Considerations on Automatic Mapping Large-Scale Heterogeneous Language Resources: *Sejong Semantic Classes* and *KorLex*

Heum Park
Center for U-Port IT
Research and Education
Pusan National University
parheum2@empal.com

Aesun Yoon
LI Lab. Dept. of French
Pusan National University
asyoon@pusan.ac.kr

Woo Chul Park and Hyuk-Chul Kwon*
AI Lab Dept. of Computer Science
Pusan National University
hckwon@pusan.ac.kr

Abstract

This paper presents an automatic mapping method among large-scale heterogeneous language resources: *Sejong Semantic Classes* (SJSC) and *KorLex*. KorLex is a large-scale Korean WordNet, but it lacks specific syntactic & semantic information. *Sejong Electronic Dictionary* (SJD), of which semantic segmentation depends on SJSC, has much lower lexical coverage than KorLex, but shows refined syntactic & semantic information. The goal of this study is to build a rich language resource for improving Korean semantico-syntactic parsing technology. Therefore, we consider integration of them and propose automatic mapping method with three approaches: 1) Information of Monosemy/Polysemy of Word senses (IMPW), 2) Instances between Nouns of SJD and Word senses of KorLex (INW), and 3) Semantically Related words between Nouns of SJD and Synsets of KorLex (SRNS). We obtain good performance using combined three approaches: recall 0.837, precision 0.717, and F1 0.773.

1 Introduction

While remarkable progress has been made in Korean language engineering on morphological level during last two decades, syntactic and semantic processing has progressed more slowly. The syntactic and semantic processing requires 1) linguistically and formally well defined argument structures with the selectional restrictions of each argument, 2) large and semantically well segmented lexica, 3) most importantly, interrelationship between the argument structures and lexica. A couple of language resources have been developed or can be used for this end. *Sejong Electronic Dictionaries* (SJD) for nouns and predicates (verbs and adjectives) along with semantic classes (Hong 2007) were developed for syntactic and semantic analysis, but the current versions do not contain enough entries for concrete applications, and they show inconsistency problem. A Korean WordNet, named *KorLex* (Yoon & al, 2009), which was built on Princeton WordNet 2.0 (PWN) as its reference model, can provide means for shallow semantic processing but does not contain refined syntactic and semantic information specific to Korean language. *Korean Standard Dictionary* (STD) provides a large number of entries but it lacks systematic description and formal representation of word senses, like other traditional dictionaries for humans. Given these resources which were developed through long-term projects (5 – 10 years), integrating them should result in significant benefits to Korean syntactic and semantic processing.

The primary goal of our recent work including the work reported in this paper is to build a language resource, which will improve Korean semantico-syntactic parsing technology. We proceed by integrating the argument structures as provided by SJD, and the lexical-semantic hierarchy as provided by KorLex. SJD is a language resource, of which all word senses are labeled according to *Sejong semantic classes* (SJSC), and in which selectional re-

* Corresponding Author

Proceedings of the 8th Workshop on Asian Language Resources, pages 14–21,
Beijing, China, 21-22 August 2010. ©2010 Asian Federation for Natural Language Processing

strictions are represented in SJSC as for the argument structures of predicates. KorLex is a large scale language resource, of which the lexical-semantic hierarchies and other language–independent semantic relations between synsets (synonym sets) share with those of PWN, and of which Korean language specific information comes from STD. The secondary goal is the improvement of three resources as a result of comparing and integrating them.

In this paper, we report on one of the operating steps toward to our goals. We linked each word sense of KorLex to that of STD by hand, when the former was built in our previous work (Yoon & al. 2009). All predicates in SJD were mapped to those of STD on word sense level by semi-automatic mapping (Yoon, 2010). Thus KorLexVerb and KorLexAdj have syntactico-semantic information on argument structures via this SJD - STD mapping. However, the selectional restrictions provided by SJD are not useful, if SJSC which represents the selectional restrictions in SJD is not linked to KorLex. We thus conduct two mapping methods between SJSC and upper nodes of KorLexNoun: 1) manual mapping by a PH.D in computational semantics (Bae & al. 2010), and 2) automatic mapping. This paper reports the latter. Reliable automatic mapping methods among heterogeneous language resources should be considered, since the manual mapping among large-scale resources is a very time and labor consuming job, and might lack consistency. Less clean resources are, much harder and more confusing manual mapping is.

