Bilingual dictionary generation for low-resourced language pairs

Varga István
Yamagata University,
Graduate School of Science and Engineering
dyn36150@dip.yz.yamagata-u.ac.jp

Yokoyama Shoichi
Yamagata University,
Graduate School of Science and Engineering
yokoyama@yz.yamagata-u.ac.jp

Abstract

Bilingual dictionaries are vital resources in many areas of natural language processing. Numerous methods of machine translation require bilingual dictionaries with large coverage, but less-frequent language pairs rarely have any digitalized resources. Since the need for these resources is increasing, but the human resources are scarce for less represented languages, efficient automatized methods are needed. This paper introduces a fully automated, robust pivot language based bilingual dictionary generation method that uses the WordNet of the pivot 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. Our 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.

1 Introduction

In recent decades automatic and semi-automatic machine translation systems gradually managed 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, certain 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 highly accurate machine translation system switches from the choice and adaptation of the translation method to the problem of available translation resources between the chosen languages.

One possible solution is bilingual corpus acquisition for statistical machine translation (SMT). However, for highly accurate SMT systems large bilingual corpora are required, which are rarely available for less represented languages. Rule or sentence pattern based systems are an attractive alternative, for these systems the need for a bilingual dictionary is essential.

Our paper targets bilingual dictionary generation, a resource which can be used within the frameworks of a rule or pattern based machine translation system. Our goal is to provide a low-cost, robust and accurate dictionary generation method. Low cost and robustness are essential in order to be re-implementable with any arbitrary language pair. We also believe that besides high precision, high recall is also crucial in order to facilitate post-editing which has to be performed by human correctors. For improved precision, we propose the usage of WordNet, while for good recall we introduce a bidirectional selection method with local thresholds.

Our paper is structured as follows: first we overview the most significant related works, after which we analyze the problems of current dictionary generation methods. We present the details of our proposal, exemplified with the Japanese-Hungarian language pair. We evaluate the generated dictionary, performing also a comparative evaluation with two other pivot-language based methods. Finally we present our conclusions.

2 Related works

2.1 Bilingual dictionary generation

Various corpus based, statistical methods with very good recall and precision were developed starting from the 1980's, most notably using the

Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, pages 862–870,
Singapore, 6-7 August 2009. ©2009 ACL and AFNLP

Dice-coefficient (Kay & Röscheisen, 1993), *correspondence-tables* (Brown, 1997), or *mutual information* (Brown et al., 1998).

As an answer to the corpus-based method's biggest disadvantage, namely the need for a large bilingual corpus, in the 1990's Tanaka and Umemura (1994) presented a new approach. As a resource, they only use dictionaries to and from a pivot language to generate a new dictionary. These so-called 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. Tanaka and Umemura's method uses bidirectional source-pivot and pivot-target dictionaries (harmonized dictionaries). Correct translation pairs are selected by means of inverse consultation, a method that relies on counting the number of pivot language definitions of the source word, through which the target language definitions can be identified (Tanaka and Umemura, 1994).

Sjöbergh (2005) also presented an approach to pivot language based dictionary generation. When generating his English pivoted 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.

These two approaches described above are the best performing ones that are general enough to be applicable with other language pairs as well. In our research we used these two methods as baselines for comparative evaluation.

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 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 et al. (2001) proposed a method with multiple pivots (English and Kanji/Hanzi characters) to translate Sino-Korean entries. 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 (2007). Besides English, they also use Chinese as a second pivot.

2.2 Lexical database in lexical acquisition

Large lexical databases are vital for many areas in natural language processing (NLP), where large amount of structured linguistic data is needed. The appearance of WordNet (Miller et al., 1990) had a big impact in NLP, since not only did it provide 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 adopted the WordNet structure, thus creating a potential large multilingual network. 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).

Multilingual projects, such as EuroWordNet (Vossen 1998; Peters et al. 1998), Balkanet (Stamou et al. 2002) or Multilingual Central Repository (Agirre et al. 2007) aim to solve numerous problems in natural language processing. EuroWordNet was specifically designed for word disambiguation purposes in cross-language information retrieval (Vossen 1998). 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 is implemented (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 creating a dictionary with a large coverage. Secondly, these multilingual databases cover only a handful of languages, with Hungarian or Japanese not being part of them. Adding a new language would require the existence of a WordNet of that language.

