Dictionary generation for less-frequent language pairs using WordNet

István Varga and Shoichi Yokoyama
Graduate School of Science and Engineering, Yamagata University, Yonezawa, Japan

Chikara Hashimoto
National Institute of Information and Communications Technology, Kyoto, Japan

Abstract

Bilingual dictionaries are vital resources in many areas of natural language processing. Numerous methods of machine translation require bilingual dictionaries of large coverage, but less-frequent language pairs rarely have any digitalized resources of such kind. Since the need for these resources is increasing, but the human resources are scarce for less represented languages, efficient automatized methods are imperative. This article presents a fully automated, robust intermediate language-based bilingual dictionary generation method that uses the WordNet of the intermediate language to build a new bilingual dictionary. We propose the usage of WordNet in order to increase accuracy; we also introduce a bidirectional selection method with a flexible threshold to maximize recall. The evaluations showed 79% accuracy and 51% weighted recall, outperforming representative pivot language-based methods. A dictionary generated with this method will still need manual post-editing, but the improved recall and precision decrease the work of human correctors.

Correspondence:
István Varga, PhD student,
Graduate School of Science and Engineering,
Yamagata University,
Yonezawa, Japan.
E-mail:
dyn36150@dipfr.dip.yz.yamagata-u.ac.jp

1 Introduction

Although the quality of machine translation is still a few steps away from what was dreamed decades ago, automatic and semi-automatic machine translation systems gradually do manage to take over costly human tasks. This much welcomed change can be attributed not only to major developments in techniques regarding translation methods, but also to important translation resources, such as monolingual or bilingual dictionaries and corpora, thesauri, and so on.

However, while widely used language pairs can fully take advantage of state-of-the-art developments in machine translation, some low-frequency, or less common language pairs lack some or even most of the above-mentioned translation resources. In that case, the key to a high-performance machine translation system switches from the choice and adaptation of the translation method to the question of whether there are any translation resources or not between the chosen languages. Since the relative small number of potential users proves to be the only bottleneck concerning the justification of such a machine translation system, low-cost translation resource development is the viable solution.

Literary and Linguistic Computing, Vol. 24, No. 4, 2009. © The Author 2009. Published by Oxford University Press on behalf of ALLC and ACH. All rights reserved. For Permissions, please email: journals.permissions@oxfordjournals.org
doi:10.1093/llc/fqp025 Advance Access published on 15 July 2009

A natural language can be best described with its content (lexical elements) and structure (grammatical elements). In a multilingual environment a bilingual dictionary represents the content, while bilingual grammatical rules or bilingual sentence patterns represent the structure. In our article we challenge the first element, the structure, by proposing a new bilingual dictionary generation method as a possible solution to low-cost translation resource development. Since low-resourced language pairs rarely have any aligned bilingual corpora, a bilingual dictionary cannot be induced from aligned texts with methods described by Kay and Röscheisen (1993), Brown (1997), or Brown *et al.* (1998).

We choose the Hungarian and Japanese languages as an example of low-frequency language pair. Besides daily usage, our dictionary will be incorporated in our Japanese – Hungarian rule-based machine translation (RBMT) system. As a result our dictionary can contain loose or similar translations as well, that are correct only in certain contexts. Our system, besides using the bilingual dictionary and a set of grammatical rules (sentence patterns) in its transfer process, will use a language model familiar from statistical machine translation (SMT) meant to score and select from the multiple outputs of the transfer.

This article is structured as follows: after we investigate pivot language-based methods in various fields of natural language processing (NLP), we analyse the problems of current dictionary-generating methods, followed by our goals and the details of our proposal. We evaluate the dictionary created using our proposed method, including a comparative evaluation with two other baseline dictionary-generation methods. Finally, we present our conclusions and prospects for the future.

2 Related Work

Pivot resource-based methods are widespread in NLP; we describe the most recent findings in intermediate language-based machine translation (MT) and cross-language information retrieval (CLIR). We also present similar methods in bilingual dictionary generation. Because our proposal uses information extracted from WordNet, we also overview the latest findings regarding WordNet.

2.1 Pivot language-based methods in NLP

Babych *et al.* explored the usage viability of transitive MT, presenting a Russian pivoted Ukrainian to English MT experiment (2007). Both MT systems are used as a black box, with no possibility to alternate the system's components. The evaluation showed 18–37% improvement over the direct MT system. Bick and Nygaard presented a successful semi-transitive method (2007). They use Danish to intermediate for their Norwegian to English system, but instead of two separate MT systems (Norwegian to Danish and Danish to English, respectively) they handle Norwegian as misspelled Danish, which in turn is used as input for the well-performing Danish to English system. Wu and Wang presented a pivot language approach for phrase-based statistical machine translation, with English and German acting as intermediates for a French–Spanish machine translation system. They reported a 22% improvement against the direct system (Wu and Wang, 2007).

Pivot language-based methods proved their relative usefulness in numerous tasks in CLIR with the reported precisions being between low and promising (Gollins and Sanderson, 2001; Ballesteros, 2000; Kraaij, 2003; Lehtokangas, 2002).

2.2 Pivot language-based bilingual dictionary generation

Pivot language-based dictionary generation methods use only dictionaries to and from an intermediate language to generate a new dictionary. They rely on the idea that the lookup of a word in an uncommon language through a third, intermediated language can be automated. 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 and Umemura's method uses bidirectional source–intermediate and intermediate–target dictionaries (harmonized dictionaries). Correct translation pairs are 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. Tanaka and Umemura report that this method is 'useful for revising and supplementing the vocabulary of existing dictionaries' (Tanaka and Umemura, 1994).

Sjöbergh's English intermediated Swedish–Japanese dictionary generation method compares each Japanese-to-English description with each Swedish-to-English description. Scoring is based on word overlap, weighted with inverse document frequency (IDF). Sjöbergh reports 'a dictionary with high recall and good precision' (Sjöbergh, 2005).

The two approaches described above are the best-performing ones that are robust enough to be applicable with other language pairs as well. In our research we used these two methods as baselines for our proposal.

There are numerous refinements of the above methods, but for various reasons they cannot be implemented with any arbitrary language pair. Shirai and Yamamoto (2001) used English to design a Korean–Japanese dictionary, but because of the usage of language-specific information, they conclude that their method 'can be considered to be applicable to cases of generating among languages similar to Japanese or Korean through English'. In other cases, only a small portion of the lexical inventory of the language is chosen to be translated: Paik and Bond and Shirai (2001) proposed a method with multiple pivots (English and Kanji/Hanzi characters) to translate Sino–Korean entries. Although owing to the introduction of the Kanji/Hanzi character information the precision was remarkably high, for the same reason the translatable entries are quite limited. Bond and Ogura describe a Japanese–Malay dictionary that uses a novel technique in its improved matching through normalization of the pivot language, by means of semantic classes, but only for nouns. Besides English, they also use Chinese as a second pivot. The generated dictionary is 'reasonably accurate (...), useful not only for humans, but with the information required by a semantic transfer-based machine translation system' (Bond and Ogura, 2008).