In this paper, we propose an automatic mapping method of those two resources with three approaches to determine mapping candidate synsets of KorLex to a terminal node of SJSC: 1) using information of monosemy/polysemy of word senses, 2) using instances between nouns of SJD and word senses of KorLex, and 3) using semantically related words between nouns of SJD and word senses of KorLex. We compared the results of automatic mapping method with three approaches with those of manual mapping aforementioned.

In the following Section 2, we discuss related studies concerning language resources and automatic mapping methods of heterogeneous language resources. In Section 3, we introduce KorLex and SJD. In Section 4, we propose an automatic mapping method with three approaches from semantic classes of SJD to synsets of KorLex. In Section 5, we compare the results of automatic mapping with those of manual mapping. In Section 6, we draw conclusions and future works.

2 Related Works

Most existing mappings of heterogeneous language resources were conducted manually by language experts. The Suggested Upper Merged Ontology (SUMO) had been fully linked to PWN. For manual mapping of between PWN and SUMO, it was considered synonymy, hypernymy and instantiation between synsets of PWN and concepts of SUMO, and found the nearest instances of SUMO for synsets of PWN. Because the concept items of SUMO are much larger than those of PWN, it could be mapped between high level concepts of PWN and synonymy concepts of SUMO easily. (Ian Niles et al 2003). Dennis Spohr (2008) presented a general methodology to mapping EuroWordNet to the SUMO for extraction of selectional preferences for French. Jan Scheffczyk et al. (2006) introduced the connection of FrameNet to SUMO. They presented general-domain links between FrameNet Semantic Types and SUMO classes in SUOKIF and developed a semi-automatic, domain-specific approach for linking FrameNet Frame Elements to SUMO classes (Scheffczyk & al. 2006). Sara Tonelli et al. (2009) presented a supervised learning framework for the mapping of FrameNet lexical units onto PWN synsets to solve limited coverage of semantic phenomena for NLP applications. Their best results were recall 0.613, precision 0.761 and F1 measure 0.679.

Considerations on automatic mapping methods among language resources were always attempted for the sake of efficiency, using similarity measuring and evaluating methods. Typical traditional evaluating methods between concepts of heterogeneous language resources were the dictionary-based approaches (Kozima & al 1993), the semantic distance algorithm using PWN (Hirst & al 1998), the scaling method by semantic distance between concepts (Sussna 1997), conceptual similarity between concepts (Wu & al 1994), the scaled semantic similarity between concepts (Leacock 1998), the semantic similarity between concepts using IS-A relation (Resnik 1995), the measure of similarity between concepts (Lin 1998), Jiang and Conrath's (1997) similarity computations to synthesize edge and node based techniques, etc.

Satanjeev et al. (2003) presented a new measure of semantic relatedness between concepts that was based on the number of shared words (overlaps) in their definitions (glosses) for word sense disambiguation. The performances of their extended gloss overlap measure with 3-word window were recall 0.342, precision 0.351 and F1 0.346. Siddharth et al. (2003) presented the Adapted Lesk Algorithm to a method of word sense disambiguation based on semantic relatedness. In addition, Alexander et al (2006) introduced the 5 existing evaluating methods for PWN-based measures of lexical semantic relatedness and compared the performance of typical five measures of semantic relatedness for NLP applications and information retrieval. Among them, Jiang-Conrath's method showed the best performances: precision 0.247, recall 0.231 and F1 0.211 for Detection.

In many studies, it was presented a variety of the adapted evaluating algorithms. Among them, Jiang-Conrath's method, Lin's the measure of similarity and Resnik's the semantic similarity show good performances (Alexander & al 2006, Daniele 2009).

3 Language resources to be mapped

3.1 KorLex 1.5

KorLex 1.5 was constructed from 2004 to 2007. Different from its previous version (KorLex 1.0) which preserves all semantic relations among synsets of PWN, KorLex 1.5 modifies them by deletion/correction of existing synsets, addition of new synsets and conversion of hierarchical structure. Currently, KorLex includes nouns, verbs, adjectives, adverbs and classifiers: KorLexNoun, KorLexVerb, KorLexAdj, KorLexAdv and KorLexClas, respectively. Table 1 shows the size of KorLex 1.5, in which 'Trans' means the number of synsets translated from PWN 2.0 and 'Total' is the number of manually added synsets including translated ones.

