Low Cost Construction of a Multilingual Lexicon from Bilingual Lists

Lian Tze Lim, Bali Ranaivo-Malançon, and Enya Kong Tang

Abstract—**Manually constructing multilingual translation lexicons can be very costly, both in terms of time and human effort. Although there have been many efforts at (semi-)automatically merging bilingual machine readable dictionaries to produce a multilingual lexicon, most of these approaches place quite specific requirements on the input bilingual resources. Unfortunately, not all bilingual dictionaries fulfil these criteria, especially in the case of under-resourced language pairs. We describe a low cost method for constructing a multilingual lexicon using only simple lists of bilingual translation mappings. The method is especially suitable for under-resourced language pairs, as such bilingual resources are often freely available and easily obtainable from the Internet, or digitised from simple, conventional paper-based dictionaries. The precision of random samples of the resultant multilingual lexicon is around 0.70–0.82, while coverage for each language, precision and recall can be controlled by varying threshold values. Given the very simple input resources, our results are encouraging, especially in incorporating under-resourced languages into multilingual lexical resources.**

Index Terms—**Lexical resources, multilingual lexicon, under-resourced languages.**

I. Introduction

Multilingual translation lexicons are very much desired in many natural language processing (NLP) applications, including multilingual machine translation and cross-lingual information retrieval, but are very costly to construct manually. On the other hand, given the abundance of bilingual machine readable dictionaries (MRD), there have been many efforts at (semi-)automatically merging these bilingual lexicons into a sense-distinguished multilingual lexicon [1]–[3].

Many of these approaches require the input bilingual lexicons to include certain types of information besides equivalents in the target language, such as gloss or definition text in the source language and domain field codes. Unfortunately, bilingual lexicons with such features are not always available, especially for under-resourced language pairs. Nor are the delineation or granularity of different sense entries indicated clearly or consistently. More often than not, the lowest common denominator across bilingual lexicons is just a simple list of mappings from a source language word to one or more target language equivalents.

Manuscript received November 2, 2010. Manuscript accepted for publication January 22, 2011.

The authors are with the Natural Language Processing Special Interest Group, Faculty of Information Technology, Multimedia University, Malaysia (e-mail: liantze@gmail.com, {ranaivo, enyakong}@mmu.edu.my).

English	**Chinese**	**Malay**	**French**
factory plant	工厂	loji kilang	fabrique manufacture usine

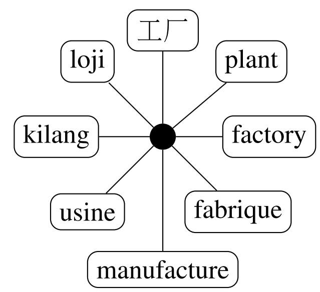

Fig. 1. Example multilingual lexicon entry for the concept *industrial plant* with lexical items from English, Chinese, Malay and French.

We aim to bootstrap a multilingual translation lexicon, given the simplest bilingual dictionaries taking the form of simple lists of bilingual translations. Such low resource requirements (as well as the low-cost method that will be described) is especially suitable for under-resourced language pairs. We first give a brief overview of the overall structure of the multilingual lexicon in Section II. Section III describes how an initial trilingual lexicon can be generated from bilingual ones, and how further languages can be added. Initial experimental results presented in Section IV show that our method is capable of generating a usable multilingual dictionary from simple bilingual resources without the need for rich information types, such as those mentioned in Section V.

II. Multilingual Lexicon Organisation

Each entry in our multilingual lexicon is similar to a *translation set* described by Sammer and Soderland [4] as 'a multilingual extension of a WordNet synset [5]' and contains 'one or more words in each k languages that all represent the same word sense'. Unlike Sammer and Soderland's translation sets, however, our lexicon entries currently do not include any gloss or contexts to indicate the intended word sense.

Figure 1 shows an example translation set entry representing the concept *industrial plant*, containing English '*factory*' and '*plant*'; Chinese '工厂' (*gōngchǎng*); Malay '*loji*' and '*kilang*'; French '*fabrique*', '*manufacture*' and '*usine*'.

