Two approaches to generating Korean Numeral Classifiers

Francis Bond,♠ Kyonghee Paik,♡ Jong-Bok Kim,♢ Jaehyung Yang♣
♠ NICT Language Infrastructure Group, MASTAR Project
♡ Media Network Center, Waseda University
♢ Kyung Hee University, ♣ Kangnam University

1 Introduction

In this paper, we compare two approaches to generating Korean numeral classi ers, both using semantic classes from CoreNet. In the rst approach, classi ers are assigned to semantic classes by hand; in the second, the mapping is learned from a corpus. Both approaches achieve comparable performance.

2 Generation of Numeral Classifiers

First, let us briefly explain the properties of numeral classi ers, focusing on Korean; then give an algorithm to generate multilingual classi ers.

2.1 What are Numeral Classifiers

Numeral classi ers are used for languages which nouns cannot be directly modi ed by numerals. However, quanti er phrases can also function as noun phrases on their own, with anaphoric or deictic reference, when what is being quanti ed is recoverable from the context. For example (2) is acceptable if the letters have already been referred to, or are clearly visible. All the three languages share this property.

(1) [some background with books salient]

(2) 2-kwon-ul satta
2-CL-ACC buy
"I bought two books"

Numeral classi ers are a subclass of nouns. The main property distinguishing them from prototypical nouns is that they cannot be used by themselves. Typically, they post x to numerals, forming a quanti er phrase.

We will call all such combinations of a numeral/quanti er/interrogative with a numeral classi er a **numeral-classifier combination**, and the noun phrase they quantify their **target**. Languages differ as to what the classi ers can follow. Korean does not allow them to post x to any quanti ers, only to numerals or the interrogative [what]myech.

(3) *2-saram* "2 people" (Numeral)

(4) *myetch-saram* "some people" (Quanti er)

(5) *myetch-mari* "how many people" (Interrogative)

Numeral classi ers characteristically premodify their target, linked by an adnominal case marker, as in (6); or appear 'floating' as adverbial phrases, typically to before the verb: (7) that governs their target.

(6) 2-kwon-eui chaek-ul satta
2-CL-ADN book-ACC bought
"I bought two books."

(7) chaek-ul 2-kwon(-ul) satta
book-ACC 2-CL(-ACC) bought
"I bought two books."

Sortal classi ers differ from each other in the restrictions they place on their target. For example the classi er *-saram* adds the restriction that its target must be `human`. That is, it can only be used to classify human referents.

2.2 An Algorithm to Generate Numeral Classifiers

The basic algorithm we use is that of Bond and Paik (2000), an extension of the algorithm proposed by Sornlertlamvanich et al. (1994). The algorithm is shown in Figure 1.

The algorithm can be used when a noun is a member of more than one semantic class or of no semantic class. In the lexicon we used, nouns are, on average, members of 2 semantic classes. However, we assume that semantic classes are ordered so that the most basic class comes rst. During contextual processing, other semantic classes may become more salient, in which case they will be used to select the default classi er.

The algorithm can also handle the generation of classi ers that quantify coordinate noun phrases. These commonly appear in appositive noun phrases such as *ABC-to XYC-no 2-sha* “the two companies, ABC and XYZ”.

1. For a simple noun phrase
   (a) If the head noun has a default classi er in the lexicon:
   use the noun’s default classi er
   (b) Else if it exists, use the default classi er of the head noun’s most salient semantic class (the class’s default classi er)
   (c) Else use the **residual** classi er (개 -kae for Korean)
2. For a coordinate noun phrase
   generate the classi er for each noun phrase
   use the most frequent classi er

Figure 1: Algorithm to generate numeral classi- ers

If a noun’s default classi er is the same as the default classi er for its semantic class, then there is no need to list it in the lexicon. This makes the lexicon smaller and it is easier to add new entries. Any display of the lexical item (such as for maintenance or if the lexicon is used as a human aid), should automatically generate the classi er from the semantic class.

We extend step (1b) in one way: if a semantic class has no classi er associated with it, then we use the classi er associated with its hypernym. If the hypernym has no classi er we continue up the hierarchy until a classi er is found, or we reach the root. This allows us to mark the classi ers even more efficiently. We can mark the upper level node for `11111:human` with -myong for example, and only mark exceptions further down the tree.