2.3 WordNets as resources in natural language processing

Large lexical databases are vital for many areas in NLP, where large amounts of structured linguistic data are needed. The appearance of WordNet (Miller *et al.*, 1990) had a big impact in NLP, since it not only provided one of the first wide-range collections of linguistic data in electronic format, but it also offered a relatively simple structure that can be implemented with other languages as well. In the last decades since the first, English WordNet, numerous languages adapted the WordNet structure.

Multilingual projects, such as EuroWordNet (Vossen, 1998; Peters *et al.*, 1998), Balkanet (Stamou *et al.*, 2002) or Multilingual Central Repository (Agirre *et al.*, 2007) aimed to solve numerous problems in natural language processing. EuroWordNet (EWN) was specifically conceived for word disambiguation purposes in CLIR (Vossen, 1998). Besides proving its relative usefulness for the task in numerous occasions (Gonzalo *et al.*, 1998; Clough and Stevenson, 2004; Santiago *et al.*, 2002), some limitations also became apparent (Clough and Stevenson, 2004; Voorhees, 1994).

The internal structure of the multilingual WordNets itself can be a good starting point for bilingual dictionary generation. In case of EuroWordNet, besides the internal design of the initial WordNet for each language, an Inter-Lingual-Index interlinks word meaning across languages (Peters *et al.*, 1998). However, there are two limitations: first of all, the size of each individual language database is relatively small (Vossen, 1998), covering only the most frequent words in each language, thus not being sufficient for a dictionary with a large coverage. Secondly, these multilingual databases cover only a handful of languages, with Hungarian or Japanese not being part of any of them. Adding a new language would require the existence of a WordNet of that language. The Japanese language is one of the most recent ones added to the WordNet family (Isahara *et al.*, 2008), but the Hungarian WordNet is still under development (Prószéky *et al.*, 2001; Miháltz and Prószéky, 2004).

3 Problems of Current Pivot Language-Based Methods

3.1 Selection method shortcomings

Pivot language-based methods usually generate and score a number of translation candidates, and the candidate's scores that exceed a certain pre-defined global threshold are selected as viable translation pairs. However, the scores highly depend on the number of translations in the intermediate language; therefore, there is a fluctuation in what that score represents. For example, the same value can represent not only an accurate translation of a certain entry, but also an inaccurate translation of another entry. For this reason, a large number of entries are entirely left out from the dictionary, since all of their translation candidates scored low, while faulty translation candidates are selected with some correct ones, because they surpass the global threshold.

3.2 Dictionaries only are not enough as a resource

Regardless of the language pair, in most cases the meanings of the corresponding entries are not identical; they only overlap to a certain extent. Therefore, the pivot language-based dictionary generation problem can be defined as the identification of the common elements or the extent of the definitions' overlap. Current pivot language-based methods fail to deliver the desired recall and precision because dictionaries are an incomplete source of information for this purpose. There are two reasons why dictionaries only as resources are not enough:

3.2.1 *Dictionaries do not provide enough information about the language*

Dictionaries do provide bidirectional or unidirectional lexical relations between words across languages, which is enough for a human to correctly identify and connect the meanings in the intermediate language. However, for the computer it is very difficult to 'understand' these overlaps, because machines need more information, desirably a full description of words. Bilingual dictionaries do not have that. The translation of a certain entry in a regular computer-readable bilingual dictionary is a simple text that represents the meanings of the entry, not a full semantic description. Humans can connect the semantic meanings that the translations represent, but computers can only compute the representation of the meaning, the character string. Therefore, instead of the desired semantic overlap, current automated methods can only perform lexical overlap on an incomplete translation.

3.2.2 *Differences in dictionary compilation*

Even semantically identical or very similar words can have different definitions in different dictionaries. For example, our Japanese–English and Hungarian–English dictionaries the entry#1 *(sok)* and entry#2 (沢山-takusan) share approximately the same meaning, but our resource dictionaries provide us with almost entirely different translations (Table 1).

Even if the translations from the source and target languages are correctly transferred to the intermediate language, due to the inconsistent translations from the target and source languages and lack of proper correspondences between the two definitions, the recall and the precision suffer.

For example, the entries #3, #4, #5, and #6 have similar meanings (Table 1). Therefore, the Hungarian–Japanese translation candidates ((#3, #5), (#4, #5), (#6, #5), and (#3, #7)) all share a single common entry in turn, but because of the ambiguities of 'to depict' and 'to design', the latter two ones are incorrect. However, because of the different definitions and the lexical nature of the overlap, current 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. Tanaka and Umemura's method failed to recognize any of the correct translation pairs, while Sjöbergh's method chose an incorrect one (#6, #5).

A similar effect can be observed with the translation candidates (#8, #10) and (#9, #10). Both of these translations are correct, but because of the presence of two common elements in the first candidate's translations ('emptiness' and 'blank'), it received much higher scores than the second one. As a result, it was correctly generated as a translation pair, but the second one, which is equally correct, was not chosen.

Table 1 Translation candidate examples with their respective English translations

No.	Hungarian or Japanese entry	English translation (according to our source dictionaries)
1	sok	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
2	沢山 (takusan)	many, a lot, much
3	rajzol	to design, to draw, drew, drawn, to lay down, to limn
4	fest	to blazon, to decorate, to dip, to dye, to limn, to paint, to stain
5	描く(kaku)	to draw, to paint, to sketch, to depict, to describe
6	ábrázol	to delineate, to depict, to illustrate, to limit, to picture, to plot, to portrait, to represent, to typify, to write down
7	図る (hakaru)	to plot, to attempt, to plan, to take in, to deceive, to devise, to design, to refer A to B
8	űr	blank, chasm, emptiness, gap, space, vacancy
9	üresség	blankness, cavity, emptiness, hollowness, vacancy, viciousness, void
10	空 (kara)	emptiness, vacuum, blank

4 Proposed Method

4.1 Goals

The main goal of our research is the proposal of a bilingual dictionary-generation method, which is applicable with most language pairs. We exemplify our proposal with the generation of a Hungarian–Japanese dictionary.

We believe that besides good precision, high recall is the most important necessity for such a bilingual dictionary for two reasons. First of all, the main usage of this dictionary will be as a translation resource for a machine translation system, therefore a large coverage is vital. As a second reason, we can argue that a dictionary with high recall is less costly to revise, as also pointed out by Bond and Ogura (2008). Although we expect high 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, with their translation becoming entirely manual.

