Some Issues in the Construction of a Multi-Lingual Lexical-Semantic Net

Yu-de Bi
KORTERM, KAIST
305-701 Daejeon, Korea
biyude@gmail.com

Key-Sun Choi
KORTERM, KAIST
305-701 Daejeon, Korea
kschoi@cs.kaist.ac.kr

Jian-guo Xiong
Luoyang FLU
471003 Luoyang, China
jianguoxiong@163.com

Yang Liu
Peiking University
100084 Beijing, China
liuyang@pku.edu.cn

Abstract

An advanced knowledge base, such as a lexical-semantic net, guarantees accuracy in semantic interpretation and setting of semantic relations. The paper provides a tentative analysis of the concepts, relation representations, conceptual system and cross- language translation as observed in the construction of a multi-lingual lexical-semantic net.

1 Introduction

Ontology is a concept in philosophy, concerned with the nature and intrinsic relations of "being". As a kind of ontological product, WordNet sums up the achievements in the psychological studies of lexical knowledge and is playing an important role in NLP and information retrieval systems.

2 Significance and Objective

2.1 Significance

A lexical-semantic net, an integral part of information retrieval systems, presents itself as the primary resource for the next generation of the internet – the Semantic Web. It will provide important linguistic reference for cross-language studies and language teaching. Also it can help to improve the accuracy rate in multi-lingual information retrieval systems, text grouping and MT.

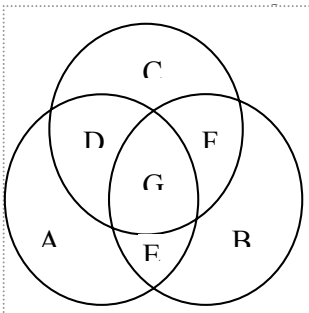
2.2 Objective

Our aim is to construct an independent and unified system. Its independence is reflected in the uniqueness of each language. Although the objective world features universality in most aspects, each language has its own names for its unique cultural phenomena due to variations in geography and culture. Therefore, independence of the conceptual system is necessary for each language. Unification facilitates cross-language communication. That is say, all independent systems must be unified on a certain basis.

3 The Issues

3.1 Preliminary remarks

In terms of ontology, concept is the mental image of objective beings. Due to cultural variations, the lexicon of each language differs. This is the so-called cultural-linguistic variation, which is reflected in conceptual vocabulary, as shown in the following diagram.



1. A--English;
   B--Korean;
   C--Chinese
2. D ∪ G=A∩C;
   E ∪ G=A∩B;
   F ∪ G=B∩C
3. G=A∩B∩C
   G—the common conceptual words for the three languages

Lexical coverage should be given primary attention, in order to reproduce the unique cultural properties of each individual language in the development of such systems, because lexical entries bear the characteristics of different languages and cultures.

3.2 Multi-lingual concept correspondence and relation representation

3.2.1 Multi-lingual concept correspondence

The first thing to consider is the establishment of concept correspondence between different languages. However, due to cultural variation, when translating from one language to another, we will unavoidably encounter lexical gaps, the non-correspondence of words and meanings between two languages. For example, WordNet has judo, but words like *taekwondo* or *wushu* (martial art) are not found in it. *Faro* has no corresponding concept in Korean or Chinese.

Doubtlessly, the differences between languages play a minor role as compared with universality. The basis of translatability is the common core of different cultures, the universality of human thinking and languages, the receptivity and open-mindedness of the interpreters and readers, as well as what the translation is all about.

In engineering practice, regarding those concepts for which there are only near corresponding words, we can use their parent concept as a temporary correspondent. For those concepts without correspondents, we can add sister

Petr Sojka, Key-Sun Choi, Christiane Fellbaum, Piek Vossen (Eds.): GWC 2006, Proceedings, pp. 293–294.

or daughter nodes as their corresponding coordinates or subordinates.

3.2.2 Relation representation

Compared with concept correspondence, the representation of the semantic relations in a multi-lingual lexical-semantic net is simpler. In fact, the major part of the system in WordNet can still be used.

3.3 Concept classification system

EuroWordNet inherits the structure of WordNet, with each language's unique properties properly represented. However, the structure and classification system of WordNet can hardly be directly applied to oriental languages. For example, in English there is a class of weather verbs like rain, snow, etc. The corresponding concepts are expressed in Korean with syntactic combinations, such as 비가 오다/rain, 눈이 오다/snow.

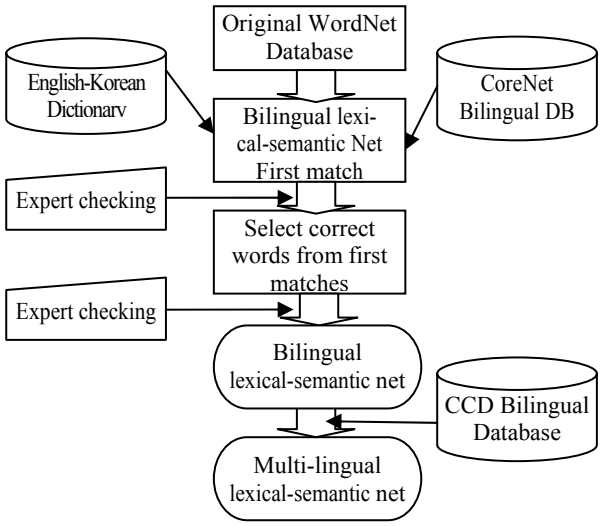
3.4 Cross-language translation

In natural languages, most words have more than one meaning. The diversity of lexical meanings brings difficulties for automatic matching. In the construction of the E-K lexical-semantic net, we extracted the Korean lexical items from CoreNet. We found that semi-automatic matching is possible. The process is divided into 3 steps:

Matching: Each English entry in WordNet synsets is used as the benchmark word. Then a bilingual electronic dictionary is retrieved for all the possible Korean wordsets. Next, all the candidate lexical items in those wordsets are extracted to set up a temporal DB.

Selection: Polysemes are checked manually, and one or more corresponding words are selected from the candidate word list.

Checking: All the selected meaning items are checked manually, and all the correct corresponding words are recorded in the relevant field in the WordNet database.



Conclusion

With the coming of the information age, and the rapid development of the internet, information exchange between different countries has become more and more important. The construction of a cross-language lexical-semantic net is a complicated job, involving various fields and disciplines. In all, computer scientists and linguists should cooperate to fulfill this task.

References

WordNet. http://wordnet.princeton.edu/
CoreNet. http://korterm.kaist.ac.kr
EuroWordNet.
http://www.illc.uva.nl/EuroWordNet.
CCD. http://icl.pku.edu.cn.