3 Problems of current pivot language based methods

3.1 Selection method shortcomings

Previous pivot language based methods generate and score a number of translation candidates, and the candidate's scores that exceed a certain predefined global threshold are selected as viable translation pairs. However, the scores highly depend on the entry itself or the number of translations in the pivot language, therefore there is a variance in what that score represents. For this reason, a large number of good entries are entirely left out from the dictionary, because all of their translation candidates scored low, while faulty translation candidates are selected, because they exceed the global threshold. Due to this effect the recall value drops significantly.

3.2 Dictionaries not enough as resource

Regardless of the language pair, in most cases the meanings of the corresponding words 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 relevant overlapping in the source-to-pivot and target-to-pivot definitions.

Current methods perform a strictly lexical overlap of the source-pivot and target-pivot entries. Even if the meanings of the source and target head words are transferred to the pivot language, this is rarely done with the same set of words or definitions. Thus, due to the different word-usage or paraphrases, even semantically identical or very similar head words can have different definitions in different dictionaries. As a result, performing only lexical overlap, current methods cannot identify the differences between totally different definitions resulted by unrelated concepts, and differences in only nuances resulted by lexicographers describing the same concept, but with different words.

4 Proposed method

4.1 Specifics of our proposal

For higher precision, instead of the familiar lexical overlap of the current methods we calculate the semantically expanded lexical overlap of the source-to-pivot and target-to-pivot translations. In order to do that, we use semantic information extracted from the WordNet of the pivot language.

To improve recall, we introduce *bidirectional selection.* As we stated above, the global threshold eliminates a large number of good translation pairs, resulting in a low recall. As a solution, we can group the translations that share the same source or target entry, and set *local thresholds* for each head word. For example, for a source language head word *entry_source* there could be multiple target language candidates: *entry_target*$_1$*, … ,entry_target*$_n$. If the top scoring *entry_target*$_k$ candidates are selected, we ensure that at least one translation will be available for *entry_source*, 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. Local thresholds depend on the top scoring *entry_target*, being set to *maxscore·c*. Constant *c* varies between 0 and 1, allowing a small window not only for the maximum, but high scoring candidates as well. It is language and selection method dependent (see §5.1 for details).

4.2 Translation resources

As an example of a less-common language pair, we have chosen Japanese and Hungarian. For translation candidate generation, we have chosen two freely available dictionaries with English as the pivot language. The Japanese-English dictionary had 197282, while the Hungarian-English contained 189331 1-to-1 entry pairs. The Japanese-English dictionary had part-of-speech (POS) information as well, but to ensure robustness, our method does not use this information.

To select from the translation candidates, we mainly use *WordNet* (Miller et. al., 1990). From WordNet we consider four types of information: *sense categorization, synonymy*, *antonymy* and *semantic categories* provided by the tree structure of nouns and verbs.

4.3 Dictionary generation method

Our proposed method consists of two steps. In step 1 we generate a number of translation pair candidates, while in step 2 we score and select from them based on semantic information extracted from WordNet.

Step 1: translation candidate generation

Using the source-pivot and pivot-target dictionaries, we connect the source and target entries that share at least one common translation in the pivot language. We consider each source-target pair a *translation candidate*. With our Japanese-English and English-Hungarian dictionaries we accumulated 436966 Japanese-Hungarian translation candidates.

Step 2: translation pair selection

We examine the translation candidates one by one, looking up the source-pivot and target-pivot dictionaries, comparing the translations in the pivot language. There are six types of translations that we label *A*-*F* and explain below. First, we perform a strictly lexical match based only on the dictionaries. Next, using information extracted from WordNet we attempt to identify the correct translation pairs.

(a) Lexically unambiguous translation pairs

Some of the translation candidates have exactly the same translations into in the pivot language; we consider these pairs as being correct by default. Also among the translation candidates we identified a number of source entries that had only one target translation; and a number of target entries that had 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 (*type A* pairs).