4.2 Specifics of our proposal

For good precision, instead of the familiar lexical overlap of the current methods we calculate the ‘semantic overlap’ of the source–intermediate and target–intermediate translations. In order to do that, we use semantic information extracted from the WordNet of the intermediate language.

To improve recall, we introduce ‘double directional selection’. We can group the intermediate language translations that share the same entry, and set a local threshold for their scores. Thus, we guarantee that the entry has at least one translation in the resulting dictionary, maintaining a high recall. Since we can group the entries in the source language and target language as well, we perform this selection twice, once in each direction.

4.3 Translation resources

For translation candidate generation, we use two dictionaries.

(1) ‘edict’ is a Japanese to English unidirectional dictionary created and maintained by Jim Breen (1995) which has 197282 1-to-1 entries after cleaning. It can be freely downloaded from the Internet (http://www.csse.monash.edu.au/∼jwb/j_edict.html).

(2) a Hungarian–English bidirectional dictionary created and maintained by Attila Vonyó that has 189331 1-to-1 entries after cleaning. This dictionary can also be freely downloaded from the Internet (http://almos.vein.hu/∼vonyoa/SZOTAR.HTM).

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. We believe that this way our method is more robust, since many computer-readable dictionaries do not include part-of-speech information.

To select from the translation candidates, we mainly use 'WordNet' (Miller *et al.*, 1990). WordNet is 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. WordNet is also freely available from the Internet *(*http://wordnet.princeton.edu/*)*. From WordNet we consider four types of semantic information:

(1) Sense classification: a detailed description for each word regarding its senses. Explanations as well as synonyms are provided for most senses of each word.

(2) Synonymy: 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).

(3) Antonymy: An antonym is a word that means the opposite of another word, or in WordNet's definition 'a lexical relation between word forms'. For adjectives and adverbs, antonymy is the main central organizing principle.

(4) Hypernymy/hyponymy: 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. With verbs hypernymy/hyponymy is more ambiguous, but to a certain extent WordNet provides with a hierarchical structure, although it is less deep than for nouns. 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.

4.4 Dictionary-generation method

Our proposed method consists of two steps. In step 1 we generate translation pair candidates that we believe will contain most of the correct translation pairs. In step 2 we first perform a limited lexical-based selection, after which we score all translation candidates based on semantic information extracted from WordNet. Finally also in step 2 we select the most appropriate candidates based on the scoring. Both steps are entirely automatic. Below is the detailed description of our method.

STEP 1: *Translation candidate generation*

We first generate the translation candidates. We look up every entry in turn from the source–intermediate dictionary, looking up also every intermediate–target entry in which the intermediate entry matches the source translation definition result from the source–intermediate dictionary. We consider every source–target pair as a translation candidate for the next step. For example, in the case of English intermediated Japanese–Hungarian dictionary generation, 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 (Fig. 1). Understandably a large number of obviously erroneous pairs and pairs with too little semantic similarity will be also included. For example, the pair 曖昧 —'halvány' is not correct, 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

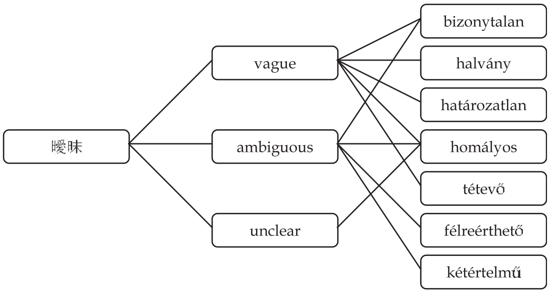

Fig. 1 An example of translation candidate generation

entry has the same meaning with 'vague' as 'dim, lacking clarity'. However, the identification of homonyms and elimination of noisy translations will be performed in the next step.

Thus, 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 target language entries too, because it would result in exactly the same translation candidates.

With the method described above we accumulated 436966 Japanese–Hungarian translation candidates.

Step 2: *Translation pair selection*

We examine the translation candidates one by one, looking up the source–intermediate and target–intermediate dictionaries, comparing the intermediate language translations. The translation candidates are scored by the correlation of the two translation 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 WordNet.

(1) Lexically unambiguous translation pair extraction

Some of the translation candidates have exactly the same definitions in the intermediate language; we consider these pairs as being correct by default. Also, among the translation candidates we identified source entries that have only one target translation and target entries that have only one source translation. Being the sole candidates for the given entries, we consider these pairs too as being correct.

37391 Japanese–Hungarian translation pairs were retrieved with this method; we call them 'type A' translations.

(2) Sense classification

We use the synonyms of WordNet's sense description to disambiguate the common meanings. The scoring method is as follows: for a given source–target translation candidate (s,t) we look up their translations into the intermediate language from the respective dictionaries $(s\rightarrow I=\{s\rightarrow i_1, s\rightarrow i_2, \ldots s\rightarrow i_n\}$ and $t\rightarrow I=\{t\rightarrow i_1, t\rightarrow i_2, \ldots t\rightarrow i_m\})$. We select the intermediate language definitions that are common in the two definitions $(s\rightarrow I \cap t\rightarrow I)$ and we look up their respective senses $(sense(s\rightarrow i), sense(t\rightarrow i))$ 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 source-to-intermediate definitions and a certain set of senses from the target-to-intermediate definitions. We mark $score_B(s,t)$ the maximum ratio of the identical and total identified sets of the common words. The higher the $score_B(s,t)$ is, the more probable that the candidate (s,t) is a valid translation.

$$Score_B(s,t) = \max_{i\in(s\rightarrow I)\cap(t\rightarrow I)} \frac{|sense(s\rightarrow i)\cap sense(t\rightarrow i)|}{|sense(s\rightarrow i)\cup sense(t\rightarrow i)|} \quad (1)$$

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 analyse *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 'correct') and 5 from the Hungarian-to-English definition (#1, #3, #5, #6, #10, with 'correct' and 'proper'). As a result, four senses are common, and one is different. Based on equation (1) $score_{B(\text{“right”})}$ (正解, $helyes) = 0.8$, when scoring is done through the word ‘right’. ‘Correct’ has four senses according to WordNet, all of them are recognized by both definitions through ‘right’, therefore the score through ‘correct’ is $score_{B(\text{“correct”})}$(正解, $helyes) = 1$. As a result, the maximized score becomes $score_B$(正解, $helyes) = 1$.