Internally, each translation set is accessed by a language-independent axis node, with language-specific lexicalisations connected to it, similar to the structural scheme used in the multiligual extension of the Lexical Markup Framework [6] and the Papillon Multilingual Dictionary [7]. Our multilingual lexicon is thus capable of handling lexical gaps (when a concept is not lexicalised in a language) as well as diversification phenomena (when a word sense in a language is more specific than its translation in another language). Nevertheless, for our current experiment, we will allow diversified meanings to be connected directly to the same axis.

III. Building the Lexicon

Our bootstrapping algorithm first generates trilingual translation triples based on the one-time inverse consultation (OTIC) procedure [8], which was proposed to generate translation lexicons for new language pairs from existing bilingual lexicons. These triples are then merged to produce the translation sets in our multilingual lexicon. New languages are added by producing translation triples containing the new language and languages already present in our multilingual lexicon, then merging the new triples into the existing entries by detecting common translation pairs.

A. One-time Inverse Consultation

Tanaka, Umemura and Iwasaki [8] first proposed OTIC to generate a bilingual lexicon for a new language pair L_1–L_3 via an intermediate language L_2, given existing bilingual lexicons for language pairs L_1–L_2, L_2–L_3 and L_3–L_2. Following is an example of a OTIC procedure for linking Japanese words to their Malay translations via English:

- For every Japanese word, look up all English translations ($\mathbb{E}_1$).
- For every English translation, look up its Malay translations ($\mathbb{M}$).
- For every Malay translation, look up its English translations ($\mathbb{E}_2$), and see how many match those in $\mathbb{E}_1$.
- For each $m \in \mathbb{M}$, the more matches between $\mathbb{E}_1$ and $\mathbb{E}_2$, the better m is as a candidate translation of the original Japanese word.

$$\text{score}(m) = 2 \times \frac{|\mathbb{E}_1 \cap \mathbb{E}_2|}{|\mathbb{E}_1| + |\mathbb{E}_2|}$$

A worked example is shown in Figure 2. The Japanese word '印' (*shirushi*) has 3 English translations, which in turn yields another three Malay translations. Among them, '*tera*' has 4 English translation, 2 of which are also present in the earlier set of 3 English translations. The one-time inverse consultation score for '*tera*' is thus $2 \times \frac{2}{3+4} = 0.57$, and indicates '*tera*' is the most likely Malay translation for '印'.

Bond et. al. [10] extended OTIC by linking through two languages, as well as utilising semantic field code and classifier information to increase precision, but these extensions may not always be possible as not all lexical resources include these information (nor do all languages use classifiers).

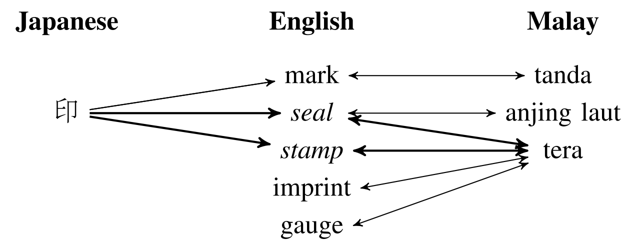

Fig. 2. Using OTIC, Malay '*tera*' is determined to be the most likely translation of Japanese '印' as they are linked by the most number of English words in both directions, with score('*tera*') $= 2 \times \frac{2}{3+4} = 0.57$. (Diagram taken from [9, Figure 1])

B. Extension to OTIC

OTIC was originally conceived to produce a list of bilingual translations for a new language pair. As our aim is a multilingual lexicon instead, we modified the OTIC procedure to produce trilingual translation triples and translation sets, as outlined in Algorithm 1.

Algorithm 1 allows partial word matches between the 'forward' and 'reverse' sets of intermediate language words. For example, if the 'forward' set contains '*coach*' and the reverse set contains '*sports coach*', the modified OTIC score is $\frac{1}{2} = 0.5$, instead of 0. This would also serve as a likelihood measure for detecting diversification in future improvements of the algorithm. The score computation for (w_h, w_t) is also adjusted accordingly to take into account this substring matching score (line 10), as opposed to the exact matching score in the original OTIC.

