Context-Dependent Multilingual Lexical Lookup for Under-Resourced Languages

Lian Tze Lim[*†]
[*]SEST, KDU College Penang
Georgetown, Penang, Malaysia
liantze@gmail.com

Lay-Ki Soon and **Tek Yong Lim**
[†]FCI, Multimedia University
Cyberjaya, Selangor, Malaysia
{lksoon,tylim}@mmu.edu.my

Enya Kong Tang
Linton University College
Seremban, Negeri Sembilan, Malaysia
enyakong1@gmail.com

Bali Ranaivo-Malançon
FCSIT, Universiti Malaysia Sarawak,
Kota Samarahan, Sarawak, Malaysia
mbranaivo@fit.unimas.my

Abstract

Current approaches for word sense disambiguation and translation selection typically require lexical resources or large bilingual corpora with rich information fields and annotations, which are often infeasible for under-resourced languages. We extract translation context knowledge from a bilingual comparable corpora of a richer-resourced language pair, and inject it into a multilingual lexicon. The multilingual lexicon can then be used to perform context-dependent lexical lookup on texts of any language, including under-resourced ones. Evaluations on a prototype lookup tool, trained on a English–Malay bilingual Wikipedia corpus, show a precision score of 0.65 (baseline 0.55) and mean reciprocal rank score of 0.81 (baseline 0.771). Based on the early encouraging results, the context-dependent lexical lookup tool may be developed further into an intelligent reading aid, to help users grasp the gist of a second or foreign language text.

1 Introduction

Word sense disambiguation (WSD) is the task of assigning sense tags to ambiguous lexical items (LIs) in a text. Translation selection chooses target language items for translating ambiguous LIs in a text, and can therefore be viewed as a kind of WSD task, with translations as the sense tags. The translation selection task may also be modified slightly to output a ranked list of translations. This then resembles a dictionary lookup process as performed by a human reader when reading or browsing a text written in a second or foreign language. For convenience's sake, we will call this task (as performed via computational means) *context-dependent lexical lookup*. It can also be viewed as a simplified version of the Cross-Lingual Lexical Substitution (Mihalcea et al., 2010) and Cross-Lingual Word Sense Disambiguation (Lefever and Hoste, 2010) tasks, as defined in SemEval-2010.

There is a large body of work around WSD and translation selection. However, many of these approaches require lexical resources or large bilingual corpora with rich information fields and annotations, as reviewed in section 2. Unfortunately, not all languages have equal amounts of digital resources for developing language technologies, and such requirements are often infeasible for under-resourced languages.

We are interested in leveraging richer-resourced language pairs to enable context-dependent lexical lookup for under-resourced languages. For this purpose, we model translation context knowledge as a second-order co-occurrence bag-of-words model. We propose a rapid approach for acquiring them from an untagged, comparable bilingual corpus of a (richer-resourced) language pair in section 3. This information is then transferred into a multilingual lexicon to perform context-dependent lexical lookup on input texts, including those in an under-resourced language (section 4). Section 5 describes a prototype implementation, where translation context knowledge is extracted from a English–Malay bilingual corpus to enrich a multilingual lexicon with six languages. Results from a small experiment are presented in 6 and discussed in section 7. The approach is briefly compared with some related work in section 8, before concluding in section 9.

2 Typical Resource Requirements for Translation Selection

WSD and translation selection approaches may be broadly classified into two categories depending

Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, pages 294–299,
Sofia, Bulgaria, August 4-9 2013.

on the type of learning resources used: knowledge- and corpus-based. Knowledge-based approaches make use of various types of information from existing dictionaries, thesauri, or other lexical resources. Possible knowledge sources include definition or gloss text (Banerjee and Pedersen, 2003), subject codes (Magnini et al., 2001), semantic networks (Shirai and Yagi, 2004; Mahapatra et al., 2010) and others.

