Language Resources and Evaluation manuscript No.
(will be inserted by the editor)

Lexicon+TX: Rapid Construction of a Multilingual Lexicon with Under-Resourced Languages

Lian Tze Lim · Lay-Ki Soon · Tek Yong Lim · Enya Kong Tang · Bali Ranaivo-Malançon

Received: date / Accepted: date

Abstract Most efforts at automatically creating multilingual lexicons require input lexical resources with rich content (e.g. semantic networks, domain codes, semantic categories) or large corpora. Such material is often unavailable and difficult to construct for under-resourced languages. In some cases, particularly for some ethnic languages, even unannotated corpora are still in the process of collection. We show how multilingual lexicons with under-resourced languages can be constructed using simple bilingual translation lists, which are more readily available. The prototype multilingual lexicon developed comprise six member languages: English, Malay, Chinese, French, Thai and Iban, the last of which is an under-resourced language in Borneo. Quick evaluations showed that 91.2 % of 500 random multilingual entries in the generated lexicon require minimal or no human correction.

Keywords Multilingual lexicon · Under-resourced languages

PACS 1.100 · 1.300

L. T. Lim
School of Engineering, Science and Technology, KDU College Penang, 32 Jalan Anson, 10400 Georgetown, Penang, Malaysia
Tel.: +604-226-6368, Fax: +604-228-0362
E-mail: liantze@gmail.com

L. T. Lim · L.-K. Soon · T. Y. Lim
Faculty of Computing and Informatics, Multimedia University, Persiaran Multimedia, 63100 Cyberjaya, Selangor, Malaysia
E-mail: liantze@gmail.com, {lksoon,tylim}@mmu.edu.my

E. K. Tang
Linton University College, Persiaran UTL, Bandar Universiti Teknologi Legenda, Batu 12, 71700 Mantin, Negeri Sembilan, Malaysia
E-mail: enyakong1@gmail.com

B. Ranaivo-Malançon
Faculty of Computer Science and Information Technology, Universiti Malaysia Sarawak, 94300 Kota Samarahan, Sarawak, Malaysia
E-mail: mbranaivo@fit.unimas.my

1 Introduction

As lexical resources are usually costly to construct by hand from scratch, a 'draft' copy is usually automatically acquired from existing resources. Much work has been done on automatic data acquisition of multilingual lexicons, but often require input lexical resources with rich information fields. These may be semantic networks (Verma and Bhattacharyya, 2003); domain codes and semantic labels (Jalabert and Lafourcade, 2002; Bond and Ogura, 2008); definition texts (Janssen, 2004); or even existing multilingual lexicons (Mausam et al, 2009).

Unfortunately, such comprehensive resources may not be available for all languages. Berment (2004) categorised human languages into three categories, based on their digital 'readiness' or presence in cyberspace and software tools:

- **'tau'-languages:** totally-resourced languages, from French *très bien dotés*,
- **'mu'-languages:** medium-resourced languages, from French *moyennement dotées*,
- **'pi'-languages:** under-resourced languages, from French *peu dotées*.

Berment's work involved creating the digital foundations for developing human language technologies for under-resourced languages, such creating fonts, input methods and segmentation tools for Khmer, Burmese and Lao. However, lexical resources for under-resourced languages, especially publicly avaialble ones, are still lacking. At the time of writing, Wiktionary contains only 828 Khmer, 585 Lao, 469 Burmese word entries, compared to 501,171 English word entries.[1]

Apart from coverage, it would also be impractical to assume that dictionaries of under-resourced languages contain rich information fields such as domain code, semantic relations, or even well-written gloss texts. In the worst case, the sole field present may only be a list of translation equivalents in a target language.

Other projects attempt to mine translation equivalents from corpora (Tufiş et al, 2004; Sammer and Soderland, 2007; Dorow et al, 2009), which may be more readily available than specialised dictionaries. For under-resourced languages, however, a sizeable corpus may still be difficult to obtain or create.

We show how a preliminary version of a multilingual lexicon can be constructed using simple bilingual translation lists, which are more easily available. Section 2 reviews a selection of multilingual lexicon projects, while section 3 outlines our multilingual lexicon data acquisition methodology. The development and quick evaluation of a prototype is described and compared to related work in section 4. A brief discussion is presented in section 5, before concluding in section 6.