We retain the intermediate language words along with the 'head' and 'tail' languages, i.e. the OTIC procedure will output translation *triples* instead of pairs. α and β on line 14 are threshold weights to filter translation triples of sufficiently high scores. Bond et. al. [10] did not discard any translation pairs in their work; they left this task to the lexicographers who preferred to whittle down a large list rather than adding new translations. In our case, however, highly suspect translation triples must be discarded to ensure the merged multilingual entries are sufficiently accurate. Specifically, the problem is when an intermediate language word is polysemous. Erroneous translation triples (w_h, w_m, w_t) may then be generated (with lower scores), where the translation pair (w_h, w_m) does not reflect the same meaning as (w_m, w_t). If such triples are allowed to enter the merging phase, the generated multilingual entries would eventually contain words of different meanings from the various member languages: for example, English '*bold*', Chinese '黑体' (*hēitǐ* bold typeface) and Malay '*garang*' (fierce) might be placed in the same translation set by error.

As an example, consider the (w_h, w_m, w_t) translation triples with non-zero scores generated by OTIC where w_h = '*garang*',

Algorithm 1: Generating trilingual translation chains

for all lexical items $w_h \in L_1$ **do**
  $\mathbb{W}_m \leftarrow$ translations of w_h in L_2
  for all $w_m \in \mathbb{W}_m$ **do**
    $\mathbb{W}_t \leftarrow$ translations of w_m in L_3
    for all $w_t \in \mathbb{W}_t$ **do**
      Output a translation triple (w_h, w_m, w_t)
      $\mathbb{W}_{m_r} \leftarrow$ translations of w_t in L_2
      $\text{score}(w_h, w_m, w_t) \leftarrow \sum_{w \in \mathbb{W}_m} \frac{\text{number of common words in } w_{m_r} \in \mathbb{W}_{m_r} \text{ and } w}{\text{number of words in } w_{m_r} \in \mathbb{W}_{m_r}}$
    end for
    $\text{score}(w_h, w_t) \leftarrow 2 \times \frac{\sum_{w \in \mathbb{W}_m} \text{score}(w_h, w, w_t)}{|\mathbb{W}_m| + |\mathbb{W}_{m_r}|}$
  end for
  $X \leftarrow \max_{w_t \in \mathbb{W}_t} \text{score}(w_h, w_t)$
  for all distinct translation pairs (w_h, w_t) **do**
    if $\text{score}(w_h, w_t) \geq \alpha X$ or $(\text{score}(w_h, w_t))^2 \geq \beta X$ **then**
      Place $w_h \in L_1$, $w_m \in L_2$, $w_t \in L_3$ from all triples $(w_h, w_{\ldots}, w_t)$ into same translation set
      Record $\text{score}(w_h, w_t)$ and $\text{score}(w_h, w_m, w_t)$
    else
      Discard all triples $(w_h, w_{\ldots}, w_t)$
    end if
  end for
end for ▷ The sets are now grouped by (w_h, w_t)
Merge all translation sets containing triples with same (w_h, w_m)
Merge all translation sets containing triples with same (w_m, w_t)

presented in Figure 3. The highest $\text{score}(w_h, w_t)$ is 0.143. When $\alpha = 0.8$ and $\beta = 0.2$, (w_h, w_t) pairs whose score is less then $\alpha \times 0.143 = 0.1144$, or whose score squared is less then $\beta \times 0.143 = 0.0286$ will be discarded. Therefore, triples containing (garang, 大胆) (and other pairs of lower scores) will be discarded as its score 0.111 and squared score 0.0123 are lower than both threshold values.

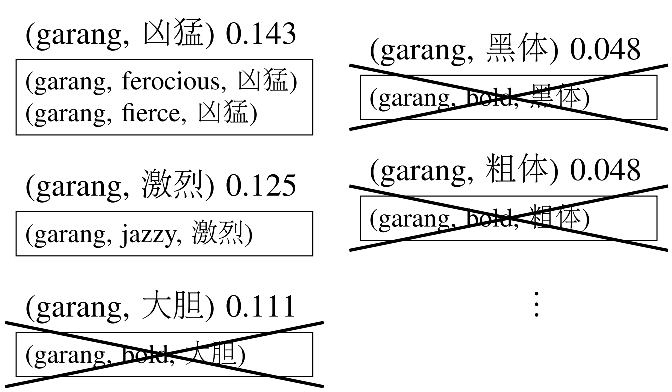

Fig. 3. Generated translation triples from Algorithm 1

The retained translation triples are then merged into translation sets based on overlapping translation pairs among the languages. An example is shown in Figure 4, where the translation triples are merged into one translation set with five members.