(b) Using sense description

For most polysemous words WordNet has detailed descriptions with synonyms for each sense. We use these synonyms of WordNet's sense descriptions to disambiguate the meanings of the common translations. For a given source-target translation candidate (s,t) we look up the source-pivot and target-pivot translations ($s\rightarrow I=\{s\rightarrow i_1,\ldots,s\rightarrow i_n\}$ and $t\rightarrow I=\{t\rightarrow i_1,\ldots,t\rightarrow i_m\}$). We select the elements that are common in the two definitions ($I'=(s\rightarrow I)\cap(t\rightarrow I)$) and we look up their respective senses from WordNet ($sns(I')$). We identify the words' senses comparing each synonym in the WordNet's synonym description with each word from the dictionary definition. As a result, for each common word we arrive at a certain set of senses from the source-pivot definitions ($sns((s\rightarrow I')$) and a certain set of senses from the target-pivot definitions ($sns((t\rightarrow I')$). We mark $score_B(s,t)$ the maximum ratio of the identical and total number of identified senses (Jaccard coefficient). The higher the $score_B(s,t)$ is, the more probable is candidate (s,t) a valid translation.

$$score_B(s,t)=\max_{i'\in(s\rightarrow I)\cap(t\rightarrow I)}\frac{|sns(s\rightarrow i')\cap sns(t\rightarrow i')|}{|sns(s\rightarrow i')\cup sns(t\rightarrow i')|} \quad (1)$$

For example, *正解 (seikai: correct, right, correct interpretation)* and *helyes (correct, proper, right, appropriate)* have two common translations (I'={*right, correct*}), thus $score_B(s,t)$ can be performed with these two words. The adjective *right* has 13 senses according to WordNet, among them 4 were identified from the Japanese to English definition (*sns*(*right*)={#1, #3, #5, #10}, all identified through *correct*) and 5 from the Hungarian to English definition (*sns*(*right*)={#1, #3, #5, #6, #10}, through *correct* or *proper*). As a result, 4 senses are common, and 1 is different. Thus the adjective *right*'s score is 0.8 ($score_B(s,t)$[*right*](正解,*helyes*)). The adjective *correct* has 4 senses, all of them are recognized by both definitions through *right*, therefore the score through *correct* is 1 ($score_B(s,t)$[*correct*](正解,*helyes*)). The maximum of the above scores is the final score: $score_B(s,t)$(正解,*helyes*)=1.

All translation candidates are verified based on all four POS available from WordNet. Since synonymy information is available for nouns (N), verbs (V), adjectives (A) and adverbs (R), four separate scores are calculated for each POS.

Scores that pass a global threshold are considered correct. 33971 Japanese-Hungarian candidates (*type B* translations) were selected, with these two languages the global threshold was set to 0.1. Even this low value ensures that at least one of ten meanings is shared by the two entries of the pair, thus being suitable as translation pair.

(c) Using synonymy, antonymy and semantic categories

We expand the source-to-pivot and target-to-pivot definitions with information from WordNet (synonymy, antonymy and semantic category, respectively). Thus the similarity of the two expanded pivot language descriptions gives a better indication on the suitability of the translation candidate. Using the three relations, the common versus total number of translations (Jaccard coefficient) will define the appropriateness of the translation candidate.

$$score_{C,D,E}(s,t)=\frac{|ext(s\rightarrow i)\cap ext(t\rightarrow i)|}{|ext(s\rightarrow i)\cup ext(t\rightarrow i)|} \quad (2)$$

Since the same word or concept's translations into the pivot language also share the same semantic value, the extension with synonyms ($ext(l\rightarrow i)=(l\rightarrow i)\cup syn(l\rightarrow i)$, where $l=\{s,t\}$) the extended translation should share more common elements.

In case of antonymy, we expand the initial definitions with the *antonyms of the antonyms* ($ext(l\rightarrow i)=(l\rightarrow i)\cup ant(ant(l\rightarrow i))$, where $l=\{s,t\}$). This extension is different from the synonymy extension, in most cases the resulting set of words being considerably larger.

Along with synonymy, antonymy is also available for nouns, verbs, adjectives and adverbs, four separate scores are calculated for each POS.