2 Multilingual Lexical Databases

Many multilingual lexicon projects have been developed as befits their importance in natural language processing (NLP) applications. A considerable number of such efforts are multilingual extensions of the Princeton WordNet (Fellbaum, 1998), such as MultiWordNet (Pianta et al, 2002), EuroWordNet (Vossen, 2004), BalkaNet (Tufiş et al, 2004) and Universal Multilingual WordNet (de Melo and Weikum, 2009), amongst

[1] Based on `https://en.wiktionary.org/wiki/Wiktionary:Statistics` (June 2013).

others. As wordnet-based lexical databases use the Princeton WordNet hierarchy (or an extension of it) as the common index, aggregated multilingual lookups can be easily performed, as provided by the Open Multilingual Wordnet[2] (Bond and Paik, 2012) and BabelNetXplorer[3] (Navigli and Ponzetto, 2012). Due to the easy availability of Princeton WordNet and rich lexical information, wordnet-based projects are very well developed, with as many as over 1,500,000 words in over 200 languages in the Universal Multilingual Wordnet.

A frequent critique against wordnet-based multilingual lexical databases is the fine-grained sense distinctions, which may not always be desirable in all NLP applications. Aligning translation equivalents from new languages to the correct senses must be precise and accurate, which can only be facilitated if rich resources, such as those mentioned in the previous section, are available (see also Sammer and Soderland, 2007; Mausam et al, 2009; Varga et al, 2009; as well as sections 4.2 and 4.3 in this paper).

In addition, due to such fine sense granularity, human contributors and evaluators would likely need to have a higher level of linguistic expertise to participate effectively in the projects. This may limit the number of eligible human contributors, thereby hampering efforts to build lexical resources, especially for under-resourced languages. In contrast, other projects use a crowd-sourcing approach, such as Papillon (Boitet et al, 2002; Mangeot-Lerebours et al, 2003), JeuxDeMots (Lafourcade, 2007) and the Arabic preterminological database (Daoud et al, 2009). Volunteers without professional linguistic training may collaboratively contribute entries to multilingual lexicons. This would lower the requirement barrier to higher participation from laypersons, which is important for building resources for under-resourced languages.

3 Building Lexicon+TX

Lexicon+TX (for **T**ranslation and cross(**X**)-lingual lookup) is a multilingual lexicon, the purpose of which is to connect under-resourced languages to richer-resourced languages by providing translation equivalents from different languages. Entries in Lexicon+TX are organised as multilingual translation sets. Each translation set corresponds to a coarse-grained concept, and is accessed by a language-independent axis node. Translation sets are similar to Sammer and Soderland's (2007) data structures of the same name: 'a multilingual extension of a WordNet synset (Fellbaum, 1998)' and contains 'one or more lexical items (LI) in each k languages that all represent the same word sense'. Synonyms from different languages are connected to the axis, structurally similar to the scheme used in the multilingual extension of Lexical Markup Framework (Francopoulo et al, 2009) and the Papillon project (Boitet et al, 2002; Mangeot-Lerebours et al, 2003).

Multilingual translation sets can be bootstrapped from simple lists of bilingual translations, which are easier for native speakers to provide, or extracted from existing bilingual dictionaries. A modified version of the one-time inverse consultation (OTIC)

[2] `http://www.casta-net.jp/~kuribayashi/multi/`
[3] `http://lcl.uniroma1.it/babelnet/`

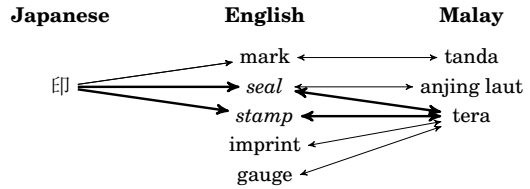

Fig. 1 Using OTIC, Malay «tera» is determined to be the most likely translation of Japanese «印» as they are linked by the highest number of English words in both directions, with score(«tera») $= 2 \times \frac{2}{3+4} = 0.57$. (Diagram from Bond and Ogura, 2008)

procedure proposed by Tanaka et al (1998) is then applied to generate a multilingual lexicon.

