Linking Korean Words with an Ontology

Min-Jae Kwon, Hae-Yun Lee, and Hee-Rahk Chae
Department of Linguistics and Cognitive Science
Hankuk University of Foreign Studies, Korea
E-mail: minjae.kwon@gmail.com, {haeyun, hrchae}@hufs.ac.kr

Abstract

This paper describes our ongoing work on linking Korean word senses with the concepts of an ontology. We have few Korean wordnets which are linked to upper-level ontologies, although the need for such wordnets/ontologies has increased not only in the academic world but also in the industry. We present a method for linking Korean senses with the concepts of SmartSUMO, which uses various language resources such as a bilingual dictionary, Princeton WordNet, and a thesaurus.

1. Introduction

The need for ontologies has increased in computer science or information science recently. Especially, NLP systems such as information retrieval, machine translation, etc. require ontologies whose concepts are connected to natural language words. Reflecting this trend, there have been many works which have focused on linking an ontology with lexical resources like WordNet (Gangemi et al. 2003, Niles and Pease 2003, Prévot et al. 2005, Reed and Lenat 2002).

There are a few Korean wordnets such as U-WIN, KorLex, CoreNet, etc. Most of them, however, stand alone without any link to an ontology, except for CoreNet (Korterm 2005). CoreNet makes use of the NTT-taxonomy as an ontology. As the taxonomy is based on the Japanese concept-system, linking Korean words to it leads to a result which does not fit with the intuition of Korean native speakers. Hence, we need a Korean wordnet which is linked to a language-neutral ontology such as SUMO, OpenCyc, DOLCE, etc.

In this paper, we will present a method of linking Korean word senses with the concepts of an ontology, which is part of an ongoing project.[1] We use a Korean-English bilingual dictionary, Princeton WordNet (Fellbaum 1998), and the ontology SmartSUMO (Oberle et al. 2007). The current version of WordNet is mapped into SUMO, which constitutes a major part of SmartSUMO. We have focused on mapping Korean word senses with corresponding English word senses by way of WordNet. Our work will lead to an extended version of SmartSumo, which reflects the conceptual system of Korean native speakers.

This paper is organized as follows. In section 2, we will give an overview of the architecture of our work. In section 3, we will introduce main algorithms of the work with application examples. Thereafter, in section 4, we will show an updated version of SmartSUMO, which is a result of the mapping between Korean word senses with SmartSumo concepts. Lastly, in section 5, we will summarize our work and outline future works to be continued.

2. Architecture

In this section, we present the overall architecture of our work and explain each stage with reference to language resources used. The whole picture of the architecture for building an ontology aligned with Korean word senses can be represented as in Figure 1.

2.1 Frequency list of Korean words

In the first stage, we get a word list which is extracted from the “Sejong-Corpus” (10 million words). The current work uses 20,000 high frequency nouns from the list. Later, we will work on words of other categories such as verbs, adjectives, and adverbs.

2.2 SmartSUMO

We chose SmartSUMO as the ontology to work on. SmartSUMO is an essential part of SWIntO (the “SmartWeb Integrated Ontology”) which contains several domain ontologies as well (Oberle et al. 2007). SmartSUMO, as an upper-level ontology within SWIntO, provides a rich taxonomy of concepts and predefined axioms. After examining various ontologies, the SWIntO project team made SmartSUMO by combining SUMO (Niles and Pease 2001) and DOLCE (Masolo et al. 2003). They pruned a top-level part of the SUMO taxonomy and aligned the remaining part with appropriate top-level categories of DOLCE. The resulting SmartSUMO contains about 950 concepts and 400 relations.

2.3 Language resources

For the mapping between Korean words and SmartSUMO, we make use of some language resources. Our mapping procedure depends basically on the information from a Korean-English bilingual dictionary, “PRIME Korean-English Dictionary (KED)”. This dictionary contains about 120,000 lemmas and provides Korean-English word correspondences and some examples as shown in Figure 2.