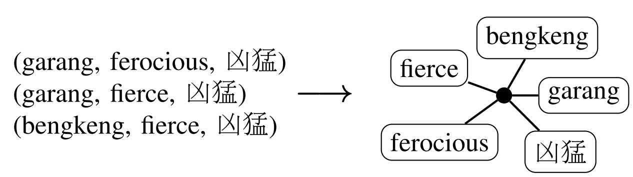

Fig. 4. Merging translation triples into translation sets

C. Adding More Languages

The algorithm described in the previous section gives us a trilingual translation lexicon for languages $\{L_1, L_2, L_3\}$. Algorithm 2 outlines how a new language L_4, or more generally, L_{k+1} can be added to an existing multilingual lexicon of languages $\{L_1, L_2, \ldots, L_k\}$. We first run OTIC to produce translation triples for L_{k+1} and two other languages already included in the existing lexicon. These new triples are then compared against the existing multilingual entries. If two words in a triple are present in an existing entry, the third word is added to that entry as well.

Figure 5 gives such an example: given the English–Chinese–Malay translation set earlier, we prepare translation triples for French–English–Malay. By detecting overlapping English–Malay translation pairs in the translation set and

Algorithm 2: Adding L_{k+1} to multilingual lexicon $\mathbb{L}$ of $\{L_1, L_2, \ldots, L_k\}$

$T \leftarrow$ translation triples of L_{k+1}, L_m, L_n generated by Algorithm 1 where $L_m, L_n \in \{L_1, L_2, \ldots, L_k\}$
for all $(w_{L_m}, w_{L_n}, w_{L_{k+1}}) \in T$ **do**
  Add $w_{L_{k+1}}$ to all entries in $\mathbb{L}$ that contains both w_{L_m} and w_{L_n}
end for

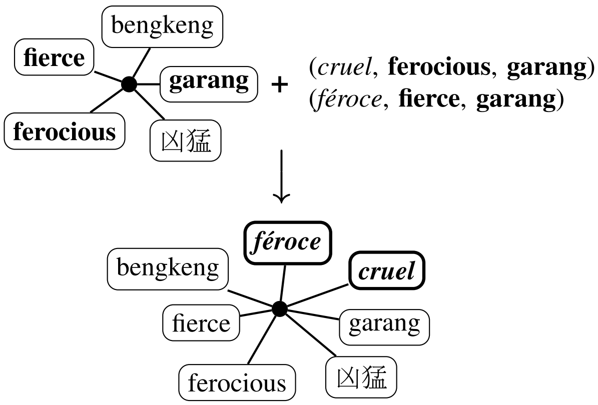

Fig. 5. Adding French members to existing translation sets

triples, two new French words '*cruel*' and '*féroce*' are added to the existing translation set.

D. Resources for Experiment

We generated a multilingual lexicon for Malay, English and Chinese using the modified OTIC procedure, with English as the intermediate language. We used the following bilingual dictionaries as input:

- *Kamus Inggeris–Melayu untuk Penterjemah*, an English to Malay dictionary published by PTS Professional Publishing. The vast majority of Malay glosses in this dictionary are single words, or simple phrases containing only a few words. We therefore reversed the direction and used it as a Malay to English dictionary.
- *XDict*, a free English to Chinese Dictionary packaged for GNU/Linux distros, including Ubuntu and Debian.
- *CC-CEDICT*[1], a free Chinese to English dictionary. We omitted Chinese lexical items marked to be archaic, idioms and family names. As CC-CEDICT entries do not include a part-of-speech (POS) field, we assigned one to each entry–gloss pair by running the Stanford POS Tagger[2] on the English glosses.

We normalised English entries with respect to American and British spelling variances[3], as well as stripping off the verb infinitive '*to*'. Chinese entries were normalised by stripping off the adjective marker '的'. (See [9] for other normalisation possibilities.)

[1] http://cc-cedict.org/wiki/start
[2] http://nlp.stanford.edu/software/tagger.shtml
[3] http://wordlist.sourceforge.net/

To add French to the generated Malay–English–Chinese lexicon, we converted entries from FeM, a French–English–Malay dictionary[4], into translation triples with default scores of 1.0.