3.1 One-Time Inverse Consultation (OTIC)

Tanaka et al (1998) first proposed the OTIC procedure to generate a bilingual dictionary for a new language pair L_1–L_3 via an intermediate language L_2, given existing bilingual dictionaries for language pairs L_1–L_2, L_2–L_3 and L_3–L_2. Following is an example of an OTIC procedure for linking Japanese words to their Malay translations via English (Figure 1):

1. For every Japanese word (e.g. «印»), look up all its English translations ($\mathbb{E}_f$), i.e. {«mark», «seal», «stamp»}.
2. For every English translation $e \in \mathbb{E}_f$, look up its Malay translations ($\mathbb{M}$). For example, $\mathbb{M}$ for «seal» is {«anjing laut», «tera»}.
3. For every Malay translation $m \in \mathbb{M}$, look up its English translations ($\mathbb{E}_r$). $\mathbb{E}_r$ for «tera» is {«seal», «stamp», «imprint», «gauge»}.
4. For each $m \in \mathbb{M}$, the more matches between $\mathbb{E}_f$ and $\mathbb{E}_r$, the better m is as a candidate translation of the original Japanese word, computed by the Dice coefficient of $\mathbb{E}_f$ and $\mathbb{E}_r$, i.e. $\text{score}(m) = 2 \times \frac{|\mathbb{E}_f \cap \mathbb{E}_r|}{|\mathbb{E}_f| + |\mathbb{E}_r|}$.
   In Figure 1, the Malay candidate «tera» has the most overlap between $\mathbb{E}_f$ and $\mathbb{E}_r$ ({«seal», «stamp»}), and therefore the highest score at 0.57. «tera» is thus the most probable Malay translation of «印».

OTIC can be extended further (Bond et al, 2001) by linking through two languages, as well as utilising semantic field codes and classifier information to increase precision, but these measures may not always be possible as not all lexical resources include these information fields (nor do all languages use classifiers).

3.2 Extending OTIC to Generate Trilingual Translation Sets

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 follows:

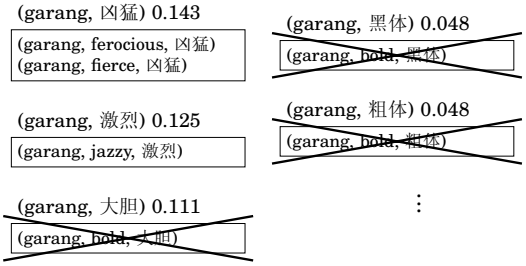

Fig. 2 Filtering Malay–English–Chinese translation triples

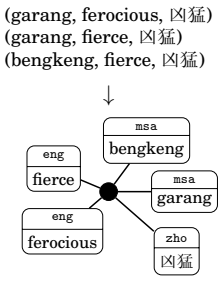

Fig. 3 Merging translation triples into translation sets

- Retain words from intermediate language (L_2), in addition to 'head' (L_1) and 'tail' (L_3) languages to get trilingual translation triples.
- Delete triples with scores below the local filtering threshold, which is the product of a threshold parameter α and the maximum OTIC score for all triples starting with the same 'head' language LI. For example, when $\alpha = 0.8$, all triples with scores lower than $0.8 \times 0.143 = 0.114$ in Figure 2 are deleted.
- Merge triples with common $(w_{L_1}, w_{L_2}, \ldots)$ and $(\ldots, w_{L_2}, w_{L_3})$ segments into the same translation set (Figure 3).

Bond et al (2001) 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 dubious 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 sets would eventually contain words of different meanings from the various member languages (e.g. the rejected triples in Figure 2): for example, English «bold», Chinese «黑体» (*hēitǐ*, 'bold typeface') and Malay «garang» ('fierce') might be placed in the same translation set by error.

3.3 Adding New Languages

The algorithm described in the previous section produces a trilingual lexicon for languages $\{L_1, L_2, L_3\}$. A new language L_4, or more generally, L_{k+1} can be added to an existing multilingual lexicon of languages $\{L_1, L_2, \ldots, L_k\}$ by processing more bilingual dictionaries. OTIC is first run to produce translation triples for L_{k+1} and two other languages already included in the existing Lexicon+TX. These new triples are then compared against the existing multilingual entries. If two words in a triple are present in an existing translation set, the third word is added to that set as well. Figure 4 shows how new French items «féroce» and «cruel» can be added to Lexicon+TX after it has been populated with English, Malay and Chinese members.