Semantic categories are provided by the tree structure (hypernymy/hyponymy) of nouns and verbs of WordNet. We transpose each entry from the pivot translations to its semantic categories ($ext(l \rightarrow i) = \Sigma semcat(l \rightarrow i)$, where $l=\{s,t\}$). We assume that the correct translation pairs share a high percentage of semantic categories. Accordingly, the translations of semantically similar or identical entries should share a high number of common semantic categories.

The scores based on these relations highly depend on the number of pivot language translations; therefore we use the bidirectional selection method with local thresholds for each source and target head word. Local thresholds are set based on the best scoring candidate for a given entry. The thresholds were $maxscore \cdot 0.9$ for synonymy and antonymy; and $maxscore \cdot 0.8$ for the semantic categories (see §5.1 for details).

Using synonymy, 196775 candidate pairs (*type C*), with antonymy 99614 pairs (*type D*); while with semantic categories 195480 pairs (*type E*) were selected.

(d) Combined semantic information

The three separate lists of type C, D and E selection methods resulted in slightly different results, proving that they cannot be used as standalone selection methods (see §5.2 for details).

Because of the multiple POS labelling of numerous words in WordNet, many translation pairs can be selected up to four times based on separate POS information (noun, verb, adjective, adverb), all within one single semantic information based methods. Since we use a bidirectional selection method, experiments showed that translation pairs that were selected during both directions, in most cases were the correct translations. Similarly, translation pairs selected during only one direction were less accurate. 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. For every translation candidate we select the maximum score from the several POS (noun, verb, adjective and adverb for synonymy and antonymy relations; noun and verb for semantic category) based scores, multiplied by a multiplication factor (*mfactor*). The multiplication factor varies between 0 and 1, awarding the candidates that were selected both times during the double directional selection; and punishing when selection was made only in a single direction. The product gives the combined score ($score_F$), c_1, c_2 and c_3 are constants. In case of Japanese and Hungarian, these method scored best with the constants set to 1, 0.5 and 0.8, respectively. The combined score also highly depends on the word entry, therefore local thresholds are used in this selection method as well, which were empirically set to $maxscore \cdot 0.85$ (see §5.1 for details).

$$score_F(s,t) = \prod_{rel} \begin{pmatrix} (c_1 + \max(score_{rel}(s,t))) \cdot \\ (c_2 + c_3 \cdot mfactor_{rel}(s,t)) \end{pmatrix} \qquad (3)$$

As an example, for the Japanese entry 購入 *(kōnyū: buy, purchase)* there are 10 possible Hungarian translations; using the above methods 5 of them (#1, #7, #8, #9, #10) are selected as correct ones. Among these, only 1 of them (#1) is a correct translation, the rest have similar or totally different meanings. However, with the combined scores the faulty translations were eliminated and a new, correct, but previously average scoring translation (#2) was selected (Table 1).

#	translation candidate	*score_F*	*score_C*				*score_D*				*score_E*	
			N	V	A	R	N	V	A	R	N	V
1	vétel *(purchase)*	**2.012**	0.193	0.096	0	0	0	**0.500**	0	0	0.154	**0.500**
2	üzlet *(business transaction)*	**1.387**	0.026	0.030	0	0	0	0.250	0	0	0.020	0.077
3	hozam *(output, yield)*	1.348	0.095	0.071	0	0	0	0	0	0	0.231	0.062
4	emelőrúd *(lever, purchase)*	1.200	0.052	0.079	0	0	0	0	0	0	0.111	0.067
5	előny *(advantage, virtue)*	1.078	0.021	0.020	0	0	0	0	0	0	0.054	0.056
6	támasz *(purchase, support)*	1.053	0.014	0.015	0	0	0	0	0	0	0.037	0.031
7	vásárlás *(shopping)*	0.818	0.153	**0.285**	0	0	0	0	0	0	0.273	0.200
8	szerzemény *(attainment)*	0.771	0.071	**0.285**	0	0	0	0	0	0	0.136	0.200
9	könnyítés *(facilitation)*	0.771	0.064	**0.285**	0	0	0	0	0	0	0.136	0.200
10	emelőszerkezet *(lever)*	0.459	**0.285**	**0.285**	0	0	0	0	0	0	**0.429**	0.200

Table 1: Translation candidate scoring for 購入: *buy, purchase* (above thresholds in bold)

161202 translation pairs were retrieved with this method (*type F*).

