244.

Paik, K., Bond, F., Shirai, S. (2001): "Using Multiple Pivots to align Korean and Japanese Lexical Resources", In NLPRS-2001, pp. 63-70, Tokyo, Japan.

Sjöbergh, J. (2005): "Creating a free Japanese-English lexicon", In Proceedings of PACLING, pp 296-300, Tokyo, Japan.

Shirai, S., Yamamoto, K. (2001): "Linking English words in two bilingual dictionaries to generate another pair dictionary", In ICCPOL-2001, pp 174-179, Seoul, Korea.

Tanaka, K., Umemura, K. (1994): "Construction of a bilingual dictionary intermediated by a third language", In Proceedings of COLING-94, 297-303, Kyoto, Japan.