3 The CoreNet Ontology

We used the ontology provided by CoreNet: A Korean-Japanese-Chinese Aligned Wordnet with Shared Semantic Hierarchy (Choi and Bae, 2003; Korterm, 2005). It is based on, and very similar to the Goi-Taikei — A Japanese Lexicon Ikehara et al. (1997). We choose it because of its rich ontology and its wide coverage of Korean, Japanese and Chinese.

The CoreNet consists of 2,937 conceptual nodes (semantic categories) with 12 depth levels and of 51,172 senses for nouns, 5,290 for verbs, and 2,081 for adjectives in Korean. The ontology has several hierarchies of concepts: with both `is-a` and `has-a` relationships. Words can be assigned to semantic classes anywhere in the hierarchy. Not all semantic classes have words assigned to them.

Each record in the dictionary has index form, part-of-speech, sense number and list of associated pronunciation, a canonical form, semantic classes. Each word can have up to ve common noun classes and ten proper noun classes. In the case of *usagi* “rabbit”, there are two common noun classes and no proper noun classes. The semantic classes are listed in order of salience (as judged by the dictionary compilers). Consider the entry for 배 *bae* which has several entries, differentiated by their semantic class. One is *bae* “ship” with the semantic class `11322912:ship` and another is *bae* “nashi pear” with the semantic class `11322531:fruit`.

4 Mapping Classifiers to the Ontology

In this section we discuss two methods to associate classi ers to semantic classes.

4.1 Rational: Introspective Method

The rst method is to associate classi ers with each of the CoreNet 2,937 semantic classes by hand. This takes around two weeks from scratch and was the same used by Paik and Bond (2001).

We show the most frequent numeral classi ers for for Korean in Table 1

4.2 Empirical: Corpus-based Method

We used a POS tagged corpus of newspaper reports (provided by KAIST). First we identi ed all sentences with numeral classi er combinations (NCL):

1. NCL = NUM+ CL POSTCL?
   where

   num is a number or interrogative word (POS `\nnc` or string 몇 *myech* “how many”)

CLASSIFIER		Referents classi ed	No.	%	Sample Semantic Class
None		Uncountable referents	799	29.5	`111:agent`
-kae	(개)	abstract/general objects	737	27.1	`11:concrete`
-hyoi	(회)	events	707	26.1	`122125141:visit`
-myong	(명)	people	296	10.9	`11111:person`
-bangul	(방울)	liquid	26	1.0	`113112722:tear`
-jang	(장)	flat objects	24	0.9	`1132221:paper`
-dae	(대)	mechanic items/ furniture	20	0.7	`113228:machinery`
-keun	(건)	incidents	14	0.5	`122125211:contract`
-mari	(마리)	animals	14	0.5	`537:beast`
Other		26 classi ers	73	2.7	

Table 1: Korean Numeral Classi ers and associated Semantic Classes

cl is a numeral classifer (marked with `\nbu`)

postcl is a bound morpheme that can follow a classi er. Currently we recognize the following

- 이상 *isan* "more than"
- 이하 *iha* "less than"
- 정도 *chongdo* "about"
- 가량 *karyang* "about"

this class is not distinguished by the tagger, they are all marked (`\ncn`)

We rejected any classi ers from a stop list of mensural classi ers, dates and currency units.

We then search for the target. If the NCL is followed by the adnominal marker 의 (`\jcm`) and the next word is noun, we take the following noun sequence, else we take the preceeding noun sequence. We take the nal noun in the noun sequence as the antecedent. An example is given in (8).

(8) 1만2000대 컴퓨터를
12000-dae komputer-lul
1-ten-thousand 2000-cl
리눅스로 교체한다
linuks-ro kyoche-handa
computer-acc linux-dat convert-do
Twelve thousand computers converted to linux.

From 314,806 sentences we were able to extract 45,937 Sortal/Event classi er tokens. There were 158 candidate classi ers: the most common was the residual classi er 개 *-kae* with 12,690 instances, then 명 *-myong* "human" with 7,730 instances and third 대 *-dae* "machine" with 5,480 instances. 27 classi ers had only one occurrence, more than half of them were tagging errors.

Up till here the approach is very similar to that of (Sornlertlamvanich et al., 1994). We then go beyond there to map the targets to semantic classes. As the corpus is not sense tagged, we looked up each target in CoreNet, and listed those we could nd by semantic class. We then hand checked that the semantic class was suitable for the classi er. This gave us a table of (classi er, target, semantic class) instances. For example, for (8), the mapping is (대, 컴퓨터, `11322823:computer`). We use these to make the empirical mapping. Where the same semantic class appears with multiple classi ers, they are ordered by token frequency, so that the most frequently occurring classi er will be generated. This mapping is richer than the rational mapping, as it has information about speci c words, but far less comprehensive.