	Word Forms	Synsets		Word Senses
		Trans	Total	
KorLexNoun	89,125	79,689	90,134	102,358
KorLexVerb	17,956	13,508	16,923	20,133
KorLexAdj	19,698	18,563	18,563	20,905
KorLexAdv	3,032	3,664	3,664	3,123
KorLexClas	1,181	-	1,377	1,377
Total	**130,992**	**115,424**	**130,661**	**147,896**

Table 1. Product of KorLex 1.5

KorLexNoun includes 25 semantic domains with 11 unique beginners with maximum 17 levels in depth and KorLexVerb includes 15 semantic domains with 11 unique beginners with maximum 12 levels in depth. Basically, KorLex synsets inherit the semantic information of PWN synsets mapped to them. The synset information of PWN consists of synset ID, semantic domain, POS, word senses, semantic relations, frame information, and so on.

We linked each word sense of KorLex 1.5 to that of STD by hand, when the former was built in our previous work (Yoon & al. 2009). STD includes 509,076 word entries with about 590,000 word senses. It contains a wide coverage for general words and a variety of example sentences for each meaning. More than 60% of word senses in KorLex 1.5 are linked to those of STD. KorLex 1.5, thus, inherits lexical relations described in STD, but both resources lack refined semantic-syntactic information.

3.2 Sejong Electronic Dictionary

SJD was developed during 1998-2007 manually by linguists for a variety of Korean NLP application as a general-purpose machine readable dictionary. Based on *Sejong semantic classes* (SJSC), approximately 25,000 nouns and 20,000 predicates (verbs and adjectives, SJPD) contain refined syntactic and semantic information.

SJSC is a set of hierarchical meta-languages classifying word senses and it includes 474 terminal nodes and 139 non-terminal nodes, and 6 unique beginners. Each unique beginner has levels from minimum 2 to maximum 7 levels in depth. Sejong Noun Dictionary (SJND) contains 25,458 entries and 35,854 word senses having lexical information for each entry: semantic classes of SJSC, argument structures, selectional restrictions, semantically related words, derivatioinal relations/words et al.

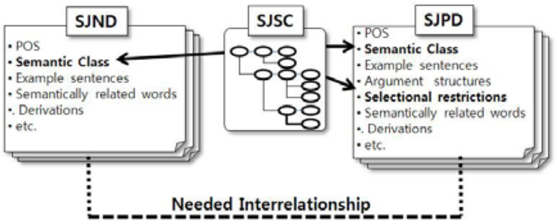


Figure 1. Correlation of lexical information among SJND, SJPD and SJSC

Figure 1 shows the correlation of lexical information among SJND, SJPD and SJSC. Certainly, that information of SJD should be applied to a variety of NLP applications: information retrieval, text analysis/generation, machine translations, and various studies and educations. However, SJD has much lower lexical coverage than KorLex. More serious problem is that SJND and SJPD are still noisy: internal consistency inside each dictionary and external interrelationship between SJND, SJPD, and SJSC need to be ameliorated, as indicated by dot line in Fig. 1.

4 Automatic Mapping from Semantic Class of SJSC to Synsets of KorLex

KorLex and SJSC have different hierarchical structures, grain sizes, and lexical information as aforementioned. For example, the semantic classes of SJSC are much bigger concepts in grain size than the synsets of KorLex: 623 concepts in SJSC vs.130,000 synsets in KorLex. Determining their semantic equivalence thus needs to be firmly based on linguistic clues.

Using following 3 linguistic clues that we found, we propose an automatic mapping method from semantic classes of SJSC to synsets of KorLex with three approaches to determine mapping candidate synsets: 1) Information of Monosemy/Polysemy of Word senses (IMPW), 2) Instances between Nouns of SJD and Word senses of KorLex (INW), and 3) Semantically Related words between Nouns of SJD and Synsets of KorLex (SRNS).

For automatic mapping method, following processes were conducted. First, to find word senses of synsets that matched to nouns of SJND for each semantic class. Second, to select mapping candidate synsets among them with three approaches aforementioned. Third, to determine the least upper bound (LUB) synsets and mapping synsets among candidates. Finally, to link each semantic class of SJSC to all lower-level synsets of LUB synsets.

4.1 Finding matched word senses between synsets and nouns of SJND

For a semantic class of SJSC, we first find word senses and synsets from KorLex that matched with nouns of SJND classified to that semantic class. Figure 2 shows the matched word senses and synsets between nouns of SJND, then synsets of KorLex for a semantic class. The left side of Figure 2 shows nodes of semantic classes with hierarchical structure and the center box shows the matched words (bold ones) among nouns of SJND with word senses of synsets in KorLex, and the right side shows matched word senses and synsets in KorLex' hierarchical structure.