Nevertheless, lexical resources of such rich content types are usually available for medium- to rich-resourced languages only, and are costly to build and verify by hand. Some approaches therefore turn to corpus-based approaches, use bilingual corpora as learning resources for translation selection. (Ide et al., 2002; Ng et al., 2003) used aligned corpora in their work. As it is not always possible to acquire parallel corpora, comparable corpora, or even independent second-language corpora have also been shown to be suitable for training purposes, either by purely numerical means (Li and Li, 2004) or with the aid of syntactic relations (Zhou et al., 2001). Vector-based models, which capture the context of a translation or meaning, have also been used (Schütze, 1998; Papp, 2009). For under-resourced languages, however, bilingual corpora of sufficient size may still be unavailable.

3 Enriching Multilingual Lexicon with Translation Context Knowledge

Corpus-driven translation selection approaches typically derive supporting semantic information from an aligned corpus, where a text and its translation are aligned at the sentence, phrase and word level. However, aligned corpora can be difficult to obtain for under-resourced language pairs, and are expensive to construct.

On the other hand, documents in a comparable corpus comprise bilingual or multilingual text of a similar nature, and need not even be exact translations of each other. The texts are therefore unaligned except at the document level. Comparable corpora are relatively easier to obtain, especially for richer-resourced languages.

3.1 Overview of Multilingual Lexicon

Entries in our multilingual lexicon are organised as multilingual translation sets, each corresponding to a coarse-grained concept, and whose members are LIs from different languages $\{L_1, \ldots, L_N\}$ conveying the same concept. We denote an LI as «item», sometimes with the 3-letter ISO language code in underscript when necessary: «item»$_{\text{eng}}$. A list of 3-letter ISO language codes used in this paper is given in Appendix A.

For example, following are two translation sets containing different senses of English «bank» ('financial institution' and 'riverside land'):

$$TS_1 = \{\text{«bank»}_{\text{eng}}, \text{«bank»}_{\text{msa}}, \text{«银行»}_{\text{zho}}, \ldots\}$$
$$TS_2 = \{\text{«bank»}_{\text{eng}}, \text{«tebing»}_{\text{msa}}, \text{«岸»}_{\text{zho}}, \ldots\}.$$

Multilingual lexicons with under-resourced languages can be rapidly bootstrapped from simple bilingual translation lists (Lim et al., 2011). Our multilingual lexicon currently contains 24371 English, 13226 Chinese, 35640 Malay, 17063 French, 14687 Thai and 5629 Iban LIs.

3.2 Extracting Translation Context Knowledge from Comparable Corpus

We model translation knowledge as a bag-of-words consisting of the context of a translation equivalence in the corpus. We then run latent semantic indexing (LSI) (Deerwester et al., 1990) on a comparable bilingual corpora. A vector is then obtained for each LI in both languages, which may be regarded as encoding some translation context knowledge.

While LSI is more frequently used in information retrieval, the translation knowledge acquisition task can be recast as a cross-lingual indexing task, following (Dumais et al., 1997). The underlying intuition is that in a comparable bilingual corpus, a document pair about finance would be more likely to contain English «bank»$_{\text{eng}}$ and Malay «bank»$_{\text{msa}}$ ('financial institution'), as opposed to Malay «tebing»$_{\text{msa}}$ ('riverside'). The words appearing in this document pair would then be an indicative context for the translation equivalence between «bank»$_{\text{eng}}$ and «bank»$_{\text{msa}}$. In other words, the translation equivalents present serve as a kind of implicit sense tag.

Briefly, a translation knowledge vector is obtained for each multilingual translation set from a bilingual comparable corpus as follows:

1. Each bilingual pair of documents is merged as one single document, with each LI tagged with its respective language code.
2. Pre-process the corpus, e.g. remove closed-class words, perform stemming or lemmatisation, and word segmentation for languages without word boundaries (Chinese, Thai).

3. Construct a term-document matrix (TDM), using the frequency of terms (each made up by a LI and its language tag) in each document. Apply further weighting, e.g. TF-IDF, if necessary.
4. Perform LSI on the TDM. A vector is then obtained for every LI in both languages.
5. Set the vector associated with each translation set to be the sum of all available vectors of its member LIs.