We provided a look-up interface to the resultant multilingual lexicon, using which users can look up a word in any member languages. All multilingual entries containing the word being looked up will be returned, with the words inside each entry being ranked by their associated OTIC scores. Figure 6 shows the look-up results for Malay '*kebun*'.

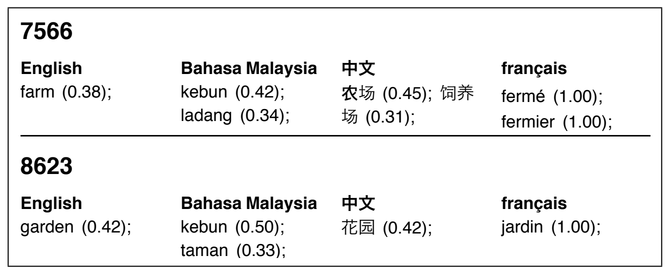

Fig. 6. Multilingual lexicon look-up result for Malay '*kebun*'

IV. Results and Discussion

As the correct addition of French lexical items depends on the accuracy of the Malay–English–Chinese lexicon generated, and also because it was harder for us to find evaluators who speak all four languages, only the Malay–English–Chinese entries are evaluated.

In general, precision increases for greater threshold values of α and β, at the expense of less words in each language being included. Our procedure produced more translation sets which should have been merged (false negatives) when α and β are high; however this is more desirable than words of different meanings being placed in the same translation set (false positives).

We performed two evaluations on the generated multilingual lexicon, described in the following subsections.

A. Evaluation on 100 Random Translation Sets

For the first evaluation, we randomly selected 100 translation sets constructed from at least two translation triples, using different α and β values. Evaluators were told to only accept as accurate translation sets in which all member Malay, English and Chinese words are (near-)synonyms. For this initial work, a translation set is deemed accurate if it contains

[4] http://www-clips.imag.fr/cgi-bin/geta/fem/fem.pl?lang=en

diversified word meanings, i.e. it is acceptable for both Malay '*beras*' (uncooked rice) and '*nasi*' (cooked rice) to occur in the same translation set as English '*rice*'. The evaluation results are summarised in Figure 7.

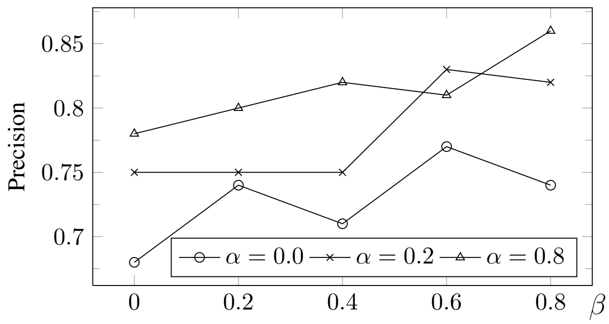

Fig. 7. Precision for 100 randomly selected translation sets with varying α and β.

Precision increases with α and β, but is generally in the range of 0.70–0.82, and can go up to as high as 0.86. Of the erroneous sets, most of the wrongly included words are not the top-ranked ones in each language, especially when α and β are high. Many errors are caused by incorrect POS assignments to *CEDICT* entries. Nevertheless, we find such results encouraging, particularly because it can be achieved with such simple bilingual translation mapping lists.

B. Evaluation on Test Word Samples

As mentioned earlier near the end of section III-B, translation sets generated using OTIC are most prone to error when the intermediate language (English in our experiment) word is polysemous, thereby selecting a 'tail' language word that does not have the same meaning as the 'head' language word.

To evaluate how effective OTIC is at detecting polysemy in the intermediate language, we selected four polysemous English words as test words, namely '*bank*', '*plant*', '*target*' and '*letter*'. We define a list of gold standard translation sets for each test word, based on all possible generated triples from our input dictionaries. All translation sets containing the test words are then retrieved. By viewing generation of translation sets as a data clustering problem, we access their accuracy by calculating the F_1 score and Rand index (RI) [11] for each list of retrieved translation sets $R = \{R_1, R_2, \ldots, R_m\}$ for a test word against that test word's golden standard $A = \{A_1, A_2, \ldots, A_n\}$:

$$\begin{aligned}
\text{TP} &= |\{\text{word pairs occurring in some } R_i \in R \text{ and some } A_j \in A\}| \\
\text{FP} &= |\{\text{word pairs occurring in some } R_i \in R \text{ but not in any } A_j \in A\}| \\
\text{TN} &= |\{\text{word pairs not occurring in any } R_i \in R \text{ nor any } A_j \in A\}| \\
\text{FN} &= |\{\text{word pairs not occurring in any } R_i \in R \text{ but in some } A_j \in A\}|
\end{aligned}$$

$$\begin{aligned}
\text{Precision} &= \frac{\text{TP}}{\text{TP} + \text{FP}} \\
\text{Recall} &= \frac{\text{TP}}{\text{TP} + \text{FN}} \\
F_1 &= \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \\
\text{RI} &= \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{FN} + \text{TN}}
\end{aligned}$$

Table I, figures 8 and 9 summarise the results for each test word. Here again, RI and F_1 increase with α and β. We also note from the graphs that the threshold β is more influential in raising both RI and F_1. However, the scores may decrease if α and β are too high, as is the case for '*plant*' when $\alpha = 0.8$ and $\beta \geq 0.4$. This is due to valid words being rejected by the high thresholds, thereby increasing the number of false negatives (FN) and lowering RI and F_1. Taking into account the Rand indices, F_1 scores as well as word coverage in each language, we found taking $\alpha \approx 0.6$ and $\beta \approx 0.2$ to offer a reasonable balance between precision, recall and coverage.

TABLE I
MINIMUM AND MAXIMUM RAND INDEX AND F_1 SCORE FOR EACH TEST WORD

Test word	Rand Index min	Rand Index max	F_1 min	F_1 max	Min. thresholds for best score α	β
'*bank*'	0.417	0.611	0.588	0.632	0.6	0.4
'*plant*'	0.818	0.927	0.809	0.913	0.6	0.2
'*target*'	0.821	1.000	0.902	1.000	0.4	0.2
'*letter*'	0.709	0.818	0.724	0.792	0.8	0.2

V. RELATED WORK

There have been many efforts to create lexical databases similar to the Princeton English WordNet [5] for other languages. To leverage the many types of rich data and resources built on top of Princeton WordNet, many such projects aim to align their entries to those in the Princeton WordNet. Notable wordnet projects include EuroWordNet (Western European languages) [3], BalkaNet (Eastern European languages) [2], and many more[5]. All these wordnets taken together can be regarded as a huge multilingual lexicon, with the Princeton Wordnet as its main hub. However, this also means these wordnets tend to suffer from a frequent critique against the Princeton WordNet: its overly fine sense distinctions often cause human lexicographers and evaluators working with the wordnets much confusion, as well as complicating NLP applications that make use of them.

Sammer and Soderland [4] constructed PanLexicon, a multilingual lexicon by computing context vectors for words of different languages from monolingual corpora, then grouping the words into translation sets by matching their

[5]see http://www.globalwordnet.org/gwa/wordnet_table.htm

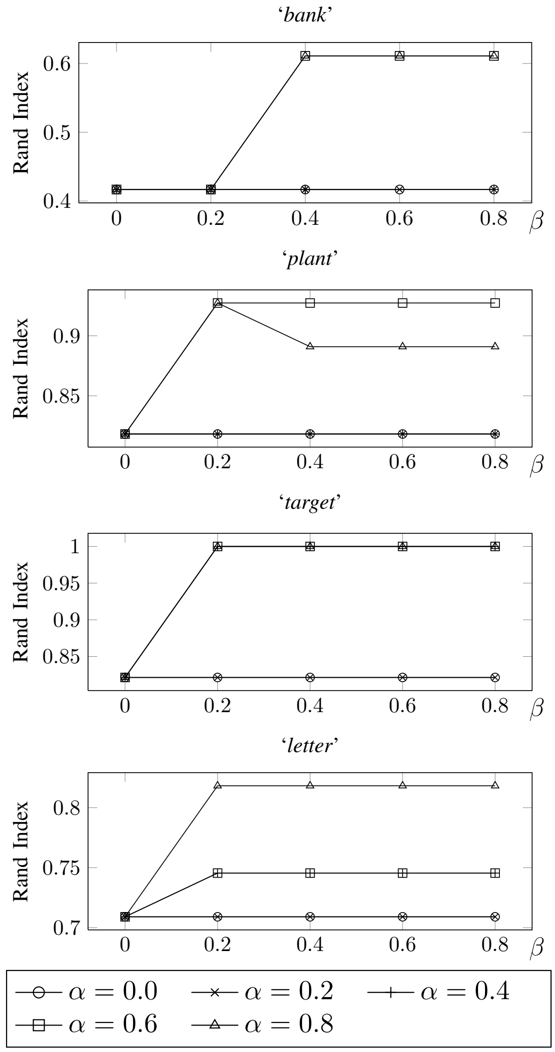

Fig. 8. Rand indices for translation sets containing '*bank*', '*plant*', '*target*' and '*letter*' with varying thresholds α and β.

