Linking CoreNet to WordNet - Some Aspects and Interim Consideration

In-Su Kang
Kyungsung Univ.
608-736 Busan
South Korea
dbaisk@ks.ac.kr

Sin-Jae Kang
Daegu Univ.
712-714 Daegu
South Korea
sjkang@daegu.ac.kr

Se-Jin Nam
KAIST
305-701 Daejeon
South Korea
jordse@gmail.com

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

Abstract

CoreNet, which is built on 2,937 semantic categories, is a multilingual lexico-semantic network aiming at bridging multiple languages/parts-of-speech for a variety of NLP applications. To foster its more widespread use, we have attempted to link semantic categories of CoreNet to Princeton WordNet. To ameliorate translation problems between CoreNet (mostly written in Korean) and English WordNet and to enhance recall of WordNet equivalents, two are partially indirectly linked through a Korean WordNet which shares most synset IDs with English WordNet. As an interim report, this paper describes a mapping methodology and current considerations.

1 Introduction

CoreNet (Choi and Bae, 2004) is a concept network of word senses, which has been constructed in KAIST for the Korean, Chinese, and Japanese languages since 1994 based on CoreNet concept hierarchy originated from NTT Goi-Taikei (Ikehara *et al.*, 1997) concept hierarchy. For Korean, it encompasses 31,384 general words (Biemann *et al.*, 2004), and a total of 62,632 senses of them are linked to one or more concepts in CoreNet concept hierarchy which is comprised of 2,937 high-level concepts mainly taxonomically organized into 12 depths. Unlike other lexico-semantic networks such as WordNet (Fellbaum, 1998) or Goi-Taikei (Ikehara *et al.*, 1997), CoreNet was designed to function as a single shared resource to bridge the semantics of not only different languages but also several parts-of-speech (POS) in the same language. For this, a single CoreNet concept hierarchy is used to link word senses of different lexical categories such as nouns, verbs, and adjectives for Korean, Chinese, and Japanese.

To start extending CoreNet's multi-lingualism into Indo-European languages and to promote its broader utilization for diverse NLP application, we have attempted to map CoreNet concept hierarchy to Princeton WordNet (PWN). This paper describes such efforts.

The rest of the paper is organized as follows. Section 2 presents a methodology to map between CoreNet and PWN. Section 3 describes current considerations while linking, and Section 4 gives a conclusion. For writing Korean expressions, Yale Romanization is used with a syllable delimiter '-'.

2 Mapping Methodology

2.1 Bottom-Up Mapping

There are two strategies in assigning the mapping-order to CoreNet concepts: top-down and bottom-up approaches. A top-down method starts from topmost nodes and determines mappings of higher-level nodes first and then lower-level nodes. A bottom-up one does the contrary.

When the former locates target-hierarchy equivalents of a source-hierarchy concept, it might be intractable to take into account all descendants of the source concept. Due to that, the top-down approach may produce the mapping result where target equivalents of source hyper-concepts do not include those of some source hypo-concepts as progeny. Although the bottom-up approach could avoid such inconsistency, it may yield numerous candidate-equivalents for higher-level source concepts, which would need to be generalized to a manageable number of higher concepts. However, such generalization can be accelerated by an automatic method such as discovering

nearest-common-ancestors from known offspring nodes within the target hierarchy.

A top-down scheme could be beneficial for the case where two concept hierarchies are similar. However, CoreNet and PWN are inherently different in languages, sizes, and concept granularities. Moreover, CoreNet concepts are not clearly defined compared to PWN synsets. These may raise problems in a downward mapping approach. Our team thus decided to employ a bottom-up method.

2.2 Indirect Mapping through KorLex

Since CoreNet concepts are described in non-English languages (mostly in Korean), linking CoreNet to PWN requires a language translation process. Translation between different languages in general involves disambiguating word senses, handling the cases that simple word-for-word correspondences do not exist.