5 Evaluation and Discussion

The algorithm was tested on the same set of 90 sentences used by Paik and Bond (2001). We only considered sentences overt targets classi ed by a sortal classi er. Noun phrases modi ed by group classi ers, such as *-soku* "pair" were not evaluated, as we reasoned that the presence of such a classi er would be marked in the input to the generator. We also did not consider the anaphoric use of numeral classi ers.

In total, there were 90 noun phrases modi ed by a sortal classi er. We assumed as input only the target word itself, and looked up its semantic class in CoreNet.

The results, with a breakdown of the errors, are summarised in Table 2. Correct stands for exact match, acceptable for a different register (for example, we generated 명 -myeong although 사람 -saram was used in the test set). Incorrect means we generated a different classi er than that in the test corpus. Overgenerated means we generated a classi er where one was used in Japanese but not in Korean.

The method based on intuition did better on this test set, mainly because of its wider cover. The corpus based method should do better on in domain data, we are currently constructing a new test set to con rm this. Another way of improving the corpus based method would be to generalize upward as well as downward, this is a topic for further work.

Result	Rational		Empirical	
	%	No.	%	No.
Correct	62%	56	54%	49
Acceptable	6%	5	3%	3
Incorrect	13%	12	23%	21
Over Generated	19%	17	19%	17
Total	100%	90	100%	90

Table 2: Results of applying the algorithm

Our classi er mapping is dwarfed by the detailed work of Hwang et al. (2008), who give a more precise mapping of classi er to semantic class using the Korean WordNet (KorLex). However, currently they do not have any frequency information, so have no way of selecting which of the possible classi ers should be used for a given noun. Guo and Zhong (2005) give a highly accurate method of selecting classi ers that uses many more context features. We agree that this gives better immediate results. However, our ultimate goal is to rst do word sense disambiguation using richer context features, and then use the appropriate semantic class with our algorithm. This should allow us to back-off for words not seen before with a classi er.

6 Conclusion

In this paper we presented an algorithm to generate Korean numeral classi ers using a rich. It was shown to select the correct sortal classi er 72% of the time using a hand mapping and only 52% of the time using a mapping based on a domain speci c corpus. The algorithm uses the ontology provided by CoreNet, and shows how accurately semantic classes can predict numeral classi ers for the nouns they subsume.

Acknowledgments

The authors thank KAIST and Key-Sun Choi for providing the corpus, and their overall support.

References

Francis Bond and Kyonghee Paik. Re-using an ontology to generate numeral classifiers. In *18th International Conference on Computational Linguistics: COLING-2000*, pages 90–96, Saarbrücken, 2000.

Key-Sun Choi and Hee-Sook Bae. A Korean-Japanese-Chinese aligned wordnet with shared semantic hierarchy. In *Proceedings of the International Conference on Natural Language Processing and Knowledge Engineering*, pages 91–96, 2003.

Hui Guo and Huayan Zhong. Chinese classifier assignment using SVMs. In *4th Sighan Workshop on Chinese Language Processing*, pages 25–31, Jeju Island, 2005.

Soonhee Hwang, Aesun Yoon, and Hyuk-Chul Kwon. Semantic representation of Korean numeral classifier and its ontology building for HLT applications. *Language Resources and Evaluation*, 42(2):151–172, 2008.

Satoru Ikehara, Masahiro Miyazaki, Satoshi Shirai, Akio Yokoo, Hiromi Nakaiwa, Kentaro Ogura, Yoshifumi Ooyama, and Yoshihiko Hayashi. *Goi-Taikei — A Japanese Lexicon*. Iwanami Shoten, Tokyo, 1997. 5 volumes/CDROM.

Korterm. *CoreNet: Multilingual WordNet*. KAIST Press, 2005. (in Korean).

Kyonghee Paik and Francis Bond. Multilingual generation of numeral classifiers using a common ontology. In *19th International Conference on Computer Processing of Oriental Languages: ICCPOL-2001*, Seoul, 2001. 141–147.

Virach Sornlertlamvanich, Wantanee Pantachat, and Surapant Meknavin. Classifier assignment by corpus-based approach. In *15th International Conference on Computational Linguistics: COLING-94*, pages 556–561, Kyoto, August 1994. (`http://xxx.lanl.gov/abs/cmp-lg/9411027`).