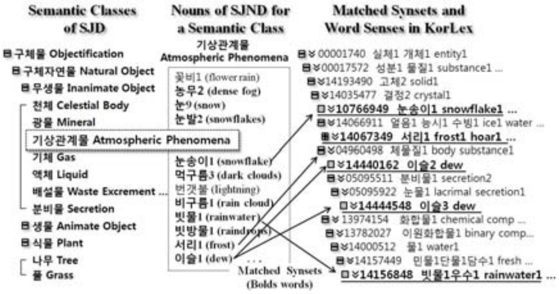

Semantic Classes of SJD
Nouns of SJND for a Semantic Class
Matched Synsets and Word Senses in KorLex
⊟ 구체물 Objectification
⊟ 구체자연물 Natural Object
⊟ 무생물 Inanimate Object
천체 Celestial Body
광물 Mineral
기상관계물 Atmospheric Phenomena
기체 Gas
액체 Liquid
배설물 Waste Excrement ...
분비물 Secretion
⊟ 생물 Animate Object
⊟ 식물 Plant
나무 Tree
풀 Grass
기상관계물 Atmospheric Phenomena
꽃비1 (flower rain)
농무2 (dense fog)
눈9 (snow)
눈발2 (snowflakes)
눈송이1 (snowflake)
먹구름3 (dark clouds)
번갯불 (lightning)
비구름1 (rain cloud)
빗물1 (rainwater)
빗방울1 (raindrops)
서리1 (frost)
이슬1 (dew)
...
Matched Synsets (Bolds words)
⊟ ¥00001740 실체1 개체1 entity1
⊟ ¥00017572 성분1 물질1 substance1 ...
⊟ ¥14193490 고체2 solid1
⊟ ¥14035477 결정2 crystal1
□ ¥10766949 눈송이1 snowflake1 ...
⊟ ¥14066911 얼음1 농시1 수빙1 ice1 water ...
⊞ ¥14067349 서리1 frost1 hoar1 ...
⊟ ¥04960498 체물질1 body substance1
□ ¥14440162 이슬2 dew
⊟ ¥05095511 분비물1 secretion2
⊟ ¥05095922 눈물1 lacrimal secretion1
□ ¥14444548 이슬3 dew
⊟ ¥13974154 화합물1 chemical comp ...
⊟ ¥13782027 이원화합물1 binary comp...
⊟ ¥14000512 물1 water1
⊟ ¥14157449 민물1단물1담수1 fresh ...
□ ¥14156848 빗물1우수1 rainwater1 ...

Figure 2. Matched word senses and synsets with nouns of SJND for a semantic class

For example, a semantic class 기상관련물 'Atmospheric Phenomena' (rectangle in the left) has nouns of SJND (words in the center), the bold words are the matched words with word senses of synsets from KorLex, and the underlined synsets of the right side are the matched ones and synset IDs in KorLex. The notations for automatic mapping process between semantic classes of SJSC and synsets of KorLex are as follows: noun of SJND is *ns*, matched noun ns_m, un-matched ns_u , semantic class of SJSC is *sc*, synset is *ss* and word sense of a synset is *ws* in KorLex, and monosemy word is w_{mono} and polysemy word is w_{poly}.

A semantic class *sc* has nouns *ns* of SJND having matched noun ns_m and un-matched ns_u by comparing with word senses *ws* of a synset *ss* in KorLex. Thus a synset has word senses as ss_1={ws_1, ws_2, …, ws_n}={ ns_{m1}, ns_{m2}, …, ns_{mk}, $ns_{u\ k+1}$, $ns_{u\ k+2}$, …}. And nouns of SJND for a semantic class sc_1 is presented $ns(sc_1)$={ns_{m1},

ns_{m2}, …, ns_{mk}, $ns_{u\,k}$, $ns_{u\,k+1}$, …}. Therefore, we can find the matched word senses ns_{m1} ~ ns_{mk} for a semantic class *sc* from nouns of SJND and word senses of a synset *ss* in KorLex.

4.2 Selecting Mapping Candidate Synsets

Using those matched synsets and word senses, we select mapping candidate synsets with three different approaches.

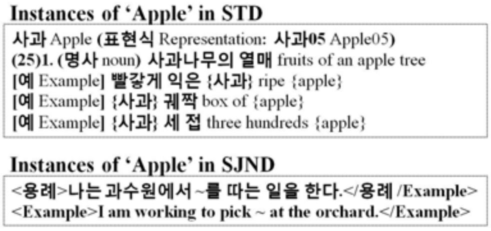
4.2.1 Using Information of Monosemy and Polysemy of KorLex

Using information of monosemy/polysemy of word senses of a synset, the first approach evaluates mapping candidate synsets. The candidate synsets are evaluated into three categories: *mc(A)* is a most relevant candidate synset, *mc(B)* is a relevant candidate synset and *mc(C)* is a reserved synset. Evaluation begins from lowest level synsets to top-level beginner. The process of first approach is as follows.