[1] The project, “A Study of the Language-Neutral Ontology” is supported by the National Research Foundation of Korea (2006-2012).

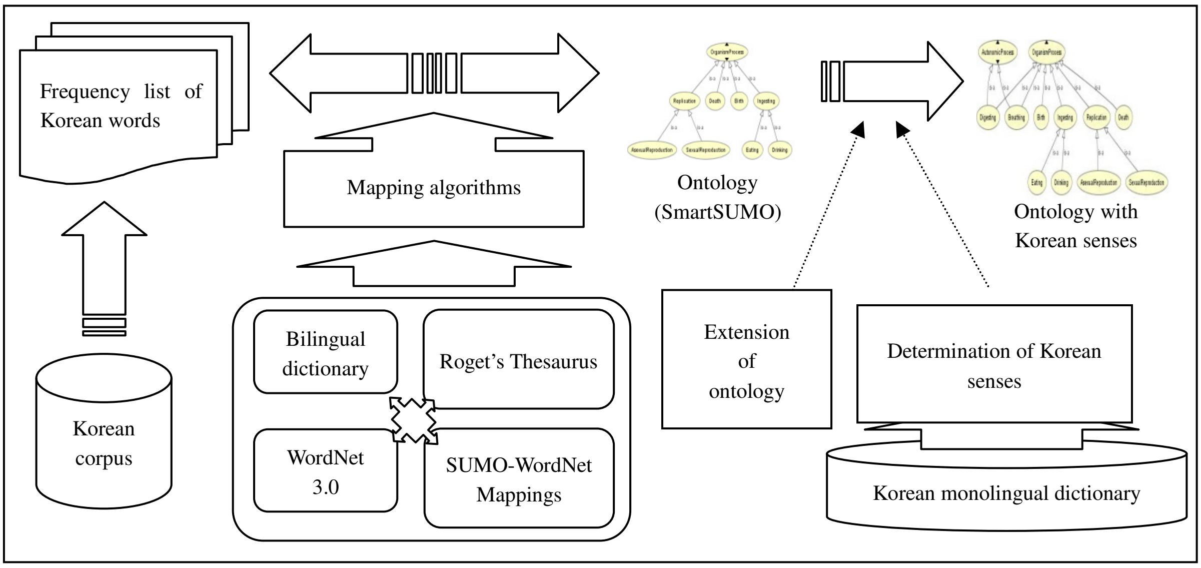

Figure 1: Architecture of linking Korean word senses with an ontology

의자 ; Korean word
a chair, a sofa, a settee, a lounge, … ; English words

의자에 걸터 앉다 sit on a chair ; Kor.-Eng.examples
의자를 한 줄로 늘어놓다 arrange chairs in a row
…

Figure 2: A lexical item in KED

In addition to the bilingual dictionary, we use WordNet 3.0 and Roget's Thesaurus (Roget 1911).[2] WordNet is already linked to SUMO (Niles and Pease 2003). Therefore, if we map Korean word senses to WordNet synsets, we can get a (indirect) mapping of Korean words to SUMO and finally to SmartSUMO. During the mapping process, the thesaurus is used, which will be explained in the next section.

2.4 After the mapping

When the mapping is completed, we get a revised version of SmartSUMO, to which those concepts reflecting Korean word senses will have been attached. At this stage, we perform two sub-tasks. Firstly, we extend the ontology, if necessary, on the basis of the alignments between Korean word senses and SmartSUMO concepts. Secondly, we must (re-)assign the word forms of a lexical entry according to their senses. Reflecting the concepts under which word forms are allotted, we determine manually the sense number of each word form in the Korean monolingual dictionary "Standard Korean Dictionary (SKD)".

3. Mapping Procedure

In this section, we will introduce the algorithms which map Korean word senses to English word senses automatically. Basically, we make use of a Korean-English bilingual dictionary as a bridge between Korean lemmas and SUMO concepts, as in Okumura and Hovy (1994). Because of the differences in the content and format of the dictionaries used, we need a different set of more sophisticated algorithms than those in Okumura and Hovy (1994).