4 Context-Dependent Lexical Lookup

Given an input text in language L_i ($1 \leq i \leq N$), the lookup module should return a list of multilingual translation set entries, which would contain $L_1, L_2, \ldots, L_N$ translation equivalents of LIs in the input text, wherever available. For polysemous LIs, the lookup module should return translation sets that convey the appropriate meaning in context.

For each input text segment Q (typically a sentence), a 'query vector', V_Q is computed by taking the vectorial sum of all open class LIs in the input Q. For each LI l in the input, the list of all translation sets containing l, is retrieved into TS_l.

TS_l is then sorted in descending order of

$$\text{CSim}(V_t, V_Q) = \frac{V_t \cdot V_Q}{|V_t| \times |V_Q|}$$

(i.e. the cosine similarity between the query vector V_Q and the translation set candidate t's vector) for all $t \in TS_l$.

If the language of input Q is not present in the bilingual training corpus (e.g. Iban, an under-resourced language spoken in Borneo), V_Q is then computed as the sum of all vectors associated with all translation sets in TS_l. For example, given the Iban sentence *'Lelaki nya tikah enggau emperaja iya, siko dayang ke ligung'* ('he married his sweetheart, a pretty girl'), V_Q would be computed as

$$\begin{aligned} V_Q = &\sum V(\text{lookup}(\text{«lelaki»}_{\text{iba}})) \\ &+ \sum V(\text{lookup}(\text{«tikah»}_{\text{iba}})) \\ &+ \sum V(\text{lookup}(\text{«emperaja»}_{\text{iba}})) \\ &+ \sum V(\text{lookup}(\text{«dayang»}_{\text{iba}})) \\ &+ \sum V(\text{lookup}(\text{«ligung»}_{\text{iba}})) \end{aligned}$$

where the function lookup(w) returns the translation sets containing LI w.

5 Prototype Implementation

We have implemented LEXICALSELECTOR, a prototype context-dependent lexical lookup tool in Java, trained on a English–Malay bilingual corpus built from Wikipedia articles. Wikipedia articles are freely available under a Creative Commons license, thus providing a convenient source of bilingual comparable corpus. Note that while the training corpus is English–Malay, the trained lookup tool can be applied to texts of any language included in the multilingual dictionary.

Malay Wikipedia articles[1] and their corresponding English articles of the same topics[2] were first downloaded. To form the bilingual corpus, each Malay article is concatenated with its corresponding English article as one document.

The TDM constructed from this corpus contains 62 993 documents and 67 499 terms, including both English and Malay items. The TDM is weighted by TF-IDF, then processed by LSI using the Gensim Python library[3]. The indexing process, using 1000 factors, took about 45 minutes on a MacBook Pro with a 2.3 GHz processor and 4 GB RAM. The vectors obtained for each English and Malay LIs were then used to populate the translation context knowledge vectors of translation set in a multilingual lexicon, which comprise six languages: English, Malay, Chinese, French, Thai and Iban.

As mentioned earlier, LEXICALSELECTOR can process texts in any member languages of the multilingual lexicon, instead of only the languages of the training corpus (English and Malay). Figure 1 shows the context-depended lexical lookup outputs for the Iban input 'Lelaki nya tikah enggau emperaja iya, siko dayang ke ligung'. Note that «emperaja» is polysemous ('rainbow' or 'lover'), but is successfully identified as meaning 'lover' in this sentence.

6 Early Experimental Results

80 input sentences containing LIs with translation ambiguities were randomly selected from the Internet (English, Malay and Chinese) and contributed by a native speaker (Iban). The test words are:

- English «plant» (vegetation or factory),

[1] http://dumps.wikimedia.org/mswiki/

[2] http://en.wikipedia.org/wiki/Special:Export

[3] http://radimrehurek.com/gensim/

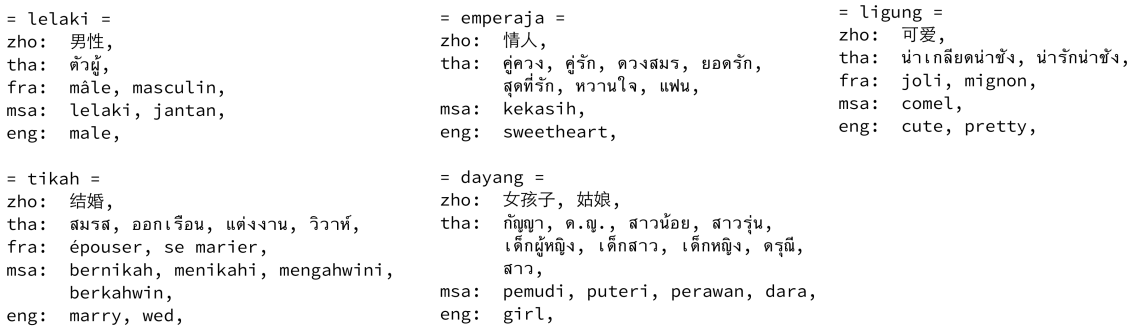

Figure 1: LEXICALSELECTOR output for Iban input 'Lelaki nya tikah enggau emperaja iya, siko dayang ke ligung'. Only top ranked translation sets are shown.

- English «bank» (financial institution or riverside land),
- Malay «kabinet» (governmental Cabinet or household furniture),
- Malay «mangga» (mango or padlock),
- Chinese «谷» (*gù*, valley or grain) and
- Iban «emperaja» (rainbow or lover).

Each test sentence was first POS-tagged automatically based on the Penn Treebank tagset. The English test sentences were lemmatised and POS-tagged with the Stanford Parser.[4] The Chinese test sentences segmented with the Stanford Chinese Word Segmenter tool.[5] For Malay POS-tagging, we trained the QTag tagger[6] on a hand-tagged Malay corpus, and applied the trained tagger on our test sentences. As we lacked a Iban POS-tagger, the Iban test sentences were tagged by hand. LIs of each language and their associated vectors can then be retrieved from the multilingual lexicon.

The prototype tool LEXICALSELECTOR then computes the CSim score and ranks potential translation sets for each LI in the input sentences (ranking strategy `wiki-lsi`). The baseline strategy (`base-freq`) selects the translation set whose members occur most frequently in the bilingual Wikipedia corpus.

As a comparison, the English, Chinese and Malay test sentences were fed to Google Translate[7] and translated into Chinese, Malay and English. (Google Translate does not support Iban currently.) The Google Translate interface makes available the ranked list of translation candidates for each word in an input sentence, one language at a time.The translated word for each of the input test word can therefore be noted. The highest rank of the correct translation for the test words in English/Chinese/Malay are used to evaluate `goog-tr`.

Two metrics were used in this quick evaluation. The first metric is by taking the precision of the first translation set returned by each ranking strategy, i.e. whether the top ranked translation set contains the correct translation of the ambiguous item. The precision metric is important for applications like machine translation, where only the top-ranked meaning or translation is considered.

The results may also be evaluated similar to a document retrieval task, i.e. as a ranked lexical lookup for human consumption. This is measured by the mean reciprocal rank (MRR), the average of the reciprocal ranks of the correct translation set for each input sentence in the test set T:

$$\text{MRR} = \frac{1}{|T|} \sum_{i=1}^{|T|} \frac{1}{\text{rank}_i}$$

The results for the three ranking strategies are summarised in Table 1. For the precision metric, `wiki-lsi` scored 0.650 when all 80 input sentences are tested, while the `base-freq` baseline scored 0.550. `goog-tr` has the highest precision at 0.797. However, if only the Chinese and Malay inputs — which has less presence on the Internet and 'less resource-rich' than English — were tested (since `goog-tr` cannot accept Iban inputs), `wiki-lsi` and `goog-tr` actually performs equally well at 0.690 precision.