1) For a synset which contains a single word sense, *ss*={ws_1}, if the word sense is a monosemy, it is categorized as a a candidate synset *mc(A)*. If it is a polysemy, categorization is postponed for evaluating relatedness among siblings: candidate *mc(C)*.
2) In the case of a synset having more than one word sense, *ss*={ws_1, ws_2, …}, if the matched words ns_m among word senses of a synset are over 60%: $P_{ss}(ws)$=(count(ns_m)/ count(*ws*)) ≥ 0.6, we evaluate whether that synset is mapping candidate in the next step.
3) If all matched words ns_m of a synset are monosemic, we categorize it as a candidate synset *mc(A)*. If monosemic words among matched words are over 50%: $P_{ss}(w_{mono}|ns_m) \geq 0.5$, it is evaluated as a *mc(B)*. A synset containing polysemies over 50%: $P_{ss}(w_{poly}|ns_m) \geq 0.5$, categorization is postponed for evaluating relatedness among siblings: candidate synset *mc(C)*.
4) To repeat from step 1) to 3) for all of synsets, in order to evaluate mapping candidate synsets. And then, to construct hierarchical structure for all those synsets.

4.2.2 Using Instances between Nouns of SJND and Word senses of KorLex

The second approach is to evaluate mapping candidate synsets using comparison of instances between nouns of SJND and word senses of a synset. As for KorLex, we used the examples of STD linked to word senses of KorLex. Figure 3 shows instances of STD and SJND for a word sense 'Apple'.


Instances of 'Apple' in STD

사과 Apple (표현식 Representation: 사과05 Apple05)
(25)1. (명사 noun) 사과나무의 열매 fruits of an apple tree
[예 Example] 빨갛게 익은 {사과} ripe {apple}
[예 Example] {사과} 궤짝 box of {apple}
[예 Example] {사과} 세 접 three hundreds {apple}

Instances of 'Apple' in SJND

<용례>나는 과수원에서 ~를 따는 일을 한다.</용례 /Example>
<Example>I am working to pick ~ at the orchard.</Example>

Figure 3. Instances of STD and SJND for word sense 'Apple' 사과

We reformulated the Lesk algorithm (Lesk 1987, Banerjee and Pedersen 2002) for comparing instances and evaluating mapping candidate synsets. The process of evaluating mapping candidate synsets is as follows.

1) To compare instances of a noun *ns* of SJND with examples of a word of STD linked to word sense *ws* of a synset *ss*, and to compute the Relatedness-A(*ns*, *ws*) = score(instance(*ns*), example(*ws*)).
2) To compare all nouns *ns* of SJND for a semantic class with all nouns in instances of STD linked to word senses *ws*, and to compute the Relatedness-B(*ns*, *ws*) = score(∀*ns*, nouns(example(*ws*))).
3) If Relatedness-A(*ns*, *ws*) ≥ λ_1 and Relatedness-B(*ns*, *ws*) ≥ λ_2, a synset is evaluated as a candidate synset *mc(A)*. If either Relatedness-A(*ns*, *ws*) ≥ λ_1 or Relatedness-B(*ns*, *ws*) ≥ λ_2, evaluated as a candidate synset *mc(B)*. When threshold λ_1 and λ_2 were 1~4, we had good performances.
4) To repeat from step 1) to 3) for all of synsets, in order to determine mapping candidate synsets. And then, to construct hierarchical structure for all those synsets.

4.2.3 Using Semantically Relatedness between Nouns of SJND and Synsets of KorLex

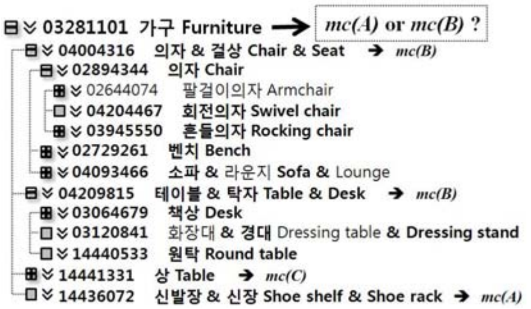
The third approach is to evaluate mapping candidate synsets using comparison of semantic relations and their semantically related words between a noun of SJND and word senses of a synset. To compute the relatedness between them, we reformulated the computational formula of relatedness based on the Lesk algorithm (Lesk 1987, Banerjee & al 2002). The process of evaluating mapping candidate synsets is as follows.