First of all, the types of matching can be divided into four groups, as follows. We will explain algorithms which are applied to each type.

- Type 1: one Korean lemma – one English word with one WordNet synset
- Type 2: one Korean lemma – more than one English word, all of which have the same WordNet synset
- Type 3: one Korean lemma – one English word with more than one WordNet synset
- Type 4: one Korean lemma – more than one English word, which have different WordNet synsets

3.1 Type 1 & 2: direct matching

Type 1 and type 2 have one or more English word(s) corresponding to a Korean word, but the English word(s) are allocated to only one concept of SUMO. In these cases, we have no problems in matching. Table 1 shows an example of type 2.

[2] The 1911 version of Roget's Thesaurus is available from Project Gutenberg site (http://www.gutenberg.org/ebooks/22). It contains 8 Classes, 39 Sections, 97 Subsections, 625 Head Groups, 1044 Heads, 3934 POSs, 10244 Paragraphs, 43196 Semicolon Groups, 98924 Total Words and 59768 Unique Words.

Korean lemma	English word(s)	WordNet synset(s)	SUMO-Concept
경찰 [kyengchal]	police	S: (n) police, police force, constabulary, law	PoliceOrganization+
	police force	S: (n) police, police force, constabulary, law	

Table 1: An example of Type 2

3.2 Type 3: comparison of examples

Type 3 has one English word which has more than one synset in WordNet. Among those English synsets, we choose one synset compatible with the given Korean lemma by examining English examples. The analysis of English examples proceeds on the basis of Roget's Thesaurus (1911), which has been proved very useful for NLP (Jarmasz and Szpakowicz 2001, Jarmasz and Szpakowicz 2003, Kennedy and Szpakowicz 2008). We calculate the so-called "main path" for each example and then compare those main paths. If there is the most compatible main path both in the examples of KED and in the examples of WordNet, then we choose the synset of the examples that have that main path. Then, we can get the link between the Korean lemma and the SUMO-concept connected with the synset.

The main path is calculated as follows: There are nine levels in the Roget's Thesaurus hierarchy, from Class down to Word. We have numbered each branching node as path from the highest level to the lowest level in the hierarchy. For example, the word *car* belongs to the category (head word) "Vehicle (272)" with the path "1.2.3.1.9". The path "1.2.3.1.9" indicates the route of the category "Vehicle" from the top: Roget's Thesaurus (1) → CLASS II. WORDS RELATING TO SPACE (1.2) → SECTION IV. MOTION (1.2.3) → Motion in General (1.2.3.1) → Vehicle 272 (1.2.3.1.9).

The algorithm proceeds as follows, and Table 2 shows an example of this algorithm.

- Extracting key words from each example of KED (We exclude the corresponding lemma, functional words, and overlapped words).
- Calculating the main path of each example on the basis of the extracted key words.
- Extracting key words from each example of synsets in WordNet (as from KED examples).
- Calculating the main path of each example (as in KED examples).
- Comparing the main path of KED with the main paths of WordNet, and choosing one synset whose examples has the closest main path to the main path of KED.

3.3 Type 4: comparison of WordNet synsets

Type 4 has more than one English word per Korean lemma, and each English word has one or more than one synset. In this case, we can use two methods depending on the situations. Firstly, we can use the method in Okumura and Hovy (1994). That is, we compare synsets of all the corresponding English words. If there is a synset which has the maximal number of common senses, we choose that synset as the most compatible English sense. The Table 3 shows the application of this sub-type. However, there are cases where it is difficult to choose one synset. For example, if we have two or more synsets which have the same maximal number of common senses, we cannot apply the present method. In this case, however, we can make use of the method which is applied to Type 3 examples, i.e. the method of example comparison.