During pre-evaluation *type A* and *type B* translations received a score of above 75%, while *type C*, *type D* and *type E* scored low (see §5.2 for details). 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.

With the described selection methods 187761 translation pairs, with 48973 Japanese and 44664 Hungarian unique entries was generated.

5 Threshold settings and pre-evaluation

5.1 Local threshold settings

As development set we considered all translation candidates whose Hungarian entry starts with "zs" (IPA: ʒ). We assume that the behaviour of this subset of words reflects the behaviour of the entire vocabulary. 133 unique entries totalling 515 translation candidates comprise this development set. After this, we manually scored the 515 translation candidates as *correct* (the translation conveys the same meaning, or the meanings are slightly different, but in a certain context the translation is possible) or *wrong* (the translation pair's two entries convey a different meaning). The scoring was performed by one of the authors who is a native Hungarian and fluent in Japanese. 273 entries were marked as *correct*. Next, we experimented with a number of thresholds to determine which ones provide with the best F-scores (Table 2). The F-scores were determined as follows: for example using synonymy information (type C) in case of threshold=0.85%, 343 of the 515 translation pairs were above the threshold. Among these, 221 were marked as correct by our manual evaluator, thus the precision being 221/343·100=64.43 and the recall being 221/273·100=80.95. F-score is the harmonic mean of precision and recall (71.75 in this case).

selection type	**threshold value** (%)				
	0.75	**0.80**	**0.85**	**0.90**	**0.95**
C	70.27	70.86	71.75	**72.81**	66.95
D	69.92	70.30	70.32	**70.69**	66.66
E	73.71	**74.90**	72.52	71.62	65.09
F	78.78	79.07	**79.34**	78.50	76.94

Table 2: Selection type F-scores with varying thresholds (best threshold values in bold)

5.2 Selection method evaluation

As a pre-evaluation of the above selection methods, we randomly selected 200 1-to-1 source-target entries resulted by each method. The same evaluator scored the translation pairs as *correct* (the translation conveys the same meaning, or the meanings are slightly different, but in a certain context the translation is possible), *undecided* (the translation pair's semantic value is similar, but a translation based on them would be faulty) or *wrong* (the translation pair's two entries convey a different meaning).

selection type	**evaluation score** (%)		
	correct	*undecided*	*wrong*
A	75.5	6.5	18
B	83	7	10
C	68	5.5	26.5
D	60	9	31
E	71	5.5	23.5
F	79	5	16

Table 3: Selection type evaluation

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).

6 Evaluation

We performed three types of evaluation:
(1) frequency-weighted recall evaluation
(2) 1-to-1 entry precision evaluation
(3) 1-to-multiple entry evaluation

For comparative purposes we also performed each type of evaluation for two other pivot language based methods whose characteristics permit to be implementable with virtually any language pair. In order to do so, we constructed two other Hungarian-Japanese dictionaries using the methods proposed by Tanaka & Umemura and Sjöbergh, using the same source dictionaries.

6.1 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 keywords, 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. As a result, we performed a frequency weighted recall evaluation. We used a Japanese frequency dictionary (F_D) generated from the Japanese EDR corpus (Isahara, 2007) to weight each Japanese entry. Setting the standard to the frequency dictionary (its recall value being 100), we automatically search for each entry (w) from the frequency dictionary, looking whether or not it is included in the bilingual dictionary (W_D). If it is recalled, we weight it with its frequency from the frequency dictionary.

$$recall_w = \frac{\sum_{w \in W_D} frequency(w)}{\sum_{w \in F_D} frequency(w)} \cdot 100 \qquad (4)$$

method	recall
our method	51.68
Sjöbergh method	37.03
Tanaka method	30.76
initial candidates	51.68
Japanese-English(*)	73.23

Table 4: Recall evaluation results (* marks a manually created dictionary)

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, it maintains the score of the initial translation candidates, therefore managing to maximize the recall value, owing to the bidirectional selection method with local thresholds. However, the recall value of a manually created Japanese-English dictionary is higher than any automatically generated dictionary's value (Table 4).