1) To compare semantically related words: between synonyms, hypernyms, hyponyms and antonyms of a noun of SJND and those of a synset of KorLex. To compute the Relatedness-C(*ns*, *ss*) = score (relations(*ns*), relations(*ss*)).
2) To compare all nouns *ns* of SJND for a semantic class with synonyms, hypernyms and hyponyms of a synset of KorLex, and compute the Relatedness-D(*ns*, *ss*) = score (∀*ns*, relations(*ss*)).
3) If Relatedness-C(*ns*, *ss*) ≥ λ_3 and Relatedness-D(*ns*, *ss*) ≥ λ_4, a synset is evaluated as a candidate synset *mc(A)*. If either Relatedness-C(*ns*, *ss*) ≥ λ_3 or Relatedness-D(*ns*, *ss*) ≥ λ_4, evaluated as a candidate synset *mc(B)*. When threshold λ_3 and λ_4 were 1~4, we have good performances.
4) To repeat from step 1) to 3) for all of synsets, in order to determine mapping synsets. And then, to construct hierarchical structure for all those synsets.

4.3 Determining Least Upper Bound (LUB) Synsets and Mapping Synsets

Next, we determine the LUB synsets using mapping candidate synsets and hierarchical structure having semantic relations: parent, child and sibling. In order to determine LUB and mapping synsets, we begin evaluation with bottom-up direction. Using relatedness among child-sibling candidate synsets, we evaluated whether their parent synset is a LUB synset or not. If the parent is a LUB synset, we evaluate its parent (grand-parent of the candidate) synset using relatedness among its sibling synsets. If the parent is not a LUB, the candidate synsets *mc(A)* or *mc(B)* are determined as mapping synsets (or LUB) and stop finding LUB. For all semantic classes, we determine LUB and mapping synsets. Finally, we link the LUB and mapping synsets to each semantic class of SJSC. The process of determining of LUB and mapping synsets is as follows.

1) Using candidate synsets and their sibling, for all candidate synsets *mc(A)*, *mc(B)* or *mc(C)* selected from the processes of "4.2 Select Mapping Candidate Synsets", to determine whether it is a LUB or not and final mapping synsets.
2) Among sibling synsets, if the ratio of count(*mc(A)*) to count(*mc(A)+mc(B)+ mc(C)*) is over 60%, the parent synset of siblings is evaluated as a candidate synset *mc(A)* and as a LUB.
3) If the ratio of count(*mc(A)+mc(B)*) to count(*mc(A)+mc(B)+mc(C)*) is over 70%, the parent of siblings is evaluated as a candidate synset *mc(A)* and as a LUB. If the ratio of count(*mc(A)+mc(B)*) to count (*mc(A) +mc(B)+mc(C)*) is between 50% and 69%, the parent of siblings is evaluated as a candidate synset *mc(B)* and as a LUB.
4) And if the others, to stop finding LUB for that synset and to determine final mapping synsets with its own level of candidate.
5) To repeat from step 1) to 4) until finding LUB synsets and final mapping synsets.



```
⊟ ≚ 03281101 가구 Furniture → mc(A) or mc(B) ?
  ⊟ ≚ 04004316 의자 & 걸상 Chair & Seat → mc(B)
    ⊟ ≚ 02894344 의자 Chair
      ⊞ ≚ 02644074 팔걸이의자 Armchair
      □ ≚ 04204467 회전의자 Swivel chair
      ⊞ ≚ 03945550 흔들의자 Rocking chair
    ⊞ ≚ 02729261 벤치 Bench
    ⊞ ≚ 04093466 소파 & 라운지 Sofa & Lounge
  ⊟ ≚ 04209815 테이블 & 탁자 Table & Desk → mc(B)
    ⊞ ≚ 03064679 책상 Desk
    □ ≚ 03120841 화장대 & 경대 Dressing table & Dressing stand
    □ ≚ 14440533 원탁 Round table
  ⊞ ≚ 14441331 상 Table → mc(C)
  □ ≚ 14436072 신발장 & 신장 Shoe shelf & Shoe rack → mc(A)
```

Figure 4. Hierarchical structure of mapping candidate synsets for a semantic class

Figure 4 shows hierarchical structure of mapping candidate synsets for a semantic class 'Furniture' and when candidate synsets' ID are '04004316' (Chair & Seat): *mc(B)*, '04209815' (Table & Desk): *mc(B)*, '14441331' (Table): *mc(C)*, and '14436072' (Shoe shelf & Shoe rack): *mc(A)*, we determine whether their parent synset '03281101' (Furniture) is a LUB or not, and evaluate it as a candidate synset *mc(A)* or *mc(B)*. In this case, synset '03281101' (Furniture) is a candidate *mc(A)* and a LUB synset.

