Lost in Translation: Word Sense Disambiguation

Lim Lian Tze
Computer Aided Translation Unit
School of Computer Sciences Universiti
Sains Malaysia
11800 Minden, Penang, Malaysia
+60(0) 4 6533888 ext 2693
liantze@cs.usm.my

Tang Enya Kong
Computer Aided Translation Unit
School of Computer Sciences Universiti
Sains Malaysia
11800 Minden, Penang, Malaysia
+60(0) 4 6533888 ext 4386
enyakong@cs.usm.my

ABSTRACT

In natural languages, a word can take on different meanings in different contexts. **Word sense disambiguation (WSD)** refers to the task of determining the correct meaning or sense of a word in context. It is an intermediate task essential to many natural language processing problems, including machine translation, information retrieval and speech processing. We focus on WSD in the context of machine translation. This involves resolving ambiguous words in the input sentence, so that words in a target language reflecting the correct senses can be chosen – a process also known as **target word selection**. The WSD module developed here will be used to complement and improve an existing example-based machine translation system. Ours will be a hybrid approach, drawing on a lexical ontology, which will be developed as part of this research and a bilingual knowledge-base of translated sentence pairs as the knowledge source. In addition, the lexical ontology to be constructed will be a reusable resource for other NLP tasks.

Keywords

Word sense disambiguation, machine translation, target word selection, lexical ontology.

1. INTRODUCTION

Word sense disambiguation (WSD) refers to the task of determining the correct meaning or sense of a word in context [7]. Approaches to WSD have moved from AI methods in the 1960's, to knowledge-based methods in the 1980's, which first used knowledge drawn from dictionaries and thesauri. Since the 1990's, statistical and corpus-based methods, which first existed in the late 19th century[4], have regained a strong popularity.

Rather than an isolated problem, WSD is an intermediate task which often forms part of other natural language processing (NLP) tasks, such as machine translation, information retrieval and speech synthesis. In particular, this research will focus on WSD in the context of machine translation.

2. RESEARCH MOTIVATION

Machine Translation (MT), or automatic translation, is the "application of computers to the translation of texts from one natural language into another" [3]. Given an input sentence in a source language (SL), an MT system must produce an output sentence in a target language (TL) that is both structurally and semantically correct. To fulfil the latter requirement, ambiguous words in the input sentence must be disambiguated, so that words in the TL reflecting the correct senses can be chosen – a process also known as **target word selection**.

Following the limited success of
ready availability of corpora resources, corpus-based methods for MT using parallel texts have become increasingly popular in recent years. Example-based Machine Translation (EBMT) uses existing translations, i.e. aligned pairs of sentences in the source language and target language, as a basis to translate new input sentences.

Al-Adhaileh and Tang's EBMT system [1] compares the dependency structure of an input sentence with those of translation examples in a Bilingual Knowledge Bank (BKB). This, in addition to part-of-speech information, solves the WSD and word selection problem to a certain extent – but still insufficient to produce satisfactory translation frequently enough. Hence, extending this system with a WSD module should improve the quality of the translation results. We propose a hybrid model, using the corpus in the said BKB (corpus-based) and a lexical ontology (knowledge-based) as our knowledge source.

3. LEXICAL ONTOLOGY

An ontology is an "explicit formal specifications of the terms in the domain and relations among them" [2]. It defines concepts, terms and vocabularies in a domain, and also the relationship among these concepts. Concepts are organized in a taxonomy, with sub-classes inheriting properties and specializing from super-classes. This similarity to the object-oriented paradigm serve to provide a knowledge representation that is naturally understood by humans while being machine-interpretable.

Ontologies are used to share common understanding of the structure of information, to enable reuse of domain knowledge, and to separate domain knowledge from operational knowledge [8]. Ontologies form a knowledge base together with instances created from defined classes. They can be a powerful tool for tasks such as translation, where information authored in a single language is converted for multiple target platforms, or for specification and development of software [5].

We propose the construction of a lexical ontology, where each lexical entry of a word's senses will contain a variety of linguistic and semantic information. We will use this information, together with the hierarchy and relationships in the ontology, to disambiguate an ambiguous input word, in order to select an appropriate translation word. In addition, the information contained in the ontology can also be reused for

USM Computer Science Postgraduate Colloquium 2004,
25-26th May 2004,
Penang, Malaysia.

[1] Direct approach (1960s), interlingua approach (1970s, and then late 1980s to early 1990s), and transfer approach (70s, 80s and 90s).

other NLP tasks, such as information retrieval and speech processing.

4. SUMMARY

We propose the construction of a lexical ontology as part of an approach to WSD in machine translation. Although many other researchers (e.g. [6] and [7]) have adopted similar methods, we will construct our ontology to suit an existing Example-Based Machine Translation system.

5. REFERENCES

[1] Al-Adhaileh, M. H. and Tang, E.K. (1999). Example-Based Machine Translation Based on the Synchronous SSTC Annotation Schema. In *Proceedings of Machine Translation Summit VII*, Singapore, pp 244-249.

[2] Gruber, T. (1993). A Translation Approach to Portable Ontology Specification. *Knowledge Acquisition, 5(2),* pp. 199-220.

[3] Hutchins, W.J. (1986). Machine Translation: Past, Present, Future. Ellis Horwood, Chichester, UK.

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

[5] Jasper, R. and Uschold, M. (1999). A Framework for Understanding and Classifying Ontology Applications. In *Proceedings of the 12th Banff Knowledge Acquisition for Knowledge Based Systems Workshop,* University of Calgary/Stanford University.

[6] Kang, S. J., Lee, J. H. (2001). Ontology-Based Word Sense Disambiguation by Using Semi-Automatically Constructed Ontology. In *Proceedings of MT Summit VIII,* Santialgo de Compostela, Galicia, Spain.

[7] Ng, H. T., Wang, B., Chan, Y. S. (2003). Exploiting Parallel Texts for Word Sense Disambiguation: An Empirical Study. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics,* Sapporo Convention Center, Sapporo, Japan, pp. 455-462.

[8] Noy, N. F. and McGuinness, D. L. (2001). Ontology Development 101: A Guide to Creating Your First Ontology. *Stanford Knowledge Systems Laboratory Technical Report KSL-01-05 and Stanford Medical Informatics Technical Report SMI-2001-0880,* Stanford University, Stanford.