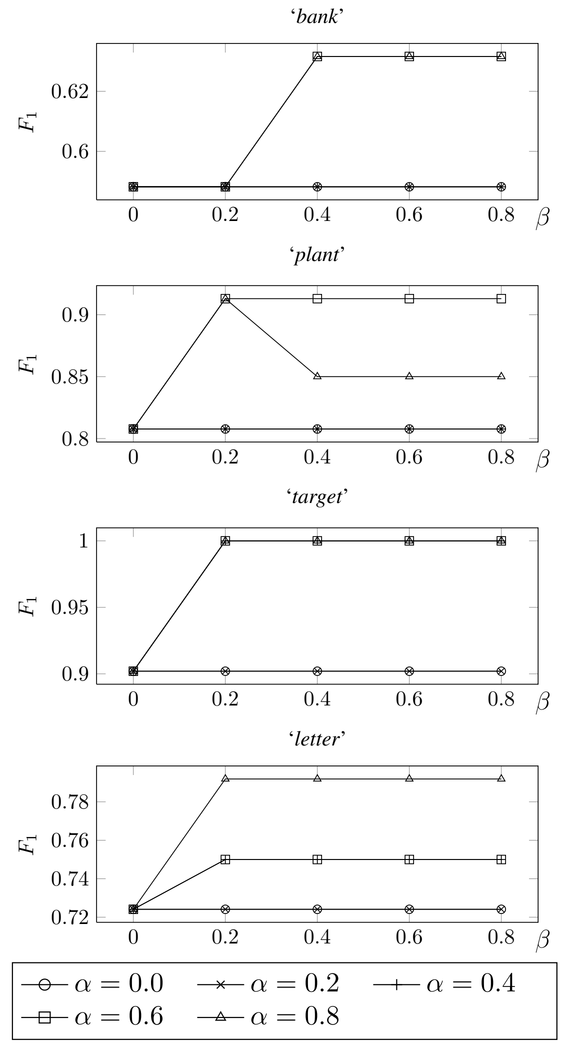

Fig. 9. F_1 scores for translation sets containing '*bank*', '*plant*', '*target*' and '*letter*' with varying thresholds α and β.

context vectors with the help of bilingual lexicons. By using a corpus-based method, good coverage of words from different languages is expected. In addition, sense distinctions are derived from corpus evidence, which are unlikely to be as fine as those of Princeton WordNet. However, their method produces many translation sets that contain semantically related but not synonymous words, e.g. '*shoot*' and '*bullet*', thus lowering the precision: the authors report 44 % precision based on evaluators' opinions (75 % if inter-evaluator agreement is not required). In addition, specific methods for identifying multi-word expressions (MWEs) in the corpus are required (which was not taken into consideration in their paper), whereas our method would also process MWEs if they are listed in the bilingual lexicons.

Markó, Schulz and Hahn [12] made use of cognate mappings to derive new translation pairs, later validated by processing parallel corpora in the medical domain. Due to the special characteristics of medical terms, each complex term is indexed on the level of sub-words, e.g. '*pseudo*$\oplus$*hypo*$\oplus$*para*$\oplus$*thyroid*$\oplus$*ism*'. The authors report up to 46 % accuracy for each language pair by checking against data from the Unified Medical Language System (UMLS). The biggest drawback in their approach is the requirement for large aligned thesaurus corpora, although such resources may be more readily available for specific domains such as medicine. Also, the cognate-based approach would not be applicable for language pairs that are not closely related.