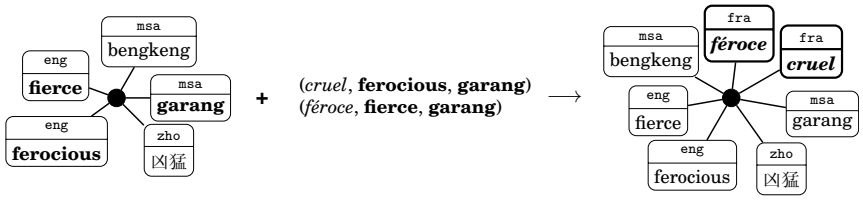

Fig. 4 Adding French members to existing translation sets

3.4 Lexicon Maintenance

Once a draft copy of Lexicon+TX has been created, maintenance is relatively straightforward and would consist of the following main operations, based on a human judge's evaluation of a translation set (see also section 4.2):

- merging translation sets;
- deleting entire translation sets;
- deleting a member LI from a translation set;
- adding a member LI to a translation set;
- splitting one translation set into more sets.

When the original input dictionaries are updated, the changes may be propagated Lexicon+TX. If new entries are added to the original input dictionaries, new translation triples can be generated and added to exiting translation sets. However, there is currently no good way of propagating deletions of entries and translation equivalence from the input dictionaries to Lexicon+TX.

4 Prototype Implementation and Evaluation

A prototype of Lexicon+TX comprising six languages (English, Malay, Chinese, French, Iban and Thai) has been constructed from six bilingual and one trilingual dictionaries, as outlined below. We plan to make the developed prototype data available for research by arrangement with the authors.

4.1 Lexicon+TX Construction using Bilingual Dictionaries

The following dictionaries were used as input, choosing open-source and free options wherever possible:

- SiSTeC-EMDict (Part of the SiSTeC-EBMT machine translation system (Boitet et al, 2011). 94,604 Malay items; 82,342 English items)
- Kamus Inggeris-Melayu Dewan (Johns, 2000) (37,618 English items, 56,368 Malay items.)
- XDict[4] (177,799 English items; 194,571 Chinese items)

[4] `http://packages.debian.org/sid/text/dict-xdict`

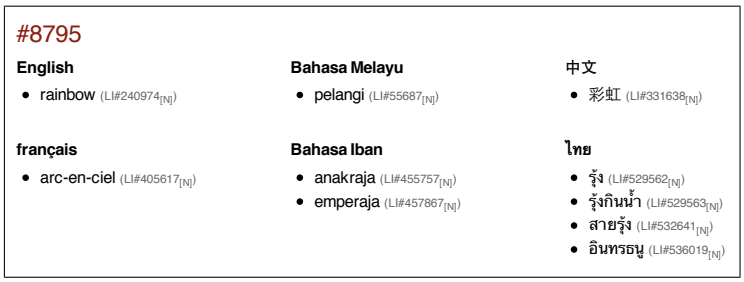

Fig. 5 Sample translation set for «rainbow»

Table 1 Number of Lexicon+TX LIs connected to new languages

Source Language	No. of LIs with new translations in target languages				No. of translation sets
	$\geq$ 2 langs.	$\geq$ 3 langs.	$\geq$ 4 langs.	5 langs.	
English	24,371	11,244	7,696	3,912	37,611
Chinese	13,226	9,023	6,044	2,774	15,562
Malay	35,640	14,987	9,919	5,053	35,297
French	17,063	7,383	5,609	3,363	26,809
Iban	5,629	5,101	4,294	3,580	7,111
Thai	14,687	13,037	10,883	6,587	8,363

Table 2 Lexicon+TX type and token coverage of 500 English and Malay Wikipedia articles

Language	Total tokens	Token coverage (%)	Total types	Type coverage (%)
English	892,224	804,184 (90.1)	70,238	31,630 (45.0)
Malay	206,682	156,105 (75.5)	33,650	12,689 (37.7)

– CC-CEDICT[5] (93,847 Chinese items, 107,228 English items)
– FeM[6] (28,288 French items, 23,148 English items, 41,519 Malay items)
– Handy Reference Dictionary of Iban and English (Sutlive and Sutlive, 1992) (9,825 Iban items, 14,201 English items)
– Yaitron[7] (32,347 Thai items, 22,660 English items)

Translation triples were generated and later aggregated using the modified OTIC procedure, to build up a prototype Lexicon+TX that eventually comprises English, Malay, Chinese, French, Iban and Thai LIs. Figure 5 shows a sample generated translation set, containing translation of «rainbow», retrieved via a PHP search interface.