正解 and *becsületes* have one common English translation: ‘correct’. As already described above, among the 13 senses 4 are identified from the Japanese-to-English definition (#1, #3, #5, #10), all through ‘correct’. However, only one sense is identified from the Hungarian-to-English definition (#4, with ‘honorable’ and ‘honourable’). Because no common senses were identified, $score_B$(正解, $becs\ddot{u}letes) = 0$ 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 parts-of-speech available from WordNet. Scores that pass a global $threshold_B$ are considered as correct translations. Empirically this $threshold_B$ was set to *0.1*; 33971 Japanese–Hungarian 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’ intermediate language descriptions with all of their synonyms. As a result, the similarity of the two expanded intermediate language descriptions gives a better indication on the suitability of the translation candidate. Since the same word or concept’s translations into the intermediate language also share the same semantic value, with the expansion by means of synonyms, the lexical representation becomes less incomplete.

The scoring method is as follows: for a given source–target translation candidate *(s,t)* we look up their translations into intermediate language from the respective dictionaries $(s \rightarrow I = \{s \rightarrow i_1, s \rightarrow i_2, \ldots s \rightarrow i_n\}$ *and* $t \rightarrow I = \{t \rightarrow i_1, t \rightarrow i_2, \ldots t \rightarrow i_m\})$. For every source-to-intermediate and target-to-intermediate translation we look up their synonyms $(syn(s \rightarrow I)$ *and* $syn(t \rightarrow I))$ using WordNet. $score_C(s,t)$ is the ratio of the common and total number of words from the newly expanded translations. The higher the $score_C(s,t)$ is, the more likely that the candidate *(s,t)* is a correct translation pair.

$$Score_C(s,t) = \frac{|(s \rightarrow i \cup syn(s \rightarrow i)) \cap (t \rightarrow i \cup syn(t \rightarrow i))|}{|(s \rightarrow i \cup syn(s \rightarrow i)) \cup (t \rightarrow i) \cup syn(t \rightarrow i))|}. \qquad (2)$$

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

Since the scores based on this relation highly depend on the number of intermediate language translations, we use the double directional selection method with a local threshold empirically set to $threshold_C = max(score_C)^{*}0.9$. 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.

A total of 196775 Japanese–Hungarian 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. 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 source–target translation candidate *(s,t)* we look up for their translations into intermediate language from the respective dictionaries $(s \rightarrow I = \{s \rightarrow i_1,$

$s\rightarrow i_2, \ldots s\rightarrow i_n\}$ and $t\rightarrow I=\{t\rightarrow i_1, t\rightarrow i_2, \ldots t\rightarrow i_m\}$). For every source-to-intermediate and target-to-intermediate translation we look up the antonyms of the antonyms *(ant(ant(s→I)) and ant(ant(t→I)))*. The resulting $score_D(s,t)$ is the ratio of the common and total number of words from the newly expanded definitions. The higher the $score_D(s,t)$ is, the most likely that the candidate *(s,t)* is a correct translation pair.

$$Score_D(s,t)= \frac{|(s\rightarrow i\cup ant(ant(s\rightarrow i)))\cap(t\rightarrow i\cup ant(ant(t\rightarrow i)))|}{|(s\rightarrow i\cup ant(ant(s\rightarrow i)))\cup(t\rightarrow i\cup ant(ant(t\rightarrow i)))|}. \quad (3)$$

Similarly with synonymy, every translation candidate is verified based on all four parts-of-speech 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 $score_D$. The empirically determined threshold is set to $threshold_D = max(score_D)^{*}0.9$. 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 specific case.

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

(5) Hypernymy/hyponymy

It is rational to think that correct translation pairs share a high percentage of semantic categories, with effect in their respective translations to the intermediate language by means of a high number of common semantic categories.

The scoring method is as follows: for a given source–target translation candidate *(s,t)* we look up the translations into the intermediate language from the respective dictionaries *($s\rightarrow I=\{s\rightarrow i_1, s\rightarrow i_2, \ldots s\rightarrow i_n\}$ and $t\rightarrow I=\{t\rightarrow i_1, t\rightarrow i_2, \ldots t\rightarrow i_m\}$).* For all intermediate language words from the source-to-intermediate translation we collect all semantic categories in which they belong (*semcat(s→I)*). This is repeated for the target-to-intermediate dictionary too (*semcat(t→I)*). 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 $score_E(s,t)$ is, the more probable that the candidate *(s,t)* is a good translation pair.

$$Score_E(s,t)=\frac{|semcat(s\rightarrow i)\cap semcat(t\rightarrow i)|}{|semcat(s\rightarrow i)\cup semcat(t\rightarrow i)|}. \quad (4)$$

Every translation candidate is verified based on the noun part and verb part of WordNet. $score_E$ also highly depends on the word entry; therefore, local threshold is used with this selection method too, with a threshold empirically set to $threshold_E = max(score_E)^{*}0.8$. 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 Japanese–Hungarian 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 candidates qualify for these two selections. Lexical limitations are obvious for 'type A'. Lexical and structural limitations apply with 'type B' too, since not all words have sense categorization in WordNet. Even the ones that have, the 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) are more widely applicable and they create an order among the candidates for a given dictionary entry.

As a pre-evaluation of our dictionary, we randomly selected 200 1-to-1 entries for each selection method. We scored the translation pairs using the following criteria:

- **good** (○): the source entry and its translation convey the same meaning, or the meanings are slightly different, but in a certain context the translation is possible;
- **undecided** (?): the source entry and its translation's semantic value are similar, but a translation based on this entry would be faulty; and
- **erroneous** (×): the source entry and its translation convey different meanings.

Table 2 Selection type evaluation (the not selected methods in italic)

Selection method	Selection type	Number 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%

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 2). 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-based methods (synonymy, antonymy, hypernymy/hyponymy) 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 are indeed correct translations in most cases. In other words, translation pairs whose target language translation was selected as a good translation for the source language entry and whose source language translation was also selected as a good translation for the target language 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 $score_{rel}(s,t)$ 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 ($fact_{rel}(s,t)$). The three separate results calculated on separate semantic relation ($rel \in \{syns,ants,hype\}$)-based scores are multiplied in turn.

$$Score_F(s,t) = \prod_{rel \in \{syns,ants,hype\}} ((c_1 + \max(score_{rel}(s,t))) \cdot (c_2 + c_3 \cdot fact_{rel}(s,t))) \quad (5)$$

c_1, c_2 and c_3 are constants used to refine the $score_F$. For the Hungarian–Japanese language pair the values of *1*, *0.5* and *0.8*, provided the most accurate results.