For all semantic classes, we find their mapping LUB and mapping synsets using information of hierarchical structure and candidate synsets. Finally, we link each semantic class of SJSC to all lower level synsets of matched LUB synsets.

5 Experiments and Results

We experimented automatic mapping between 623 semantic classes of SJSC and 90,134 noun synsets of KorLex using the proposed automatic mapping method with three approaches. To evaluate the performances, we used the results of manual mapping as correct answers, that was mapped 474 semantic classes (terminal nodes) of SJSC to 65,820 synsets (73%) (include 6,487 LUB) among total 90,134 noun synsets of KorLex. We compared the results of automatic mapping with those of manual mapping. For evaluation of performances, we employed Recall, Precision and the F1 measure: F1 = (2*Recall*Precision)/(Recall+ Precision).

Approaches	Recall	Precision	F1
1)	0.904	0.502	0.645
2)	0.774	0.732	0.752
3)	0.670	0.802	0.730
1)+2)	**0.805**	**0.731**	**0.766**
1)+3)	**0.761**	**0.758**	**0.759**
2)+3)	0.636	0.823	0.718
1)+2)+3)	**0.838**	**0.718**	**0.774**

Table 2. Performances of automatic mapping with three approaches

Table 2 shows the performances of automatic mapping with three approaches: 1) IMPW, 2) INW, and 3) SRNS. The '1)', '2)' or '3) in the Table present the results using for each approach method and '1)+2)', '1)+3)' or '2)+3)' present those of combining two approaches. The '1)+2)+3)' presents those of the combining three approaches and we can see the best performances using the last approach among results: recall 0.837, precision 0.717 and F1 0.773. The first approach '1)' method shows high recall, but low precision and the third approach '3)' method present low recall and high precision. '1)+3)' and '2)+3)' shows good performances overall. Thus, we could see good performances using the combined approach methods.

Second, we compared the numbers of semantic classes, nouns entries of SJND, noun synsets and word senses of KorLex for each approach, after mapping processes.

As shown in Table 3, we can see the most numbers of mapping synsets using the '1)' approach. The '1)+2)+3)' shows the results similar to '1)', but has the best performances (see Table 2). The percentages in the round bracket present the ratio of the results of automatic mapping to original lexical data of *Sejong* and KorLex: 474 semantic classes of SJSC, 25,245 nouns of SJND and 90,134 noun synsets and 147,896 word senses in KorLex.

	SJD		**KorLex**	
Approaches	**SC (SJSC)**	**Nouns (SJND)**	**Synsets**	**Word Senses**
1)	**473**	**18,575**	**54,943**	**69,970**
2)	445	18,402	52,109	66,936
3)	413	18,047	49,768	64,003
1)+2)	463	18,521	52,563	67,109
1)+3)	457	18,460	51,786	66,157
2)+3)	383	17,651	48,398	62,063
1)+2)+3)	**466 (98.3%)**	**18,542 (72.8%)**	**54,083 (60%)**	**69,259 (46.8%)**

Table 3. Numbers of semantic class, noun of SJD, synset and word sense of KorLex

In manual mapping, we mapped 73% (65,820) synsets of KorLex for 474 semantic classes of SJSC. The 24,314 synsets was excluded in manual mapping among 90,134 total nouns synsets. The reasons of excluded synsets in manual mapping were 1) inconsistency of inheritance for lexical relations of parent-child in SJSC or KorLex, 2) inconsistency between criteria for SJSC and candidate synsets, 3) candidate synsets belonging to more than two semantic classes, 4) specific proper nouns (chemical compound names), and 5) polysemic abstract synsets (Bae & al. 2010).

In automatic mapping, we could map 60% (54,083) synsets among total nouns synsets (90,134) of KorLex, and it is 82.2% of the results of manual mapping. The 11,737 synsets was excluded in automatic mapping by comparing with manual mapping. Most of them were 1) tiny-grained synsets found in the lowest levels, 2) synsets having no matched word senses with those of SJND, 3) synsets with polysemic word senses, 4) word senses having poor instances in KorLex and in SJND, 5) word senses in SJND having poor semantic relations.