To alleviate such difficulties in Korean-to-English (K-E) translation and to increase recall of PWN equivalents, we pursue to indirectly associate CoreNet with PWN through KorLex[1], a Korean WordNet. Although KorLex does not cover all PWN synsets or vice versa, overlapped synsets between KorLex and PWN contribute not only to reduce K-E translation problems but also to recall most synonymous K-E translations.

KorLex-based indirect-linking corresponds only to the case where a KorLex synset is found for a CoreNet concept and there exists a PWN synset of which ID is identical to that of the KorLex synset.

2.3 Mapping of Individual Concepts

To find PWN equivalents for a CoreNet concept, we exploit a CoreNet concept term (in Korean), a list of CoreNet concept words (in Korean), K-E dictionaries, and some utility programs such as PWN/KorLex/CoreNet browsers. First, for a CoreNet concept c and its concept term t_c, an equivalent-generation program produces a worksheet that includes the following.

- [I-1] Information (a concept term, a taxonomic hierarchy, concept words) of c
- [I-2] Information (taxonomic hierarchy, gloss, usage) of KorLex-noun/verb /adjective synsets $Ksyn_c$ such that at least one synonym of $Ksyn_c$ partially matches t_c.
- [I-3] Information (taxonomic hierarchy, gloss, usage) of PWN-noun/verb/adjective synsets $Esyn_c$ such that ID[2] of $Esyn_c$ is the same as that of $Ksyn_c$.
- [I-4] Information (taxonomic hierarchy, gloss, usage) of PWN-verb/noun/adjective synsets $Esyn_c$ such that at least one synonym of $Esyn_c$ partially matches one of English translations of t_c.

Each of [I-2], [I-3], and [I-4] may have multiple records for each POS (part-of-speech) such as noun, verb, and adjective. In [I-4], English translations are automatically obtained from Korean-to-English translation dictionary. The above worksheet is presented to human judges who review its contents to decide some PWN equivalents for c. If a human cannot find any related synsets within the worksheet, he/she looks up K-E dictionaries for relevant translations and browse synsets' information to select PWN equivalents of c by using PWN browsers.

As semantic relations used for linking between a CoreNet concept and a PWN synset, we employ the following seven relation types used in PWN: synonymy, hypernymy, hyponymy, troponymy, proper inclusion, presupposition, cause (Fellbaum, 1998). For instance, when a CoreNet concept corresponds synonymously to PWN synset {department_store,emporium}, its mapping is represented as follows.

{department_store,emporium^noun:03061806 ^synonym[3]}

3 Considerations

While associating CoreNet concepts to PWN synsets, the following considerations and issues have been raised. First, note that CoreNet concepts are designed to be applied to multiple POSs, but PWN concept system is divided into PWN-noun, PWN-verb, PWN-adjective, etc., according to POS. This requires us to link CoreNet to PWN separately for each of three parts-of-speech.

[1] KorLex is PWN-referenced Korean WordNet which has been developed since 2004, and it contains about 130,000 synsets and 150,000 word senses for nouns, verbs, adjectives, adverbs, and classifiers (Yoon et al., 2009).

[2] KorLex is constructed by using the synset IDs of PWN 2.0.

[3] {synset^POS:synset_ID^mapping_relation}

Second, some non-terminal concepts of CoreNet are represented in the form of enumerating child concept terms using a delimiter '/' (e.g. '*kyu-chik/pep-lyul/co-ngyak*' (rule/law/treaty), '*ngyun-li/cong-kyo*' (ethics/religion), '*mun-cang/ku/tan-nge*' (sentence/phrase/word), etc.) rather than a single concept term. So, it is non-trivial to find PWN equivalents for this type of concepts (called an enumeration-concept). For example, we have to search a PWN synset which signifies both 'ethics' and 'religion' for a CoreNet concept termed '*ngyun-li/cong-kyo*' (ethics/religion). Enumeration-concepts cover about 429 (14.6%) out of a total of 2,937 CoreNet concepts. To solve this, a set representation is introduced, which assigns an enumeration-concept a set of all PWN equivalents of its child concepts with a hyponymy relation. The mapping result for '*ngyun-li/cong-kyo*' (ethics/religion) is as follows.