In our evaluation, the MRR score of `wiki-lsi` is 0.810, while `base-freq` scored 0.771. `wiki-lsi` even outperforms `goog-tr` when only the Chinese and Malay test sentences are considered for the MRR metric, as `goog-tr`

[4] `http://www-nlp.stanford.edu/software/lex-parser.shtml`

[5] `http://nlp.stanford.edu/software/segmenter.shtml`

[6] `http://phrasys.net/uob/om/software`

[7] `http://translate.google.com` on 3 October 2012

Table 1: Precision and MRR scores of context-dependent lexical lookup

Strategy	Incl. Eng. & Iban		W/o Eng. & Iban	
	Precision	MRR	Precision	MRR
wiki-lsi	0.650	0.810	0.690	0.845
base-freq	0.550	0.771	0.524	0.762
goog-tr	0.797	0.812	0.690	0.708

did not present the correct translation in its list of alternative translation candidates for some test sentences. This suggests that the LSI-based translation context knowledge vectors would be helpful in building an intelligent reading aid.

7 Discussion

`wiki-lsi` performed better than `base-freq` for both the precision and the MRR metrics, although further tests is warranted, given the small size of the current test set. While `wiki-lsi` is not yet sufficiently accurate to be used directly in an MT system, it is helpful in producing a list of ranked multilingual translation sets depending on the input context, as part of an intelligent reading aid. Specifically, the lookup module would have benefited if syntactic information (e.g. syntactic relations and parse trees) was incorporated during the training and testing phase. This would require more time in parsing the training corpus, as well as assuming that syntactic analysis tools are available to process test sentences of all languages, including the under-resourced ones.

Note that even though the translation context knowledge vectors were extracted from an English–Malay corpus, the same vectors can be applied on Chinese and Iban input sentences as well. This is especially significant for Iban, which otherwise lacks resources from which a lookup or disambiguation tool can be trained. Translation context knowledge vectors mined via LSI from a bilingual comparable corpus, therefore offers a fast, low cost and efficient fallback strategy for acquiring multilingual translation equivalence context information.

8 Related Work

Basile and Semeraro (2010) also used Wikipedia articles as a parallel corpus for their participation in the SemEval 2010 Cross-Lingual Lexical Substitution task. Both training and test data were for English–Spanish. The idea behind their system is to count, for each potential Spanish candidate, the number of documents in which the target English word and the Spanish candidate occurs in an English–Spanish document pair. In the task's 'best' evaluation (which is comparable to our 'Precision' metric), Basile and Semeraro's system scored 26.39 precision on the trial data and 19.68 precision on the SemEval test data. This strategy of selecting the most frequent translation is similar to our `base-freq` baseline strategy.

Sarrafzadeh et al. (2011) also tackled the problem of cross-lingual disambiguation for under-resourced language pairs (English–Persian) using Wikipedia articles, by applying the *one sense per collocation* and *one sense per discourse* heuristics on a comparable corpus. The authors incorporated English and Persian wordnets in their system, thus achieving 0.68 for the 'best sense' ('Precision') evaluation. However, developing wordnets for new languages is no trivial effort, as acknowledged by the authors.

9 Conclusion

We extracted translation context knowledge from a bilingual comparable corpus by running LSI on the corpus. A context-dependent multilingual lexical lookup module was implemented, using the cosine similarity score between the vector of the input sentence and those of candidate translation sets to rank the latter in order of relevance. The precision and MRR scores outperformed Google Translate's lexical selection for medium- and under-resourced language test inputs. The LSI-backed translation context knowledge vectors, mined from bilingual comparable corpora, thus provide an fast and affordable data source for building intelligent reading aids, especially for under-resourced languages.

Acknowledgments

The authors thank Multimedia University and Universiti Malaysia Sarawak for providing support and resources during the conduct of this study. We also thank Panceras Talita for helping to prepare the Iban test sentences for context-dependent lookup.

A 3-Letter ISO Language Codes

Code	Language	Code	Language
eng	English	msa	Malay
zho	Chinese	fra	French
tha	Thai	iba	Iban

References

Satanjeev Banerjee and Ted Pedersen. 2003. Extended gloss overlaps as a measure of semantic relatedness. In *Proceedings of the 18th International Joint Conference on Artificial Intelligence*, pages 805–810.