4. Extension of SmartSUMO

As a result of the mapping process described above, we get an ontology linked with Korean lemmas. But if we examine the distribution of Korean lemmas in the ontology, we see some cases where too many lemmas to be differentiated are allocated to just one concept. In those cases, we have to split the concept into a few sub-concepts. For example, SmartSUMO has only one concept ('EmotionalState') which is related to words of emotion. As a result, that concept comprises many diverse Korean lemmas, as in (1).

(1) 분노 (anger), 성 (indignation), 화 (rage), …; 사랑 (love), 애정 (affection), …; 증오 (hatred), 질투 (jealousy), …; 기쁨 (joy), 희열 (delight), …; 두려움 (fear), 겁 (cowardice), …; 슬픔 (sorrow), …

However, ethnological and/or psychological researches show that human beings have emotions like *anger*, *love*, *hatred*, *joy*, *fear*, and *sorrow* in common. Therefore, we need to create some sub-concepts under the concept 'EmotionalState.' According to the specification of SmartSUMO, we can present some new concepts in RDF as in (2).

(2)
```
  <rdfs:Class rdf:about=
    "http://smartweb.semanticweb.org/ontology/
     smartsumo#Anger"
     rdfs:label="smartsumo:EmotionalState">
     <rdfs:subClassOf rdf:resource=
     "http://smartweb.semanticweb.org/ontology/
       smartsumo#EmotionalState" />
  </rdfs:Class>
  <rdfs:Class rdf:about=
    "http://smartweb.semanticweb.org/ontology/
     smartsumo#Love"
     rdfs:label="smartsumo:EmotionalState">
     <rdfs:subClassOf rdf:resource=
     "http://smartweb.semanticweb.org/ontology/
       smartsumo#EmotionalState" />
  </rdfs:Class>
```

	Lemma/synsets	Examples	Main path	SUMO-Concept
KED	은행[unhayng] - bank	deposit money in a bank; open[close] an account with a bank; draw money from a bank; have a bank account of $10,000 at a bank	1.5.2.4.4.1	Corporation+
WordNet	**bank**#1	they pulled the canoe up on the bank; he sat on the bank of the river and watched the currents	1.4.1.2	ShoreArea=
	depository financial institution, **bank#2**, banking concern, banking company	he cashed a check at the bank; that bank holds the mortgage on my home	1.5.2.4.4.1	Corporation+
	bank#3	a huge bank of earth	1.3	UplandArea+
	bank#4	he operated a bank of switches	1.1.8.3	Collection+
	bank#5	--	--	Keeping+
	bank#6	he tried to break the bank at Monte Carlo	1.5.2	CurrencyMeasure+
	bank#7	--	--	LandArea+
	savings bank, coin bank, money box, **bank#8**	the coin bank was empty	1.5.1.2.2.1 / 1.2.1	SafeContainer+
	bank#9, bank building	the bank is on the corner of Nassau and Witherspoon	1.2 / 1.5	Building+
	bank#10	the plane went into a steep bank	1.2.2.3.3	Motion+

Table 2: An example of Type 3

Korean lemma	English word(s)	WordNet synset	SUMO-Concept
바다[pata]	sea	sea#1	
		ocean#2, sea#2	SubjectiveAssessmentAttribute+
		sea#3	
	ocean	ocean#1	
		ocean#2, sea#2	SubjectiveAssessmentAttribute+

Table 3: An example of Type 4

5. Determining Korean senses

In the procedure of mapping, we used the bilingual dictionary, in which senses of Korean lemmas are not differentiated. Therefore, we had to make a distinction between the senses of Korean words to complete the linking procedure.

We chose one among the senses which the Korean monolingual dictionary (SKD) provides, making reference to the mapping result at the previous stage. Let's see the determination procedure with an example of *unhayng* (cf. Table 4). SKD provides 3 senses for the lemma *unhayng*. But we got the information that the Korean lemma corresponds to the concept 'Corporation'. Based on such a mapping result, we can allot the SKD-based sense number to the lemma *unhayng*.

