Improving Translation Selection using Conceptual Vectors

Lim Lian Tze
Computer Aided Translation Unit
School of Computer Sciences
Universiti Sains Malaysia
Penang, Malaysia
liantze@cs.usm.my

ABSTRACT

Natural language is ambiguous, in that the same word may mean different things in different contexts. This poses problems in various natural language processing (NLP) applications and tasks, including machine translation (MT), where translations for ambiguous words in the input text must be chosen to reflect the correct meaning. We seek to improve translation selection for a MT system by using the Conceptual Vector (CV) model to represent semantic themes of lexical items, which is produced from a dictionary source and also a translation corpus. We opted to consider semantic information of words on the level of translation units at translation runtime, as opposed to the level of sense numbers as found in dictionaries. Experiment results show that the improved MT system produces translations that better reflect the meaning of the original text.

KEYWORDS

Translation selection, example-based machine translation, conceptual vectors.

1. Introduction

Natural language is highly ambiguous in nature, in that a word may have different meanings in different contexts. For example, the English noun "bank" may mean a financial institution, or sloping land beside a body of water, depending on the context in which it is used.

Such words are said to be ambiguous, and the determination of the correct meaning (or sense) of ambiguous words in a particular context is termed word sense disambiguation (WSD). Many natural language processing (NLP) tasks require WSD to produce appropriate results. Machine translation (MT), or the automatic translation of natural text from a source language (SL) to a target language (TL), is one of them. Here, translation words for ambiguous words in the SL input text must be chosen to correctly reflect the original meaning, a process also known as translation selection.

WSD applications typically assign a sense number to a particular ambiguous word occurrence in context, chosen from a list of all possible senses for the word. However, as different NLP tasks require different levels of sense granularity [1], we feel that it is beneficial to adapt WSD approaches specifically for individual applications, which in this case is that of MT. This paper will give a high-level description of the approach we took to improve translation selection in an existing example-based machine translation (EBMT) system on this basis, using dictionary and translation corpus resources. (See [2] for a more in-depth discussion, including background information and reviews about WSD and MT in general.)

2. Motivation and Objectives

As Lee *et al.* [3] had observed, each word in a SL may have multiple senses (as listed in a dictionary). Each sense in turn may have multiple translations in a TL, only one of which is acceptable as the translation of the original SL word in a particular context. Lee and Kim [4] thus addressed translation selection with a two-stage process, where ambiguous words in the input text are first sense-tagged (i.e. assigned a dictionary sense number), followed by the selection of a suitable translation word for the assigned sense number.

Our main objective is to improve the quality of translations of an existing MT system by adapting a WSD approach specifically for translation selection. We propose to short-circuit the two-stage disambiguation process described above, by choosing the most suitable translation

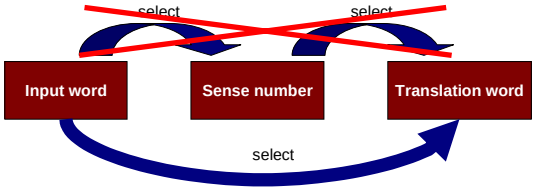


Figure 1: Optimising WSD for translation selection

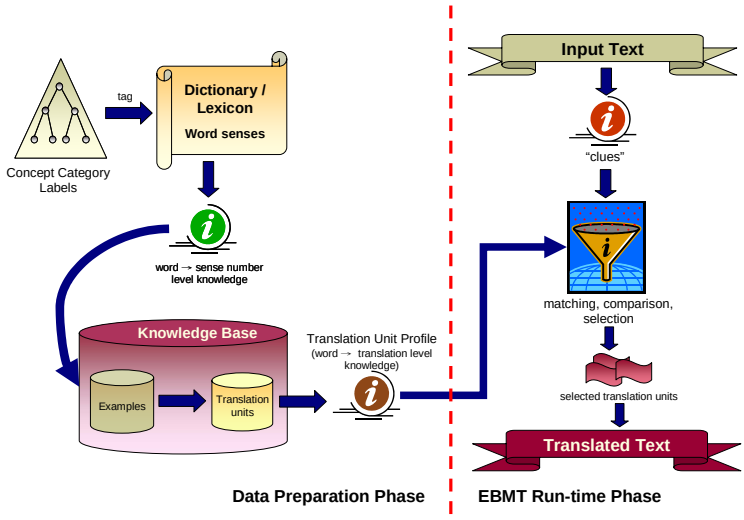


Figure 2: Overview of data preparation and translation runtime phases

unit from a database of previously translated text at runtime (**Figure 1**). This means no sense number determination is required at translation runtime.
To achieve this main objective, lexical semantic data need to be prepared on two levels as sub-objectives: the SL word sense level based on a lexicon, and the translation unit level based on the lexicon and the example database.

3. Methodology

Figure 2 shows an overview of the tasks involved in the data preparation and translation runtime phase. As mentioned briefly in the last section, semantic information about word senses and translation pairs need to be captured computationally.