Table 1 shows the number of new target languages that LIs in each source language are connected to. Note that since the Iban–English dictionary contained fewer entries compared to other input dictionaries, the number of LIs connected to all five other languages are therefore limited. Nevertheless, as far as the authors are aware, Iban, an

[5] http://cc-cedict.org/wiki/start
[6] http://www-clips.imag.fr/cgi-bin/geta/fem/fem.pl
[7] https://github.com/veer66/Yaitron

Table 3 Satisfaction score of 500 randomly selected translation sets

Score	Description	No. of sets	(%)
3	No further work needed	365	73.0
2	Minor correction: delete errant LIs	91	18.2
1	Major correction: regroup into multiple translation sets	10	2.0
0	Bad: unintelligible translation set, discard	34	6.8
	Total	500	100.0

under-resourced language,[8] is now connected for the first time to French, Thai and Chinese with relatively minimal effort and cost (albeit with precision trade-offs), all of which are rare language pairings.

Table 1 also shows the number of translation sets in which LIs of each member language appear. Comparing the number of LIs and translation sets (equivalent to cross-lingual synsets), English and French show high polysemy, while Malay and Chinese LIs are more monosemous. This is because Malay has more derivational processes than inflectional, in which derived Malay words often bear different senses, and are therefore treated as distinct LIs. For example, from the root «lari» we have «berlari» ('to run'), «berlari-larian» ('to run or chase aimlessly'), «melarikan» ('to abduct'), «pelarian» ('refugee'). On the other hand, Chinese LIs often consists of compound characters, where each combination bear distinct meanings, e.g. «蔬菜» ('vegetable') and «菜肴» ('culinary dish'). In contrast, the Iban and Thai source dictionaries contain many synonyms, thus resulting in translation sets containing many member LIs from those two languages.

To gauge the coverage of Lexicon+TX, 500 English articles and 500 Malay articles were downloaded from Wikipedia. The total number of lemmatised tokens and types in each language were then counted, as well as the coverage of Lexicon+TX entries. The results are summarised in Table 2. In addition, Lexicon+TX contains 5,078 (92.9 %) of the 5,464 most frequent English lemmas in the British National Corpus (Kilgariff, 1996).

4.2 Quick Evaluation of Lexicon+TX

As a quick evaluation, 500 translation sets were randomly extracted from Lexicon+TX and manually assessed.[9] Each translation set is given a satisfaction score of 0 to 3 depending on the amount of work required to improve it. The summarised results are shown in Table 3. As the table shows, 91.2% of the translation sets require minimal or no correction (i.e. with satisfaction score ≥ 2).

Table 4 compares the precision of the translation sets from Lexicon+TX to multilingual lexicons generated by other related work. Here, the precision metric only counts

[8] According to `https://en.wiktionary.org/wiki/Wiktionary:Statistics`, Wiktionary contains only 39 Iban entries at the time of writing (June 2013).

[9] Due to limitations of the evaluator's linguistics capabilities, only the English, Chinese and Malay members of each translation set are considered.

Table 4 Comparison of precision of merged translation sets with related work

Cited work	Precision	Resources used
Proposed method	0.73	Translation lists
Sammer and Soderland (2007)	0.20	Translation lists, monolingual corpora
Mausam et al (2009)	0.90	Pre-existing sense-distinguished multilingual lexicons

translation sets in which all member LIs indicate the same meaning, i.e. entries with score = 3 in Table 3. Sammer and Soderland's (2007) low precision score is mainly due to many semantically related words that are not synonyms being included in the same translation set, e.g. «bullet» and «shot». The multilingual lexicon produced by Mausam et al.'s (2009) graph-walking algorithm has a very high precision, but their approach merges *pre-existing* sense-distinguished multilingual lexicons (crowd-sourced Wiktionaries), in which the presence and coverage of under-resourced languages are not guaranteed. In contrast, the proposed modified OTIC method attempts to build a sense-distinguished multilingual lexicon from unaligned bilingual dictionaries, and may thus be more suitable for under-resourced languages.

4.3 Evaluation for Language Pairs Involving Under-Resourced Languages

As mentioned in the previous section, it is not always possible to find multilingual evaluators who are also fluent in an under-resourced languages. Instead, bilingual evaluators are easier to come by. We therefore conducted evaluations for newly generated under-resourced language pairs, in addition to evaluating multilingual translation sets.