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-based method selects a certain translation candidate two times based on adjectival and adverbial information in the Japanese-to-Hungarian direction, but doesn’t select it during the Hungarian-to-Japanese direction, the translation candidate

Table 3 Translation candidate scoring for 購入 *(kōnyū: buy, purchase)* (values above threshold in bold)

No.	Translation candidate	$score_F$	$score_C$				$score_D$				$score_E$	
			N	V	A	R	N	V	A	R	N	V
1	vétel	**2.012**	0.193	0.096	0	0	0	**0.500**	0	0	0.154	**0.500**
2	üzlet	**1.387**	0.026	0.030	0	0	0	0.250	0	0	0.020	0.077
3	hozam	1.348	0.095	0.071	0	0	0	0	0	0	0.231	0.062
4	emelőrúd	1.200	0.052	0.079	0	0	0	0	0	0	0.111	0.067
5	előny	1.078	0.021	0.020	0	0	0	0	0	0	0.054	0.056
6	támasz	1.053	0.014	0.015	0	0	0	0	0	0	0.037	0.031
7	vásárlás	0.818	0.153	**0.285**	0	0	0	0	0	0	0.273	0.200
8	szerzemény	0.771	0.071	**0.285**	0	0	0	0	0	0	0.136	0.200
9	könnyítés	0.771	0.064	**0.285**	0	0	0	0	0	0	0.136	0.200
10	emelőszerkezet	0.459	**0.285**	**0.285**	0	0	0	0	0	0	**0.429**	0.200

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 ($score_F$). $score_F$ also highly depends on the word entry,;therefore, local thresholds are used with this selection method too, with the thresholds being empirically set to $threshold_F = max(score_F)^*0.85$. Scores that pass this threshold are selected as translation candidates, regardless of their minimal value.

As an example, for the Japanese entry 購入 *(kōnyū: buy, purchase)* there are 10 possible Hungarian translations; using the methods based on synonymy, antonymy, and the hypernymy/hyponymy information 5 of them (#1, #7, #8, #9, #10) are selected as correct ones. Candidate #1 is selected as the only entry that passes the threshold set for the selection method, which uses antonymy information ($score_D$), when the entry is considered to be a verb (V). It is also selected by the method, which uses the hypernymy/hyponymy information ($score_E$). Candidates #7, #8, #9, #10 are selected as the best candidates using the synonymy information ($score_C$), when the entry is considered to be a verb. Moreover, candidate #10 is selected using two other selection methods as well ($score_C$, $score_E$), when it is considered to be a noun (N). Among these, only 1 of them (#1) is a correct translation, the rest have slightly similar or totally different meanings. However, with the combined scores the faulty translations were eliminated and a new, previously average scoring translation (#2) was selected (Table 3). 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 the dictionary, while the remaining three methods ('type C, D, and E') are used only indirectly for 'type F' selection (Table 2). With the described selection methods a dictionary with 48973 Japanese and 44664 Hungarian headwords were generated, totalling 187761 translation pairs.

5 Evaluation

As in many areas of natural language processing, with automatically generated bilingual dictionaries also 'recall' and 'precision' are the two most important evaluation criteria. 'Recall' indicates how many of the answers the system was able to recognize, while 'precision' shows the correctness of the answers it recognized. (Jurafsky and Martin, 2000). *F-score* is the harmonic mean of these two measures. In case of bilingual dictionary evaluation, recall is the ratio of the number of dictionary entries and the number of possible entries in a language; precision is the ratio of correct versus total number of translations.

$$\text{Recall} = \frac{\text{Number of correct answers given by system}}{\text{Total number of possible correct answers}} \tag{6}$$

$$\text{Precision} = \frac{\text{Number of correct answers given by system}}{\text{Number of answers given by system}} \tag{7}$$

$$F\text{-score} = \frac{2(\text{Precision} \times \text{Recall})}{\text{Precision} + \text{Recall}} \tag{8}$$

However, since there is no widely accepted evaluation method, it is usual for most developers to define their own criteria. As a result, the reported evaluation scores are difficult to interpret and needless to say they are almost impossible to use in comparative evaluation.

There are a few common problems in previous dictionary evaluation methods. First of all, precision evaluations are usually insufficient in their diversity and in their number of evaluated data (sample size). We believe that it is important to have multiple, diversified precision evaluation to fully understand the strong and weak points of a certain method.

Another problem is the inaccurate recall evaluation. It is well known that one of the most challenging aspects of dictionary generation is word ambiguity. It is relatively easy to automatically generate the translations of low-frequency headwords, because they tend to be less ambiguous. On the contrary, the ambiguity of the high-frequency words is much higher than their low-frequency counterparts, and as a result conventional methods fail to translate a considerable number of them. However, this discrepancy is not reflected in the traditional recall evaluation, since each word has an equal *weight*, regardless of its frequency of use. The methods verified by us managed to include in the generated dictionary words and expressions such as カソード (*kasōdo*; Hungarian: *katód*; English: 'cathode'), 感応作用 (*kannōsayō*; Hungarian: *adatelemzés*; English: 'induction, responsive effect'), or 鼓室 (*koshitsu*; Hungarian: *dobhártya*; English: 'eardrum'), but some of them failed to do so with 辞書 (*jisho*; Hungarian: *szótár*; English: 'dictionary'), 人 (*hito*; Hungarian: *ember*; English: 'man, human being'), or 食べる (*taberu*; Hungarian: *eszik*; English: 'to eat').

Considering into account the above-mentioned problems, we performed the following evaluations:

- frequency-weighted recall evaluation;
- 1-to-1 entry precision evaluation;
- 1-to-multiple entry evaluation.

For comparative purposes we also performed each type of evaluation for the two baseline methods. In order to do so, we re-implemented the methods proposed by Tanaka and Umemura, and Sjöbergh, using the same source dictionaries. With the Tanaka and Umemura method we managed to generate a Hungarian–Japanese dictionary with 105632 1-to-1 entries, size of which is comparable with our dictionary's size. Sjöbergh reports that 'with a threshold of 90% overlap well over 90% of the (...) words have a correct translation among the top ranked suggestions'. However, with the threshold set to 0.9 the number of 1-to-1 entries was only 25218, obviously at the recall's expense. Since a good recall is vital for a dictionary, especially if it's used as a translation resource, we lowered the threshold to retrieve a similar amount of translation pairs with our method's dictionary. Setting the threshold to 0.283 we managed to generate a dictionary with 187610 1-to-1 entries.

5.1 Recall evaluation

We argued that current recall evaluations don't reflect the true value of the dictionaries, because current evaluation methods don't consider the frequency in use of the entries. As a solution we automatically weighted each dictionary entry based on a Japanese frequency dictionary that we developed.