6. Conclusion

In this paper, we have shown an automatic method of linking Korean word senses with SmartSUMO concepts by using a bilingual dictionary. We saw that we need to apply different algorithms of linking, depending on the information types that the given Korean-English word pairs contain.

As our ongoing project aims to construct an ontology which is linked to most of the Korean word senses, we are going to apply the methods introduced in this paper into other syntactic categories such as verbs, adjectives, and adverbs in the future

Korean lemma	SUMO-Concept	SKD	
		senses	definition
은행[unhayng]	**Corporation**	은행 01	a vassal who finds favor with his sovereign
		은행 02	**a bank (as Corporation)**
		은행 03	a gingko nut

Table 4: An example of determining Korean senses

7. References

Fellbaum, C. ed. (1998). *WordNet: An Electronic Lexical Database*. Cambridge, MA: MIT Press.

Gangemi, A., Guarino, N., Masolo, C., and Oltramari, A. (2003). Sweetening WordNet with DOLCE. *AI Magazine* 3: 13-24.

Jarmasz, M. and Szpakowicz, S. (2001). Roget's Thesaurus: a Lexical Resource to Treasure. In *The NAACL WordNet and Other Lexical Resources Workshop*. Pittsburgh.

Jarmasz, M. and Szpakowicz, S. (2003). Roget's Thesaurus and Semantic Similarity. In *Conference on Recent Advances in Natural Language Processing (RANLP 2003)*. Borovets, Bulgaria.

KED. (2005). *Prime Korean-English Dictionary,* Seoul: Dong-A Press.

Kennedy, A. and Szpakowicz, S. (2008). Evaluating Roget's Thesauri. In *ACL-08: HLT*. Columbus, Ohio, USA.

Korterm. (2005). *CoreNet: Core Multilingual Semantic Word Net*. Daejeon: KAIST Press.

Masolo, C., Borgo, S., Gangemi, A., Guarino, N., Oltramari, A., and Schneider, L. (2003). WonderWeb Deliverable D17: ISTC-CNR.

Niles, I. and Pease, A. (2001). Towards a Standard Upper Ontology. Paper presented at *the 2nd International Conference on Formal Ontology in Information Systems*.

Niles, I. and Pease, A. (2003). Linking Lexicons and Ontologies: Mapping Wordnet to the Suggested Upper Merged Ontology. Paper presented at *the 2003 International Conference on Information and Knowledge Engineering (IKE 03)*, Las Vegas, Nebada.

Oberle, D., Ankolekar, A., Hitzler, P., Cimiano, P., Sintek, M., Kiesel, M., Mougouie, B., Baumann, S., Vembu, S., Romanelli, M., Buitelaar, P., Engel, R., Sonntag, D., Reithinger, N., Loos, B., Zorn, H., Micelli, V., Porzel, R., Schmidt, C., Weiten, M., Burkhardt, F., and Zhou, J. (2007). DOLCE ergo SUMO: On foundational and domain models in the SmartWeb Integrated Ontology (SWIntO). *Web Semantics: Science, Services and Agents on the World Wide Web* 5: 156-174.

Okumura, A. and Hovy, E. (1994). Ontology-Concept Association using a Bilingual Dictionary. Paper presented at *the 1st AMTA Conference*, Columbia, MD.

Prévot, L., Borgo, S., and Oltramari, A. (2005). Interfacing Ontologies and Lexical Resources. Paper presented at *the 2nd International Joint Conference on Natural Language Processing (IJCNLP-05)*, Jeju, Korea.

Reed, S. and Lenat, D. (2002). Mapping ontologies into cyc. In *AAAI 2002 Conference Workshop on Ontologies For The Semantic Web*. Edmonton, Canada.

Roget, P. M. (1911). *Roget's Thesaurus*. Burnt Mill, Harlow, Essex: Longman Group Limited.

SKD. (1999). *The Standard Korean Dictionary*, Seoul: Dong-A Press.