6.2 1-to-1 precision evaluation

With 1-to-1 precision evaluation we determine the translation accuracy of our method, compared with the two baseline methods. 200 random pairs were selected from each of the three Hungarian-Japanese dictionaries, scoring them manually the same way as with selection type evaluation (*correct*, *undecided*, *wrong*) (Table 5). 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, after a random identification code being assigned to each of the 600 selected translation pairs (200 from each dictionary), they were mixed. Therefore the evaluator did not know the origin of the translation pairs, only after manual scoring the total score for each dictionary was available, after regrouping based on the initial identification codes. The process was repeated 10 times, 2000 pairs were manually checked from each dictionary.

code	Japanese entry	Hungarian entry	classification
k9g6n5d8	報告 (hōkoku: information, report)	hír (report, information, news)	*correct*
j8h0k1x5	初 (ubu: innocent, naive)	zöld (green, verdant)	*undecided*
a5b6n8i3	エントリ (entori: entry <a contest>)	bejárat (entry, entrance)	*wrong*

Table 5: 1-to-1 precision evaluation examples

method	evaluation score (%)		
	correct	*undecided*	*wrong*
our method	79.15%	6.15%	14.70%
Sjöbergh method	54.05%	9.80%	36.15%
Tanaka method	62.50%	7.95%	29.55%

Table 6: 1-to-1 precision evaluation results

To rank the methods we only consider the *correct* translations. Our method performed best with an average of 79.15%, outscoring Tanaka method's 62.50% and Sjöbergh method's 54.05% (Table 6). The maximum deviance of the *correct* translations during the 10 repetitions was less than 3% from the average.

6.3 1-to-multiple evaluation

While with 1-to-1 precision evaluation we estimated the accuracy of the translation pairs, with 1-to-multiple we calculate the true reliability of the dictionary, with the initial translation candidates set as recall benchmark. When looking up the meanings or translations of a certain head word, the user, whether he's a human or a machine, expects all translations to be accurate. Therefore we evaluated 200 randomly selected Japanese entries from the initial translation candidates, together with all of their Hungarian translations, scoring them as *correct* (all translations are correct), *acceptable* (the good translations are predominant, but there are up to 2 erroneous translations), *wrong* (the number or wrong translations exceeds 2) or *missing* (the translation is missing) (Table 7).

The same type of mixed, manual evaluation was performed by the same author on samples of 200 entries from each Japanese-Hungarian dictionary. This evaluation was also repeated 10 times.

To rank the methods, we only consider the *correct* translations. Our method scored best with

71.45%, outperforming Sjöbergh method's 61.65% and Tanaka method's 46.95% (Table 8).

code	Japanese entry	Hungarian translations	classification
j4h8m9x5	圧縮 (asshuku: compression, squeeze)	összenyomás (compression, crush, squeeze: *correct*) összeszorítás (compression, confinement: *correct*) zsugorítás (shrinkage: *correct*)	*correct*
h9j913v1	底面 (teimen: base)	alap (base, bottom, foundation: *correct*) alapzat (base, bed, bottom: *correct*) lúg (alkali, base: *undecided*) támpont (base: *correct*)	*acceptable*
l0k6m3n7	鳴らす (narasu: to sound, to ring, to beat)	bekerít (to encircle, to enclose, to ring: *wrong*) cseng (to clang, to clank, to ring, to tinkle: *correct*) hangzik (to ring, to sound: *correct*) horkan (to snort: *wrong*) üt (to bang, to knock, to ring: *wrong*)	*wrong*

Table 7: 1-to-multiple entry evaluation examples

method	**evaluation score** (%)			
	correct	*acceptable*	*wrong*	*missing*
our method	71.45	13.85	14.70	0
Sjöbergh method	61.65	11.30	15.00	12.05
Tanaka method	46.95	3.35	9.10	40.60

Table 8: 1-to-many evaluation results

7 Discussion

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 maintains 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 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 through the pivot language, while precision problems discuss the reasons why erroneous pairs are produced.

7.1 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 link from the source language to the target language through the pivot language simply doesn't exist. The main reasons are: the entry is missing from at least one of the dictionaries; translations in the pivot language are expressions or explanations; or there is no direct translation or link between the source and target entries. The entries that could not be recalled are mostly expressions, rare entries, words specific to a language (ex: *tatami: floor-mat,* or *gulyás: goulash*).