5.1.1 *Japanese frequency dictionary*

The EDR (Electronic Dictionary Research) corpus (Isahara, 2007) is an annotated corpus with 207360 sentences. It is considered to be a large-scale resource covering vocabulary used in general sentences, with high objectivity based on a large amount of texts. Its 15 part-of-speech categories cover 124071 unique words, with an average frequency of 39.6. Among these, 51.12% had only one 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 errors in the corpus.

A frequency-based score might be difficult to interpret as-is, but we believe that the score will be useful for comparative evaluation, the higher score pointing out the dictionary with the better recall.

Besides the recall value of our method's dictionary we calculated the values for the two baseline methods' dictionaries. We also calculated the recall value of the initial translation candidates to verify whether we managed to maximize the recall. Finally, a manually created Japanese–English dictionary's recall value was also computed for comparative purposes.

With frequency-weighted recall *(recall$_{weighted}$)* all entries *(w)* from the frequency dictionary *(F_D)* that have a frequency *(freq(w))* higher than one were verified whether they are translated or not in the verified dictionaries *(W_D)*. The lowest possible value is 0, when no word from the frequency dictionary is translated; the highest possible value is 100, when all entries from the frequency dictionary are translated.

$$\text{Recall}_{\text{weighted}} = \frac{\sum_{w \in W_D} freq(w)}{\sum_{w \in F_D} freq(w)} \qquad (9)$$

The frequency-weighted recall value results show that our method's dictionary (51.68) outscores every other automatically generated method's dictionary (37.03, 30.76) with a significant advantage. Moreover, our method's dictionary maintained the score of the initial translation candidates, therefore managing to maximize the recall value. No entry was lost, owing to the double selection method with local thresholds.

However, the recall value of the manually created dictionary is considerably higher than any automatically generated dictionary's value (Table 4).

5.2 1-to-1 precision evaluation

With 1-to-1 precision evaluation we determined the translation accuracy of our method, in the same time comparing it against the two baseline methods. During this process we considered 1-to-1 translation pairs, thus in case of entries that have multiple translations each translation was treated separately.

Table 4 Recall evaluation results

Dictionary	Recall$_{weighted}$
Our Method's Dictionary	51.68
Sjöbergh's Dictionary	37.03
Tanaka & Umemura's Dictionary	30.76
Initial Translation Candidates	51.68
Japanese–English Dictionary*	73.23

Asterisk indicates a manually created dictionary.

The manual scoring was performed by one of the authors, who is a native Hungarian and fluent in Japanese. Since no independent evaluator was available for these two languages, a blind evaluation was performed. We randomly selected 2000 translation pairs from each of the three Hungarian–Japanese dictionaries. After a random identification code being assigned to each of the 6000 selected translation pairs (2000 from each dictionary), they were mixed into a single sample data. As a result the evaluator did not know which method produced the translation pairs, thus influencing the score difference between the dictionaries was not possible. Only after manual scoring and regrouping based on the identification codes did the score for each dictionary became available.

The scoring criteria was the same as during selection type evaluation:

- **good** (○): the source entry and its translation convey the same meaning, or the meanings are slightly different, but in a certain context the translation is possible;
- **undecided** (?): the source entry and its translation's semantic value are similar, but a translation based on this entry would be faulty; and
- **erroneous** (×): the source entry and its translation convey different meanings.

Table 5 illustrates the evaluation standard. Examples #1 and #2 are labelled as 'good', since the same meanings are conveyed. In example #3 the Japanese entry has a more generalized meaning than its Hungarian counterpart, but since they share

Table 5 1-to-1 entry precision evaluation excerpts

No.	Code	Japanese entry	Hungarian entry	Classification
1	4t7y3b1p	地球儀 (chikyūgi: globe)	földgömb (globe)	good (◯)
2	8i4b8m8x	描く (kaku: to draw, to paint, to sketch)	fest (to paint)	good (◯)
3	6r3v7l1c	髭 (hige: mustache, beard, whisker, sideburn)	bajusz (mustache)	good (◯)
4	2b7b0q7j	グローヴ (gurōvu: baseball glove)	kesztyű (glove)	undecided (?)
5	7n2a7l3h	鳥 (tori: bird, fowl, poultry)	baromfi (fowl, poultry)	undecided (?)
6	9l0h6o1z	学校 (gakkō: school, educational institution)	halraj (school, a large group of fish)	erroneous (×)
7	5g7n2z9k	予約 (yoyaku: to reserve, to subscribe, to book)	könyv (book, volume)	erroneous (×)

Table 6 1-to-1 precision evaluation results

Dictionary	1-to-1 precision evaluation (%)		
	◯	?	×
Our Method's Dictionary	79.15	6.15	14.70
Sjöbergh's Dictionary	54.05	9.80	36.15
Tanaka & Umemura's Dictionary	62.50	7.95	29.55

a common meaning, the pair was marked as being 'good'. In fact, there is no word in Japanese, which perfectly corresponds with *bajusz* ('mustache') and there is no Hungarian word that describes 髭 *(hige:* 'mustache, beard, whisker, sideburn') in a single word. In examples #4 and #5 the source entries and their translations share similar meanings, but translations based on these entries could be misleading or incorrect. Thus, we marked them as 'undecided'. In examples #6 and #7 the source entries and their translations share common English translations, but these are homonyms, therefore they are 'erroneous'.

To rank the methods we only considered the correct ('good') translations. Our method performed best with 79.15%, outscoring Tanaka and Umemura's method's 62.50% and Sjöbergh's method's 54.05% (Table 6).

5.3 1-to-multiple evaluation

During 1-to-multiple evaluation we evaluated a sample of source entries together with all of their translations, attempting to determine the true reliability of the dictionary.

The scoring was performed manually by the same author. To eliminate any doubt of result manipulation, this evaluation was also blind. We randomly selected 2000 from the 48973 Japanese entries that appear in the initial translation candidates, together with their Hungarian translations from all three dictionaries. This resulted in three separate translation sets for each Japanese entry. Next, random identification numbers were assigned to each of the resulting 6000 entries for blind evaluation, after which the entries were mixed into a single sample data. We manually compared the meanings of each Japanese source entry with their Hungarian translations, based on the following criteria:

- **correct** (◯): all translations of the source entry are good;
- **similar** (Δ): the good translations are predominant, but there are up to 2 erroneous or undecided translations;
- **wrong** (×): the number of erroneous or undecided translations exceed 2;
- **missing** (–): the translation is missing.

For example, as illustrated in Table 7, all Hungarian translations of the Japanese entry 支持 *(shiji:* 'support, backing') are good, therefore the entry itself is marked as 'correct'. 思想 *(shisō:* 'thought, idea, ideology') produced one erroneous and one undecided translation, and as a result the entry is marked as 'similar'. Among the Hungarian translations of 壊す (*kowasu:* 'to break, to destroy, to smash, to ruin') there are correct translations as well, but since there are two erroneous ones, the entry is 'wrong'.