We use the conceptual vector (CV) model proposed by Lafourcade [5] for this purpose. In this framework, the themes or concepts related to a lexical item is modelled as a mathematical vector. The thematic similarity between two lexical items is then measured as the angular distance between the CVs representing those two lexical items.

To re-iterate, we seek to conceptually profile translation units, i.e. pairs of words or phrases which are translations of each other in a selected SL and a TL, as found in a parallel translation corpus. To this end, the following steps need to be undertaken in the data preparation phase:

(i) tagging senses from a SL lexicon with semantic concept labels,
(ii) computing definition CVs for lexicon senses,
(iii) sense-tagging the SL portion of translation examples (pairs of sentences in the SL and TL),
(iv) computing profile CVs of translation units (fragments of translation examples).

At translation runtime, clue CVs will be computed from the context of each ambiguous SL word occurrence. These clue CVs will be matched against the profile CVs of candidate translation units. The translation unit whose profile CV is thematically most similar to the clue CVs will be selected.

3.1 Tagging Word Senses with Concepts

We start with some chosen SL lexicon, and a set of semantic concept labels, which preferably forms a hierarchy. Each word sense from the lexicon is assigned a selection of concept labels, using the definition text as a guide. For example, if a sense of the English noun **circulation** is defined as:

> **circulation [n] 6.** the spread or transmission of something (as news or money) to a wider group or area. ...

Possible concept labels for **circulation#6** may then include MONEY, INFORMATION, SPREAD_MOVEMENT, and TRANSMISION_OF_INFORMATION,.

3.2 Computing Definition CVs for Senses

Definition CVs ($V_{\mathcal{L}}$) are next computed for each word sense. This is done by first initialising $V_{\mathcal{L}}$ as a boolean vector with reference to the concept labels assigned in the previous step, followed by iterative computations based on the distribution of the concepts in the original semantic hierarchy, as outlined in [5]. **Figure 3** illustrates this with the example of **circulation#6**.

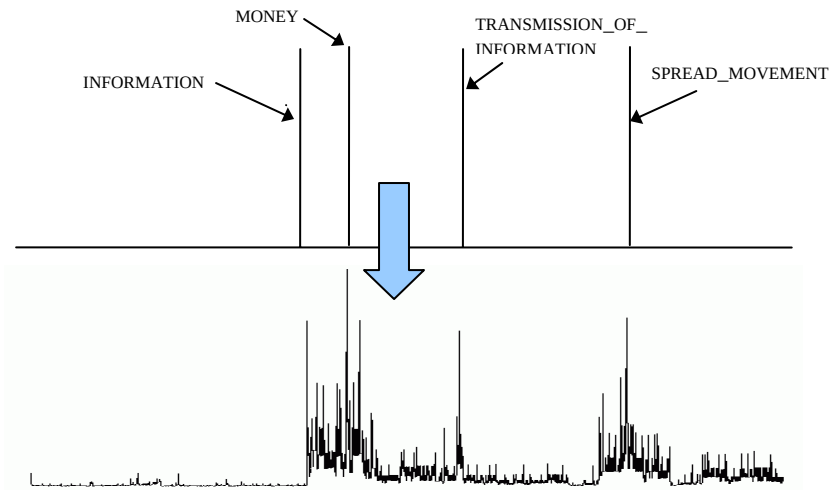
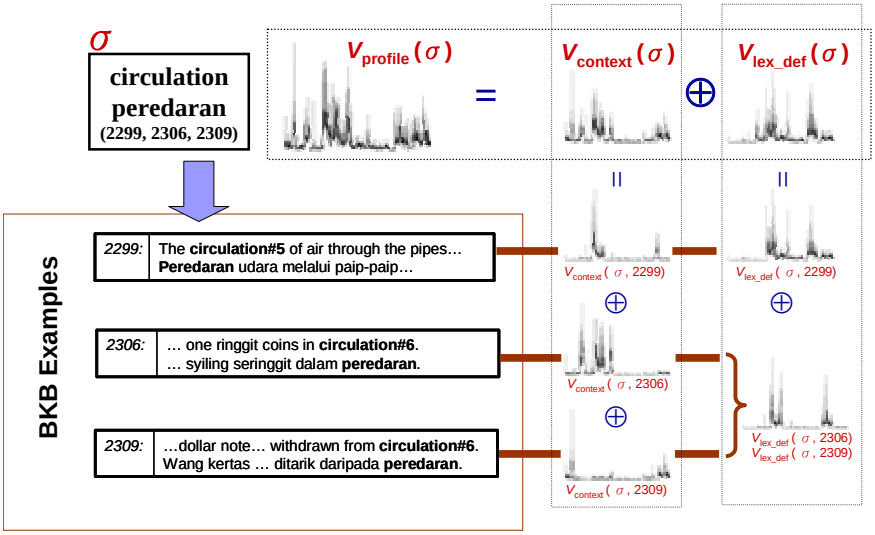


Figure 3: Computing definition CV for circulation#6



circulation#5: free movement or passage through a seires of vessels (as of water through pipes or sap through a plant

circulation#6: the spread or transmission of something (as news or money) to a wider group or area

Figure 4: Computing profile CV for *circulation–peredaran*

3.3 Sense-tagging Translation Examples

The existing MT system which we seek to improve draws on a translation corpus, in which the correspondences between segments of the SL and TL texts, as well as between the text and its tree representation, are explicitly recorded using the S-SSTC annotation schema [6]. Before we can profile the translation units which are extracted from this bilingual knowledge bank, the translation examples need to be sense-tagged. This can be performed automatically using conventional WSD techniques, followed by manual checking.