{{morality^noun:04614989^hyponymy}, {religion,faith^noun:07591116^hyponymy}}

Third, there are CoreNet concepts termed '*pin-pu*' (rich/poor), '*cu-kayk*' (host/guest), etc. This type of concepts (called an antonyms-concept) exists as a single word in Korean but contains both meanings of antonym words in English. Antonyms-concepts cover about 29 (1%) out of entire CoreNet concepts. For this, we use the set representation again to encompass all PWN equivalents of each antonym. For example, the mapping result for '*cu-kayk*' (host/guest) is as follows.

{{host^noun:09530955^hyponymy},
{guest,invitee^noun:09498008^ hyponymy}}

Forth, there are CoreNet concepts called a complementary-concept. For concept *c*, this refers to the remnant of the scope of *c* that all children concepts of *c* specify. Complementary-concepts cover about 184 (6.3%) among all CoreNet concepts. For example, 'the other workers' concept in Figure 1 is a complementary concept.

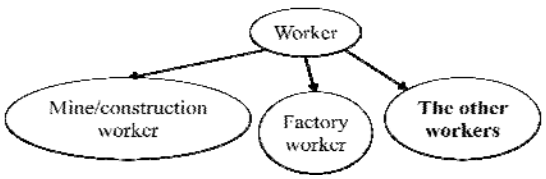

Figure 1. Example of a complementary concept

This type of concepts is mapped to the synset of its parent concept with a hypernymy relation. For example, the mapping result for 'The other workers' in the above is as follows.

{worker^noun:09025575^hypernymy}

Note that 'Worker' in Figure 1 is linked to PWN synset ID 09025575.

Fifth, there are some phrasal-concepts, which are hard to find their PWN equivalents (e.g. '*ka-kyek-ngin-sang*' (price advance), etc). In this case, we first find the headword of a phrasal-concept term, and map the phrasal-concept to synsets corresponding to the headword with a hypernymy relation. For '*ka-kyek-ngin-sang*' (price advance), its headword is '*ngin-sang*' (advance), which is mapped to PWN synset {advance,gain} with synset ID 00152358. Thus, the mapping result for '*ka-kyek-ngin-sang*' (price advance) is as follows.

{advance,gain^verb:00152358^hypernymy}

4 Conclusion

We have almost completed mapping terminal concepts (2,100) and non-terminal concepts (837) in CoreNet, except the cases belonging to the previously mentioned issues. After completing our work, we plan to distribute CoreNet with mappings to PWN in the form of LMF (Lexical Markup Framework)[4], the ISO standard for Natural Language Processing (NLP) lexicons and Machine Readable Dictionaries (MRD).

References

Aesun Yoon, Soonhee Hwang, Eunryoung Lee, and Hyuk-Chul Kwon. (2009). "Construction of Korean Wordnet KorLex 1.5". *Journal of KIISE (Korean Institute of Information Scientists and Engineers): Software and Applications*, 36(1):92-108.

Chris Biemann, Sa-Im Shin, Key-Sun Choi. (2004). "Semiautomatic extension of CoreNet using a bootstrapping mechanism on corpus-based co-occurrences". *Proceedings of the 20th International Conference on Computational Linguistics*, Geneva, Switzerland.

Christiane Fellbaum. (1998). WordNet: An Electronic Lexical Database (Language, Speech, Communication), MIT Press.

[4] http://www.lexicalmarkupframework.org/

Key-sun Choi and Hee-sook Bae. (2004). "Procedures and Problems in Korean-Chinese-Japanese Wordnet with Shared Semantic Hierarchy". *Proceedings of Global WordNet Conference*, Brno, Czech Republic, pp.91-96.

Satoru Ikehara, *et al*. (1997). The Semantic System, volume 1 of *Goidaikei: A Japanese Lexicon*, Iwanami Shoten, Tokyo.