To rank the methods, we only considered the correct translations. Our method scored best with 71.45%, outperforming Sjöbergh's method's 61.65% and Tanaka and Umemura's method's 46.95%. The latter methods suffered because of the missing Hungarian translations, especially the Tanaka and

Table 7 1-to-multiple entry evaluation excerpts

Code	Japanese entry	Hungarian translation	Classification
8f0j9a6t	支持 (shiji: support, backing)	eltartás (support: ◯) segély (aid, assistance, support: ◯) támogatás (backing, favour, support: ◯)	correct (◯)
0a8v1q7j	思想 (shisō: thought, idea, ideology)	gondolat (idea, thought: ◯) érzés (feeling, idea: ×) ismeret (cognition, idea, knowledge: ◯) ötlet (idea: ◯) sejtés (conjecture, idea: ?)	similar (Δ)
1g7j8w4q	壊す (kowasu: to break, to destroy, to smash, to ruin)	letör (to break, to chip: ◯) leáll (to stall, to stop, to break down: ?) összeomlik (to collapse, to crack up: ×) megszakad (to break, to discontinue: ×)	wrong (×)

Table 8 1-to-multiple evaluation results

Dictionary	1-to-multiple evaluation (%)			
	◯	Δ	×	–
Our method's Dictionary	71.45	13.85	14.70	0
Sjöbergh's Dictionary	61.65	11.30	15.00	12.05
Tanaka & Umemura's Dictionary	46.95	3.35	9.10	40.60

Table 9 *F*-score results

Dictionary	*F*-score
Our method's Dictionary	62.53
Sjöbergh's Dictionary	43.94
Tanaka & Umemura's Dictionary	41.22

Umemura method. In fact, if we consider only the Japanese entries that the Tanaka and Umemura method managed to translate, the results are quite good. However, the high percentage of unrecalled (40.6%) Japanese entries is the Tanaka and Umemura methods's greatest disadvantage (Table 8).

5.4 *F*-score

We calculated the *F*-scores based on the weighted recall and 1-to-1 precision. Our dictionary proved to be the best overall with 62.50, against the Sjöbergh method's 43.93 and Tanaka and Umemura method's 41.22 (Table 9).

6 Discussions

Based on the recall evaluations, the traditional methods showed their major weakness by losing substantially from the initial recall values, determined 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 over the traditional methods, our method outscoring the other two methods with the 1-to-1 precision evaluation. 1-to-multiple evaluation was also the highest, proving that WordNet-based methods outperform dictionary-based methods.

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

6.1 Recall problems

We managed to maximize the recall of our initial translation candidates, but in many cases certain translation pairs still could not be generated because the linkage from the source language to the target language through the intermediate language is not present. There could be numerous reasons for this: (1) the entry is missing from at least one of the dictionaries; (2) the entries are present in the source-to-intermediate and target-to-intermediate dictionaries, but its translations are expressions, explanations or lexically too different; and (3) no direct translations.

The entries that could not be recalled are mostly the following type: (1) expressions; (2) inflected words; (3) rare words; and (4) words that are specific to the cultural aspect of the language (畳 *(tatami;* ‘floor-mat’*),* 寿司 *(sushi*; Japanese dish, sushi), or *gulyás* (‘goulash’) and so on); (5) words that are specific to the linguistic aspect of the language (words in Japanese that include opposite concepts: 上下 *(jōge;* ‘up and down’*),* 男女 *(danjo*; ‘man and woman, both sexes’; verbs with prefixes in Hungarian: *kiröpít (*‘to let fly’*), elfogyaszt (*‘to eat up, to use up’*)*; and others); and (6) certain parts-of-speech (ex: particle, auxiliary verb, and others in Japanese).

Another problem concerning recall is the high number of words not retrievable from Wordnet, which are not counted for scoring.

6.2 Precision problems

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 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 parts-of-speech, 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. For certain head words these dictionaries contain a great number of secondary meanings or even irrelevant translations, shadowing the main, more important senses. For example, our Hungarian–English dictionary contains the following translations for *ember*: ‘bleeder, man, men, mortal, number, person, soul, walla, wallah’. The closest to the original meaning would be ‘man or person’, but the presence of other, less representative translations, shifts the correct meaning. When this definition is expanded by synonyms (combined with the fact that WordNet uses a looser definition for synonyms), some irrelevant, but highly polysemous translations, such as ‘number’ expand the initial definition with more unrelated words, such as ‘figure, act, routine, turn, bit, numeral, issue’, etc. The resulting set of words make the identification of the correct meaning extremely difficult.

In other cases the resource dictionaries don’t contain translations of all meanings. Even the unambiguous ‘type A’ translations sometimes fail to produce the desired accuracy, although they are the unique candidate for a given word entry. For example, in our English-Japanese dictionary the English entry ‘loaf’ has only the meaning similar with ‘be lazy, hang around’, while in our English–Hungarian dictionary has only the meaning of ‘a shaped mass of baked bread’. This deficiency produced the supposedly unambiguous translation pair of the Japanese word ぶらぶら (*burabura:* ‘loaf’, ‘dangle’, ‘hang around’*)* and the Hungarian *vekni (*‘bread, loaf’), which is obviously erroneous. A secondary reason for this phenomenon is the meaning shift that occurs across Japanese, English, and Hungarian words.

Surprisingly ‘type A’ precision (‘lexically unambiguous’, 75.5%) proved to be lower than type B (‘sense classification’, 83.0%) or type F (‘combined score of synonymy, antonymy, and hypernymy’ 79.0%) precisions, proving that shifting the selection method from the dictionaries to the ontology is an efficient method for automatized dictionary generation (Table 2).

Other lexical databases of 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.

7 Conclusions and Future Plans

We proposed a new pivot language-based method to create bilingual dictionaries that can be used as translation resource for machine translation. Opposed to conventional methods that use dictionaries, our method uses WordNet as main resource of the intermediate language to select the suitable translation pairs. As a result, we eliminated most of the weaknesses caused by the structural differences of dictionaries, while profiting from the semantic relations provided by WordNet. We believe that because of the robust 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. With comparative evaluations we also proved that with the usage of a large lexical database, such as WordNet better results can be achieved than with dictionaries only.

Our future plans include improvement of our dictionary by means of manual correction. Besides manual supplementation of our dictionary with the currently missing translations or translation pairs, we will also examine whether significant dictionary improvement can be obtained with a community-based dictionary system that is currently under development. We will also examine whether a dictionary generated with our method can be used as a starting platform for a community-based online dictionary.