3.4 Computing Profile CVs for Translation Units

Translation units from the bilingual knowledge bank are grouped based on the SL and TL lexical form, without regard to the sense numbers. Therefore, all three translation examples in **Figure 4** will take part in the calculation of profile CV for the translation unit **circulation–peredaran**, even if two different senses of circulation are involved.

The profile CV for a translation unit is obtained by summing up the definition CV of all senses of the SL word of that translation unit (V_{lex_def}), and also those of the context words appearing in the

translation examples that contain that translation unit ($V_{context}$).

3.5 Translation Selection Based on CVs

Given an SL input text (typically a sentence) to be translated, the summation of CVs of all possible senses for each word in the text is attached to each word token. A clue CV is then computed for each text segment that has multiple possible translations in the TL, by summing up the attached CVs of its context. The system can then select translation units from the bilingual knowledge bank, whose profile CVs have the least angular distance (and hence greatest similarity) with the clue CVs, to be used in producing the final translation output.

4. Results

The following tables show the Malay translation for the English word **circulation** chosen by the translation selection procedure (TS) using CVs as described in Section 3.5, when given these two test inputs:

1. **circulation** of magazine.
2. **circulation** of gossip.

The baseline strategy (BS) is to select the translation that occurs more frequently in the bilingual knowledge bank. The correct Malay translation for each test input is highlighted in bold.

Table 1: Translating 'circulation' to Malay using baseline strategy (BS) and proposed translation selection procedure (TS)

Strategy	***BS***	***TS***
Input 1	penyebaran	**edaran**
Input 2	edaran	**penyebaran**

5. Discussion

Initial test results show that the modified selection procedure produces translation outputs that better reflect the meaning of the input text. In addition, the use of information from both dictionary and corpus sources produced better results than the use of either the dictionary or the corpus in isolation.

By using concepts instead of words as the base of CVs, words of similar meanings are generalised under common groupings. This gives a better coverage for our translation selection procedure, which is further reinforced when profile CVs include concepts from both dictionary definitions and corpora.

To summarise, the following contributions are observed in this work:

- Adaptation of a WSD approach for the specific aim of translation selection;
- Proposal of specific guidelines for assigning related concepts for word meanings from dictionaries [6];
- Production of knowledge about word meanings on two levels i.e. word senses as in dictionaries, and translations as in parallel text.

6. Conclusion

We have presented a translation selection procedure for a MT system, using hybrid knowledge sources and conceptual vectors as the semantic representation model. By preparing lexical semantic data on two levels, we were able to select translation words for ambiguous input words directly, without first performing WSD on the input words.

As not all ambiguities can be solved using the conceptual vector model, an interesting line to pursue in future researches is to investigate how it can be integrated, or be complemented with other WSD techniques to increase the accuracy.

7. References

[1] Ide, N. and Véronis, J., "Word Sense Disambiguation: The State of the Art", *Computational Linguistics*, 24(1), 1998, pp. 1–41.

[2] Lim, L. T., "Improving Translation Se-lection in an Example-based Machine Translation (EBMT) System using Conceptual Vectors", M. Sc. thesis, School of Computer Sciences, Universiti Sains Malaysia, Penang, Malaysia, 2006.

[3] Lee, H. A., Park, J. C. and Kim, G. C. "Lexical selection with a target language monolingual corpus and an MRD." *Proceedings of the 8th International Conference on Theoretical and Methodological Issues in Machine Translation (TMI-'99)*, 1999.

[4] Lee, H. A. and Kim, G. C., "Translation Selection through Source Word Sense Disambiguation and Target Word Selection", *Proceedings of the 17th International Conference on Computational Linguistics (COLING 2002)*, Taipei, Taiwan, 2002.

[5] Lafourcade, M., "Lexical Sorting and Lexical Transfer by Conceptual Vectors", *Proceedings of the 1st International Workshop on Multimedia Annotation (MMA2001)*, Tokyo, Japan, 2001.

[6] Al-Adhaileh, M. H., Tang, E. K. and Zaharin, Y., "A synchronization structure of SSTC and its applications in machine translation", *Proceedings of COLING 2002 Post-Conference Workshop on Machine Translation in Asia*, Taipei, Taiwan, 2002.