500 random Malay–Chinese and Iban–Malay translation pairs generated from OTIC (before filtering) were extracted. There were graded by human evaluators[10] as *accept*, *reject* or *unsure*. Only two out of the ten evaluators have a background in computational linguistics; none had special training as translators nor linguists. The gold standard was then obtained by taking the majority vote to reach an *accept* or *reject* verdict for each translation pairing. An *accept* verdict is assumed in case of a tie.

OTIC filtering was then run with varying threshold parameters. By comparing the decisions of the OTIC filtering to the gold standard, we compute the precision, recall and harmonic mean (F_1) scores of OTIC filtering as:

$$\text{Precision} = \frac{tp}{tp+fp} \qquad \text{Recall} = \frac{tp}{tp+fn}$$

$$F_1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

$$\text{where } tp = \text{true positive}, \qquad fp = \text{false positive},$$

$$tn = \text{true negative}, \qquad fn = \text{false negative}.$$

[10] Five evaluators for each language pair.

Note that because the level of overlap between dictionaries depends on the sets of dictionaries used, the precision and recall can vary for different language pairs and input dictionaries.

The best precision and F_1 score achieved are shown in Table 5, with the corresponding precision and recall in parentheses.

Table 5 Best precision and F_1 scores achieved by OTIC in filtering Malay–Chinese and Iban–Malay translation pairs

Translation pairs	Best precision (recall)	Best F_1 (precision/recall)
Malay–Chinese	0.770 (0.380)	0.725 (0.636/0.843)
Iban–Malay	0.565 (0.354)	0.660 (0.492/1.000)

While a higher precision score from higher filter threshold parameters is undoubtedly desirable, this would also mean a lower recall as more translation triples (and hence equivalence links) are rejected. Some trade-off between precision and recall is therefore required in determining the filter threshold parameters, to ensure the multilingual translation sets are sufficiently accurate and contains a reasonable number of LIs, as indicated by that F_1 score. The threshold parameters that yields the best F_1 scores for each language pair is used to generate the final translation triples and translation sets.

Table 6 compares the results achieved by the proposed method with two related work on aligning translation pairs while maintaining the senses. Note again that the numbers reported in this table may not be suitable for comparative evaluation due to differences in the experiment methodology and language differences. Rather, the performance of the two related work are cited here to provide a context.

Sammer and Soderland (2007) generated English–Spanish–Chinese sets using bilingual dictionaries and monolingual corpora. (Their method is considered low cost as monolingual corpora are more readily available.) Varga et al (2009) generated Japanese–English–Hungarian sets using bilingual dictionaries and the English WordNet, which may not be applicable for under-resourced languages due to the WordNet requirement. As the table shows, the proposed method performed quite favourably, especially in view of the richness of resource types used in each work.

An unexpected outcome from this exercise was the relatively short time the evaluators took for grading the translation pairs. Although they were initially asked to evaluate only 100 pairs each, most of the evaluators took about 2–4 hours to return their decisions for all 500 pairs.

Table 6 Precision comparison with related work

Cited work	Precision	Resources used
Proposed method	0.77	Translation lists
Sammer and Soderland (2007)	0.73	Translation lists, monolingual corpora
Varga et al (2009)	0.79	Translation lists, WordNet

5 Discussion

OTIC typically works well for less polysemous words, and not so well for polysemous words. Consequently, almost all errors in the generated translation sets were due to the presence of a polysemous 'mid' language LI in the translation set, which may cause a 'tail' language LI to be connected to the 'head' language LI erroneously during the OTIC process. This is especially pronounced as English is the 'mid' language for all our triple generation tasks, and English words exhibit high polysemy. Support verbs, such as «go», «make», «take» and «come», are particularly problematic. As a rough estimate by manual inspection, about 60 % to 70 % of the membership of translation sets containing the 50 most frequent (and therefore highly polysemous) English LIs are usable.

To reduce such problems, a language containing less polysemy should be chosen as the 'mid' language, if possible. Nevertheless, the choice may be quite limited for under-resourced languages: most bilingual dictionaries would involve a major language, which often happens to be English. Alternatively, the effects of polysemy may be reduced if richer resources – such as using domain codes, classifiers or linking via a second 'mid' language – were available (Bond and Ogura, 2008). Nonetheless, such resources may not be available or applicable to all languages, especially under-resourced ones.