Moreover, a number of head words don't have any synonym, antonym and/or hypernymy/hyponymy information in WordNet, and as a result these words could not participate in the type B, C, D, E and F scoring.

7.2 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 pivot 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 POS to handle. Because semantically they are more flexible than other POS categories, and the meaning range is also highly flexible across languages, the identification of the correct translation is increasingly difficult. For this reason, the number of faulty translations and the number of meanings that are not translated was relatively high.

One other source of erroneous translations is the quality of the initial dictionaries. Even the unambiguous *type A* translations fail to produce the desired accuracy, although they are the unique candidate for a given word entry. The main 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. In other cases the resource dictionaries don't contain translations of all meanings; homonyms are present as pivot entries with different meanings, sometimes creating unique, but faulty links.

8 Conclusions

We proposed a new pivot language based method to create bilingual dictionaries that can be used as translation resource for machine translation. In contrast to conventional methods that use dictionaries only, our method uses WordNet as a main resource of the pivot 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 WordNet. We believe that because of the nature of our method it can be re-implemented with most language pairs.

In addition, owing to features such as the bidirectional selection method with local thresholds we managed to maximize recall, while maintaining a precision which is better than any other compared method's score. During exemplification, we generated a mid-large sized Japanese-Hungarian dictionary with relatively good recall and promising precision.

The dictionary is freely available online (http://mj-nlp.homeip.net/mjszotar), being also downloadable at request.

References

Agirre, E., Alegria, I., Rigau, G, Vossen, P. 2007. MCR for CLIR, *Procesamiento del lenguaje natural* 38, pp 3-15.

Bond, F., Ogura, K. 2007. Combining linguistic resources to create a machine-tractable Japanese-Malay dictionary, *Language Resources and Evaluation*, 42(2), pp. 127-136.

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, *Proceedings of COLING-88*, pp. 71-76.

Brown, R.D. 1997. Automated Dictionary Extraction for Knowledge-Free Example-Based Translation, *Proceedings of the 7th International Conference on Theoretical and Methodological Issues in Machine Translation*, pp. 111-118.

Isahara, H., Bond, F., Uchimoto, K., Uchiyama, M., Kanzaki, K. 2008. Development of Japanese WordNet, *Proceedings of LREC-2008*.

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), pp. 121-142.

Miháltz, M., Prószéky, G. 2004. Results and Evaluation of Hungarian Nominal WordNet v1.0, *Proceedings of the Second Global WordNet Conference*, pp. 175-180.

Miller G.A., Beckwith R., Fellbaum C., Gross D., Miller K.J. (1990). Introduction to WordNet: An Online 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, *NLPRS-2001*, pp. 63-70, Tokyo, Japan.

Peters, W., Vossen, P., Díez-Orzas, P., Adriaens, G. 1998. Cross-linguistic Alignment of Wordnets with an Inter-Lingual-Index, *Computers and the Humanities* 32, pp. 221–251.

Prószéky, G., Miháltz, M., Nagy, D. 2001. Toward a Hungarian WordNet, *Proceedings of the NAACL 2001 Workshop on WordNet and Other Lexical Resources*, Pittsburgh, June 2001.

Sjöbergh, J. 2005. Creating a free Japanese-English lexicon, *Proceedings of PACLING*, pp. 296-300.

Shirai, S., Yamamoto, K. 2001. Linking English words in two bilingual dictionaries to generate another pair dictionary, *ICCPOL-2001*, pp. 174-179.

Stamou, S., Oflazer, K., Pala, K., Christoudoulakis, D., Cristea, D., Tufiş, D., Koeva, S., Totkov, G., Dutoit, D., Grigoriadou, M. 1997. BalkaNet: A Multilingual Semantic Network for the Balkan Languages, *In Proceedings of the International Wordnet Conference*, Mysore, India.

Tanaka, K., Umemura, K. 1994. Construction of a bilingual dictionary intermediated by a third language, *Proceedings of COLING-94*, pp. 297-303.

Vossen, P. 1998. Introduction to EuroWordNet. *Computers and the Humanities* 32: 73-89 Special Issue on EuroWordNet.