We also plan to verify the efficiency of our dictionary generation method by implementing it in our future Japanese–Hungarian machine translation system.

References

Agirre, E., Alegria, I., Rigau, G., and Vossen, P. (2007). MCR for CLIR. *Procesamiento Del Lenguaje Natural*, **38**: 3–15.

Babych, B., Hartley, A., and Sharoff, S. (2007). Translating from Under-Resourced Languages: Comparing Direct Transfer Against Pivot Translation. *In Proceedings of MT Summit XI*. Copenhagen, Denmark: European Association for Machine Translation (EAMT), pp. 29–35.

Ballesteros, L. A. (2000). Cross-Language Retrieval via Transitive Translation. *Advances in Information Retrieval: Recent Research from the CLIR*. Springer US **7**: 203–234.

Bick, E. and Nygaard, L. (2007). Using Danish as a CG Interlingua: A Wide-Coverage Norwegian-English Machine Translation System. *In Proceedings of the 16th Nordic Conference of Computational Linguistics*. Tartu, Estonia: North European Association for Language Technology (NEALT).

Bond, F. and Ogura, K. (2008). Combining Linguistic Resources to Create a Machine-Tractable Japanese-Malay Dictionary. *Language Resources and Evaluation*, **42**(2): 127–36.

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., and Roossin, P. (1998). A Statistical Approach to Language Translation. *In Proceedings of COLING-88*. Budapest, Hungary: International Committee on Computational Linguistics (ICCL), pp. 71–6.

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, USA: Computing Research Laboratory, New Mexico State University, pp. 111–8.

Clough, P. and Stevenson, M. (2004). Evaluating the Contribution of EuroWordNet and Word Sense Disambiguation to Cross-language Information Retrieval. *In Proceedings of the Second International Global WordNet Conference (GWC-2004)*. Brno, Czech Republic: Masaryk University, Brno, pp. 97–105.

Gollins, T. and Sanderson, M. (2001). Improving Cross Language Retrieval with Triangulated Translation. *In Proceedings of the 24th ACM SIGIR*. New Orleans, USA: Association for Computing Machinary (ACM), pp. 90–5.

Gonzalo, J., Verdejo, F., Peters, S., and Calzolari, N. (1998). Applying EuroWordNet to Cross-Language Text Retrieval. *Computers and the Humanities*, **32**: 185–207.

Isahara, H. (2007). EDR Electronic Dictionary – Present Status (EDR 電子化 辞書の現状). *NICT-EDR Symposium*. Tokyo, Japan: National Institute of Information and Communications Technology (NICT), pp. 1–14 (*in Japanese*).

Isahara, H., Bond, F., Uchimoto, K., Uchiyama, M., and Kanzaki, K. (2008). Development of Japanese

WordNet. *In Proceedings of LREC-2008*. Marrakech, Morocco: European Language Resources Association (ELRA), pp. 2420–2423.

Jurafsky, D. and Martin, J. (2000). Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition. *Prentice Hall Series in Artificial Intelligence*, **1**: 455–456.

Kay, M. and Röscheisen, M. (1993). Text-Translation Alignment. *Computational Linguistics*, **19**(1): 121–42.

Kraaij, W. (2003). Exploring Transitive Translation Methods. *In Proceeding of DIR 2003*. Amsterdam, Netherlands: Association for Computing Machinary (ACM).

Landau, S. I. (1974). Of Matters Lexicographical: Scientific and Technical Entries in American Dictionaries. *American Speech*, **49**(3–4): 241–4.

Lehtokangas, R. and Airio, E. (2002). Translation via A Pivot Language Challenges Direct Translation in Clir. *In Proceedings of the SIGIR2002 Workshop: Cross-Language Information Retrieval: A Research Roadmap*. Tampere, Finland: Association for Computing Machinary (ACM), pp. 23–28.

Miháltz, M. and Prószéky, G. (2004). Results and Evaluation of Hungarian Nominal WordNet v1.0. *In Proceedings of the Second Global WordNet Conference,* pp. 175–80.

Miller, G.A., Beckwith, R., Fellbaum, C., Gross, D., and Miller, K.J. (1990). Introduction to WordNet: An On-line Lexical Database. *International Journal of Lexicography*, **3**(4): 235–44.

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

Peters, W., Vossen, P., Díez-Orzas, P., and Adriaens, G. (1998). Cross-Linguistic Alignment of Wordnets with an Inter-Lingual-Index. *Comput Hum*, **32**: 221–51.

Prószéky, G., Miháltz, M., and Nagy, D. (2001). Toward a Hungarian WordNet. *In Proceedings of the NAACL 2001 Workshop on WordNet and Other Lexical Resources*. Carnegie Mellon University: Pittsburgh, USA.

Santiago, F., L. López, A., Galiano, M., Valdivia, M., and Vega, M. (2002). Intelligent Information Access Systems (SINAI) at CLEF 2001: Calculating Translation Probabilities with SemCor. *In Evaluation of Cross-Language Information Retrieval Systems.* Springer (editors: Carol Peters, Martin Braschler, Julio Gonzalo, Michael Kluck), pp. 185–192.

Sjöbergh, J. (2005). Creating a Free Japanese-English Lexicon. *In Proceedings of PACLING*. Tokyo, Japan: Meisei University, pp. 296–300.

Shirai, S. and Yamamoto, K. (2001). Linking English Words in Two Bilingual Dictionaries to Generate Another Pair Dictionary. *In ICCPOL-2001*. Seoul, Korea: Springer, pp. 174–9.

Stamou, S., Oflazer, K., Pala, K., Christoudoulakis, D., Cristea, D., Tufis, D., Koeva, S., Totkov, G., Dutoit, D., and Grigoriadou, M. (2002). Balkanet: A multilingual Semantic Network for Balkan Languages. *In Proceedings of the First International WordNet Conference*. Mysore, India: Central Institute of Indian Languages.

Tanaka, K. and Umemura, K. (1994). Construction of A Bilingual Dictionary Intermediated by A Third Language. *In Proceedings of COLING-94*. Kyoto, Japan: International Committee on Computational Linguistics (ICCL), pp. 297–303.

Voorhees, E. M. (1994). Query Expansion Using Lexical Semantic Relations. *In Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. Dublin, Ireland: Association for Computing Machinary (ACM), pp. 61–69.

Vossen, P. (1998). *A Multilingual Database with Lexical Semantic Networks*. Dordrecht: Kluwer Academic Publishers.

Wu, W. and Wang, H. (2007). Pivot Language Approach for Phrase-Based Statistical Machine Translation. *Machine Translation,* pp. 165–181.