Lafourcade [1] also uses a vector-based model for populating the Papillon multilingual dictionary [7]. Instead of constructing context vectors from corpora, Lafourcade computes conceptual vectors for each translation pair from a bilingual dictionary, based on the gloss text (written in the source language) and associated class labels from a semantic hierarchy. Translation pairs of different language pairs are then compared based on their conceptual vectors to determine if they express the same meaning. By using class labels as the vector space generator, the conceptual vector model is able to merge dictionary entries whose gloss text contain synonymous words. It does, however, require the class labels to be assigned to the dictionary entries. Such resources are not always available, and the additional task of assigning class labels is time-consuming and costly.

VI. Conclusion and Future Work

We have described a low cost procedure for constructing a multilingual lexicon using only simple bilingual translation lists, suitable especially for including under-resourced languages in lexical resources. Precision of random samples of the generated translation sets averages in the range of 0.70–0.82. Based on the experimental Rand indices and F_1 scores for selected lexical samples, we found threshold values of $\alpha \approx 0.6$ and $\beta \approx 0.2$ give reasonable balance between precision, recall and word coverage.

Manually validating and correcting an automatically constructed lexicon, entry by entry, can be very costly both in time and human expertise. We plan to take another approach, by deploying the bootstrapped multilingual lexicon in a machine translation system and capturing user actions when they edit the translation to update the lexicon entries.

Acknowledgment

The work reported in this paper is supported by a Fundamental Research Grant (FRGS/1/10/TK/MMU/02/02) from the Malaysian Ministry of Higher Education. We thank the evaluators who participated in the results evaluation, and the two anonymous reviewers for their comments on improving this paper.

References

[1] M. Lafourcade, "Automatically populating acception lexical database through bilingual dictionaries and conceptual vectors," in *Proceedings of PAPILLON-2002*, Tokyo, Japan, 8 2002.

[2] D. Tufiş, D. Cristeau, and S. Stamou, "BalkaNet: Aims, methods, results and perspectives – a general overview," *Romanian Journal of Information Science and Technology Special Issue*, vol. 7, no. 1, pp. 9–43, 2004.

[3] P. Vossen, "EuroWordNet: A multilingual database of autonomous and language-specific wordnets connected via an Inter-Lingual-Index," *Special Issue on Multilingual Databases, International Journal of Linguistics*, vol. 17, no. 2, 2004.

[4] M. Sammer and S. Soderland, "Building a sense-distinguished multilingual lexicon from monolingual corpora and bilingual lexicons," in *Proceedings of Machine Translation Summit XI*, Copenhagen, Denmark, 2007, pp. 399–406.

[5] C. Fellbaum, Ed., *WordNet: An Electronic Lexical Database*, ser. Language, Speech, and Communication. Cambridge, Massachusetts: MIT Press, 1998.

[6] G. Francopoulo, N. Bel, M. George, N. Calzolari, M. Monachini, M. Pet, and C. Soria, "Multilingual resources for NLP in the lexical markup framework (LMF)," *Language Resources and Evaluation*, vol. 43, no. 1, pp. 57–70, 3 2009.

[7] C. Boitet, M. Mangeot, and G. Sérasset, "The PAPILLON project: Cooperatively building a multilingual lexical database to derive open source dictionaries & lexicons," in *Proceedings of the 2nd Workshop on NLP and XML (NLPXML'02)*, 2002, pp. 1–3.

[8] K. Tanaka, K. Umemura, and H. Iwasaki, "Construction of a bilingual dictionary intermediated by a third language," *Transactions of the Information Processing Society of Japan*, vol. 39, no. 6, pp. 1915–1924, 1998, in Japanese.

[9] F. Bond and K. Ogura, "Combining linguistic resources to create a machine-tractable Japanese–Malay dictionary," *Language Resources and Evaluation*, vol. 42, pp. 127–136, 2008.

[10] F. Bond, b. S. Ruhaida, T. Yamazaki, and K. Ogura, "Design and construction of a machine-tractable Japanese-Malay dictionary." in *Proceedings of MT Summit VIII*, Santiago de Compostela, Spain, 2001, pp. 53–58.

[11] W. M. Rand, "Objective criteria for the evaluation of clustering methods," *Journal of the American Statistical Association*, vol. 66, no. 336, pp. 846–850, 1971.

[12] K. Markó, S. Schulz, and U. Hahn, "Multilingual lexical acquisition by bootstrapping cognate seed lexicons," in *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP) 2005*, Borovets, Bulgaria, 2005.