Level	LUB	Ratio	Level	LUB	Ratio
1	18	0.6%	9	230	7.3%
2	18	0.6%	10	98	3.1%
3	174	5.5%	11	32	1.0%
4	**452**	**14.3%**	12	20	0.6%
5	**616**	**19.5%**	13	4	0.1%
6	**570**	**18.0%**	14	4	0.1%
7	**486**	**15.4%**	15	2	0.1%
8	**442**	**14.0%**	16-17	0	0%

Table 4. Numbers and Ratio of LUB synsets excluded in automatic mapping

Table 4 shows the numbers and ratio of the LUB synsets excluded in automatic mapping for each level in depth. Most synsets are 4-8 levels synsets among 17 levels in depth.

6 Conclusions

We proposed a novel automatic mapping method with three approaches to link Sejong Semantic Classes and KorLex using 1) information of monosemy/polysemy of word senses, 2) instances of nouns of SJD and word senses of KorLex, 3) semantically related words of nouns of SJD and synsets of KorLex. To find common clues from lexical information among those language resources is important process in automatic mapping method. Our proposed automatic mapping method with three approaches shows notable performances by comparing with other studies on automatic mapping among language resources: recall 0.837, precision 0.717 and F1 0.773. Therefore, from those studies, we can improve Korean semantico-syntactic parsing technology by integrating the argument structures as provided by SJD, and the lexical-semantic hierarchy as provided by KorLex. In addition, we can enrich three resources: KorLex, SJD and STD as results of comparing and integrating them. We expect to improve automatic mapping technology among other Korean language resources through this study.

Acknowledgement

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government(MEST) (No. 2007-0054887).

References

Jan Scheffczyk, Adam Pease, Michael Ellsworth. 2006. *Linking FrameNet to the Suggested Upper Merged Ontology*. Proc of the 2006 conference on Formal Ontology in Information Systems (FOIS 2006): 289-300.

Ian Niles and Adam Pease. 2003. *Linking lexicons and ontologies: Mapping wordnet to the suggested upper merged ontology*. In Proceedings of the 2003 International Conference on Information and Knowledge Engineering (IKE 03).

KorLex, 2007. *Korean WordNet*, Korean Language processing Lab, Pusan National University. Available at http://korlex.cs.pusan.ac.kr

C. Hong. 2007. *The Research Report of Development 21th century Sejong Dictionary*, Ministry of Culture, Sports and Tourism, The National Institute of the Korean Language.

Dennis Spohr. 2008. *A General Methodology for Mapping EuroWordNets to the Suggested Upper Merged Ontology*, Proceedings of the 6th LREC 2008:1-5.

Alexander Budanitsky and Graeme Hirst. 2006. *Evaluating WordNet-based Measures of Lexical Semantic Relatedness*, Computational Linguistics,Vol 32: Issue 1:13- 47.

Siddharth Patwardhan, Satanjeev Banerjee and Ted Pedersen. 2003. *Using Measures of Semantic Relatedness for Word Sense Disambiguation*, CICLing 2003, LNCS(vol 2588):241-257.

Satanjeev Banerjee and Ted Pedersen. 2002. *An Adapted Lesk Algorithm for Word Sense Disambiguation Using WordNet*, Proceedings of CICLing 2002, LNCS 2276:136-145

Sara Tonelli and Daniele Pighin. 2009. *New Features for FrameNet -WordNet Mapping*, Proceedings of the 13th Conference on Computational Natural Language Learning: 219-227.

Aesun Yoon, Soonhee Hwang, E. Lee, Hyuk-Chul Kwon. 2009. *Consruction of Korean WordNet 'KorLex 1.5'*, JourNal of KIISE: Sortware and Applications, Vol 36: Issue 1:92-108.

Soonhee Hwang, A. Yoon, H. Kwon. 2010. *KorLex 1.5: A Lexical Sematic Network for Korean Numeral Classifiers*, JourNal of KIISE: Software and Applications, Vol 37: Issue 1:60-73.

Sun-Mee Bae, Kyoungup Im, Aesun Yoon. 2010. *Mapping Heterogeneous Ontologies for the HLT Applications: Sejong Semantic Classes and KorLexNoun 1.5*, Korean Journal of Cognitive Science. Vol. 21: Issue 1: 95-126.

Aesun Yoon. 2010. *Mapping Word Senses of Korean Predicates Between STD(STandard Dictionary) and SJD(SeJong Electronic Dictionary) for the HLT Applications*, Journal of the Linguistic Society of Korea. No 56: 197-235.

Hyopil Shin. 2010. *KOLON: Mapping Korean Words onto the Microkosmos Ontology and Combining Lexical Resources*. Journal of the Linguistic Society of Korea. No 56: 159-196.