Pierpaolo Basile and Giovanni Semeraro. 2010. UBA: Using automatic translation and Wikipedia for cross-lingual lexical substitution. In *Proceedings of the 5th International Workshop on Semantic Evaluation (SemEval 2010)*, pages 242–247, Uppsala, Sweden.

Scott C. Deerwester, Susan T. Dumais, Thomas K. Landauer, George W. Furnas, and Richard A. Harshman. 1990. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407.

Susan T. Dumais, Michael L. Littman, and Thomas K. Landauer. 1997. Automatic cross-language retrieval using latent semantic indexing. In *AAAI97 Spring Symposium Series: Cross Language Text and Speech Retrieval*, pages 18–24, Stanford University.

Nancy Ide, Tomaz Erjavec, and Dan Tufiş. 2002. Sense discrimination with parallel corpora. In *Proceedings of the SIGLEX/SENSEVAL Workshop on Word Sense Disambiguation: Recent Successes and Future Directions*, pages 54–60, Philadelphia, USA.

Els Lefever and Véronique Hoste. 2010. SemEval-2010 Task 3: Cross-lingual word sense disambiguation. In *Proceedings of the 5th International Workshop on Semantic Evaluation (SemEval 2010)*, Uppsala, Sweden.

Hang Li and Cong Li. 2004. Word translation disambiguation using bilingual bootstrapping. *Computational Linguistics*, 30(1):1–22.

Lian Tze Lim, Bali Ranaivo-Malançon, and Enya Kong Tang. 2011. Low cost construction of a multilingual lexicon from bilingual lists. *Polibits*, 43:45–51.

Bernardo Magnini, Carlo Strapparava, Giovanni Pezzulo, and Alfio Gliozzo. 2001. Using domain information for word sense disambiguation. In *Proceedings of the 2nd International Workshop on Evaluating Word Sense Disambiguation Systems (SENSEVAL-2)*, pages 111–114, Toulouse, France.

Lipta Mahapatra, Meera Mohan, Mitesh M. Khapra, and Pushpak Bhattacharyya. 2010. OWNS: Cross-lingual word sense disambiguation using weighted overlap counts and Wordnet based similarity measures. In *Proceedings of the 5th International Workshop on Semantic Evaluation (SemEval 2010)*, Uppsala, Sweden.

Rada Mihalcea, Ravi Sinha, and Diana McCarthy. 2010. SemEval-2010 Task 2: Cross-lingual lexical substitution. In *Proceedings of the 5th International Workshop on Semantic Evaluation (SemEval 2010)*, Uppsala, Sweden.

Hwee Tou Ng, Bin Wang, and Yee Seng Chan. 2003. Exploiting parallel texts for word sense disambiguation: An empirical study. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 455–462, Sapporo, Japan.

Gyula Papp. 2009. Vector-based unsupervised word sense disambiguation for large number of contexts. In Václav Matoušek and Pavel Mautner, editors, *Text, Speech and Dialogue*, volume 5729 of *Lecture Notes in Computer Science*, pages 109–115. Springer Berlin Heidelberg.

Bahareh Sarrafzadeh, Nikolay Yakovets, Nick Cercone, and Aijun An. 2011. Cross-lingual word sense disambiguation for languages with scarce resources. In *Proceedings of the 24th Canadian Conference on Advances in Artificial Intelligence*, pages 347–358, St. John's, Canada.

Hinrich Schütze. 1998. Automatic word sense discrimination. *Computational Linguistics*, 24(1):97–123.

Kiyoaki Shirai and Tsunekazu Yagi. 2004. Learning a robust word sense disambiguation model using hypernyms in definition sentences. In *Proceedings of the 20th International Conference on Computational Linguistics (COLING 2004)*, pages 917–923, Geneva, Switzerland. Association for Computational Linguistics.

Ming Zhou, Yuan Ding, and Changning Huang. 2001. Improviging translation selection with a new translation model trained by independent monolingual corpora. *Computational Linguistics and Chinese language Processing*, 6(1):1–26.