While errors can be also reduced (i.e. raising the precision) by increasing the OTIC filtering threshold parameter α, this would also entail a lower recall as more translation triples (and hence equivalence links) are rejected. Some trade-off between precision and recall is therefore required in determining the filter thresholds to ensure the multilingual dictionary is sufficiently accurate and contains a reasonable number of LIs. Another drawback is that the number of acquired translation equivalence links are constrained by the degree of overlap between the input dictionaries (see also Table 1).

Overall, the results are highly satisfactory, considering the simplicity of the input data required. Specifically, the proposed modified OTIC procedure provides a fast, cheap and effective way for generating a first draft of a multilingual lexicon, which will then be improved by human evaluators. The method requires only simple bilingual translation lists as input data, and is therefore suitable for under-resourced languages (e.g. Iban).

One drawback of the proposed method is that the number of acquired translation equivalence links are constrained by the degree of overlap between the input dictionaries, impacting recall. Table 1 shows that as the number of target languages increases, the number of LIs having translations in all other target languages decreases. For example, less than one-third of the English LIs and half of the Iban LIs are included in Lexicon+TX. In addition, since the Iban–English dictionary contains far fewer entries than the other input dictionaries, the number of LIs with translations in all 5 target languages are limited.

Another reason affecting the recall is the presence of multi-word expressions: as much as half the English LIs in some input dictionaries are multi-word expressions, which includes technical terms, idioms and proverbial expressions. These were included in the OTIC process as we were interested to discover as many translation equivalents as possible. However, such multi-word LIs often do not have equivalent

LIs in other languages, only gloss texts. OTIC would therefore fail to generate any translation triples for such LIs.

6 Conclusion and Future Work

We have proposed a method for constructing multilingual lexicons using low cost means and resources, such that under-resourced languages can be rapidly connected to richer, more dominant languages. Lexicon+TX, a prototype multilingual lexicon containing six languages (English, Malay, Chinese, French, Thai and Iban) was successfully constructed using simple input bilingual dictionaries. As far as the authors are aware, this is the first time that Iban, an under-resourced languages, is connected to more widely spoken languages like Chinese, French and Thai.

As future work, we would like to improve the precision of the OTIC filtering process by using various heuristics. For example, since English is most likely to be the mid language, English resources could be exploited more to reduce the effects of polysemy. Another possibly helpful heuristic is to assume the first sense of polysemous mid-language LIs. There are also plans to link or align translation sets from Lexicon+TX to Princeton's WordNet (Miller et al, 1990), Papillon (Boitet et al, 2002) or the UNL (Uchida et al, 2005) Universal Words dictionary. We also plan to integrate Lexicon+TX into a machine translation system and an intelligent reading aid.

Acknowledgements

The authors would like to thank volunteers who took part in evaluating the OTIC filtering results on Malay–Chinese and Iban–Malay. We also thank the three anonymous reviewers for their extremely useful comments in improving this paper.

References

Berment V (2004) Méthods pour informatiser les langues et les groupes de langues 'peu dotées'. PhD thesis, Université Joseph Fourier, Grenoble, France

Boitet C, Mangeot M, Sérasset G (2002) The PAPILLON project: Cooperatively building a multilingual lexical database to derive open source dictionaries & lexicons. In: Proceedings of (NLPXML'02), Taipei, Taiwan, pp 1–3

Boitet C, Zaharin Y, Tang EK (2011) Learning-to-translate based on the S-SSTC annotation schema. In: Proceedings of 25th Pacific Asia Conference on Language, Information and Computation (PACLIC 2011), Singapore

Bond F, Ogura K (2008) Combining linguistic resources to create a machine-tractable Japanese–Malay dictionary. Language Resources and Evaluation 42:127–136

Bond F, Paik K (2012) A survey of wordnets and their licenses. In: Proceedings of the 6th Global WordNet Conference (GWC 2012), Matsue, Japan, pp 64—-71

Bond F, Ruhaida bS, Yamazaki T, Ogura K (2001) Design and construction of a machine-tractable Japanese-Malay dictionary. In: Proceedings of MT Summit VIII, Santiago de Compostela, Spain, pp 53–58

Daoud M, Daoud D, Boitet C (2009) Collaborative construction of Arabic lexical resources. In: Choukri K, Maegaard B (eds) Proceedings of the Second International Conference on Arabic Language Resources and Tools, The MEDAR Consortium, Cairo, Egypt

