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. ©2013 Association for Computational Linguistics