Dorow B, Laws F, Michelbacher L, Scheible C, Utt J (2009) A graph-theoretic algorithm for automatic extension of translation lexicons. In: Proceedings of the EACL 2009 Workshop on GEMS: GEometical Models of Natural Language Semantics, Athens, Greece, pp 91–95

Fellbaum C (ed) (1998) WordNet: An Electronic Lexical Database. Language, Speech, and Communication, MIT Press, Cambridge, Massachusetts

Francopoulo G, Bel N, George M, Calzolari N, Monachini M, Pet M, Soria C (2009) Multilingual resources for NLP in the lexical markup framework (LMF). Language Resources and Evaluation 43(1):57–70, DOI 10.1007/s10579-008-9077-5

Jalabert F, Lafourcade M (2002) From sense naming to vocabulary augmentation in Papillon. In: Proceedings of PAPILLON-2003 Workshop, Sapporo, Japan

Janssen M (2004) Multilingual lexical databases, lexical gaps, and SIMuLLDA. International Journal of Lexicography 17:136–154

Johns AH (ed) (2000) Kamus Inggeris Melayu Dewan: An English-Malay Dictionary. Dewan Bahasa dan Pustaka, Kuala Lumpur, Malaysia

Kilgariff A (1996) BNC database and word frequency lists. URL `http://www.kilgarriff.co.uk/bnc-readme.html`

Lafourcade M (2007) Making people play for lexical acquisition. In: Proceedings of the 7th Symposium on Natural Language Processing (SNLP 2007), Pattaya, Thailand

Mangeot-Lerebours M, Sérasset G, Lafourcade M (2003) Construction collaborative d'une base lexicale multilingue – le project Papillon. Traitement Automatiques des Langues 44(2):151–176

Mausam, Soderland S, Etzioni O, Weld D, Skinner M, Bilmes J (2009) Compiling a massive, multilingual dictionary via probabilistic inference. In: Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP, Suntec, Singapore, pp 262–270

de Melo G, Weikum G (2009) Towards a universal wordnet by learning from combined evidence. In: Proceedings of the 18th ACM Conference on Information and Knowledge Management (CIKM 2009), ACM, New York, NY, USA, pp 513–522, DOI http://doi.acm.org/10.1145/1645953.1646020

Miller GA, Beckwith R, Fellbaum C, Gross D, Miller KJ (1990) Introduction to WordNet: An on-line lexical database. International Journal of Lexicography (special issue) 3(4):235–312

Navigli R, Ponzetto S (2012) BabelNetXplorer: A platform for multilingual lexical knowledge base access and exploration. In: Proceedings of the 21st International World Wide Web Conference, complementary volume (WWW 2012), Lyon, France, pp 393–396

Pianta E, Bentivogli L, Girardi C (2002) MultiWordNet: Developing an aligned multilingual database. In: Proceedings of the First International Conference on Global WordNet, Mysore, India

Sammer M, Soderland S (2007) Building a sense-distinguished multilingual lexicon from monolingual corpora and bilingual lexicons. In: Proceedings of Machine Translation Summit XI, Copenhagen, Denmark, pp 399–406

Sutlive V, Sutlive J (1992) Handy Reference Dictionary of Iban and English. Tun Jugah Foundation Association

Tanaka K, Umemura K, Iwasaki H (1998) Construction of a bilingual dictionary intermediated by a third language. Transactions of the Information Processing Society of Japan 39(6):1915–1924, in Japanese

Tufiş D, Barbu AM, Ion R (2004) Extracting multilingual lexicons from parallel corpora. Computers and the Humanities 38(2):163–189

Tufiş D, Cristea D, Stamou S (2004) BalkaNet: Aims, methods, results and perspectives – a general overview. Romanian Journal of Information Science and Technology Special Issue 7(1):9–43

Uchida H, Zhu M, Senta TD (2005) Universal Networking Language. UNDL Foundation

Varga I, Yokoyama S, Hashimoto C (2009) Dictionary generation for less-frequent language pairs using WordNet. Literary and Linguistic Computing 24(4):449–466

Verma N, Bhattacharyya P (2003) Automatic generation of multilingual lexicon by using WordNet. In: Proceedings of International Conference on Convergence of Knowledge, Culture, Language and IT, Library of Alexandria, Egypt

Vossen P (2004) 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 17(2)