Generation of Chinese Numeral Classifiers using Semantic Classes

Kyonghee Paik
ATR Spoken Language Translation Research Lab
2-2-2 Hikari-dai, Kyoto 619-0237, JAPAN
kyonghee.paik@atr.co.jp

Francis Bond
NTT Communication Science Labs
2-4 Hikari-dai, Kyoto 619-0237, JAPAN
bond@cslab.kecl.ntt.co.jp

Abstract

In this paper, we present a solution to the problem of generating Chinese numeral classifiers using semantic classes from an ontology. In order to select an appropriate classifier, we propose an algorithm which associates classifiers with semantic classes and uses inheritance to list only exceptional classifiers with individual nouns. In this paper, we show that our proposed algorithm is effective not only for Korean and Japanese but for the genetically unrelated language Chinese.

1 Introduction

In this paper we extend a multi-lingual approach to generate numeral classifiers from Japanese and Korean to the unrelated language Chinese. We have already shown that using the Goi-Taikei ontology is a effective way to generate classifiers of Japanese and Korean (Bond and Paik, 2000; Paik and Bond, 2001). One immediate application for the generation of classifiers is machine translation, and we shall take examples from there, but it is needed for the generation of any quantified noun phrase with an uncountable head noun.

2 An Algorithm to Generate Numeral Classifiers

We use the algorithm given in Paik and Bond (2001), 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, the semantic classes are ordered so that the most basic class comes first (Ikehara et al., 1997, vol 1, p25). For example, *usagi* "rabbit" is marked as both `animal` and `meat`, with `animal` coming first. During contextual processing, other semantic classes may become more salient, in which case they will be used to select the appropriate semantic classifier.

If a noun's default classifier is the same as the default classifier for its semantic class, then there is no need to list it in the lexicon. This makes the lexicon smaller. Further, 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 classifier from the semantic class. Alternatively (and equivalently), in a lexicon with multiple inheritance and defaults, the class's default classifier can be added as a defeasible constraint on all members of the semantic class.

We use semantic classes from the ontology provided by Goi-Taikei — A Japanese Lexicon (Ikehara et al., 1997). We choose it because of its rich ontology, its extensive use in many other NLP applications, its wide coverage of Japanese, and the fact that it is being extended to other numeral clas-

1. For a simple noun phrase
   (a) If the head noun has a default classifier in the lexicon:
   use the noun's default classifier
   (b) Else if it exists, use the default classifier of the head noun's most salient semantic class (the class's default classifier)
   (c) Else use the **residual** classifier
   (つ *-tsu* for Japanese; 개 *-kae* for Korean; 个 *-ge* for Chinese)
2. For a coordinate noun phrase
   generate the classifier for each noun phrase
   use the most frequent classifier

Figure 1: Algorithm to generate numeral classifiers

sifier languages, such as Malay and Chinese.

The ontology has several hierarchies of concepts: with both `is-a` and `has-a` relationships. There are 2,710 semantic classes in a 12-level tree structure for common nouns. Words can be assigned to semantic classes anywhere in the hierarchy. Not all semantic classes have words assigned to them.

3 Classifiers and the Ontology

In this section we investigate how far the semantic classes can be used to predict default classifiers for Chinese using Japanese semantic classes. Because most sortal classifiers select for some kind of semantic class, nouns grouped together under the same semantic class will typically share the same classifier.

We show the most frequent numeral classifiers for Japanese in Table 1, for Korean in Table 2, followed by Table 3 for Chinese. We ended up with 47 classifiers used as semantic classes' default classifiers for Japanese. This is in line with the fact that most speakers of Japanese know and use between 30 and 80 sortal classifiers (Downing, 1996). Note that, we expect to add more specific classifiers at the noun level. For Chinese, there a total of 82 classifiers, more than for Korean and Japanese but still within the range predicted by Downing.

As we can see from the Table 1, 794 semantic classes were not assigned classifiers in Japanese. This included classes with no words associated with them, and those that only contained nouns with referents so abstract we considered them to be uncountable, such as freedom, greed, and so on. For Korean and Chinese, 799 and 765 classes were assigned no classifier.

The mapping we created is not complete because some of the semantic classes have nouns which do not share the same classifiers. In order to generate classifiers accurately, it is necessary to add more specific defaults at the noun level (noun default classifiers). As well as more specific sortal classifiers, there are cases where a group classifier may be more appropriate. For example, among the nouns counted with 人 *-nin* in Japanese there are entries such as couple, twins and so on which are often counted with 組 *-kumi* "pair".

In addition, the choice of classifier can depend on factors other than just semantic class, for example, in Chinese, unlike Japanese and Korean, the residual classifier (个 *-ge4*) can also be used to count people. There are three classifiers specific to counting people: 位 *-wei4* and 名 *-ming2* are politer terms. 人 *-ren2* is limited to appositive uses, it is only used when the numeral classifier comes after the target noun phrase, as in (1). It is also the classifier used when the target is a pronoun.

(1) 学生 3-人
student 3-CL
"3 students"

The most frequent numeral classifiers for Korean are shown in Table 2. Even though there are similarities between Japanese and Korean, we find

CLASSIFIER		Referents classified	No.	%	Sample Semantic Class
None		Uncountable referents	794	29.3	3:agent
-kai	(回)	events	703	25.9	1699:visit
-tsu	(つ)	abstract/general objects	565	20.9	2:concrete
-nin	(人)	people	298	11.0	5:person
-ko	(個)	concrete objects	124	4.6	854:edible fruit
-hon	(本)	long thin objects	52	1.9	673:tree
-mai	(枚)	flat objects	32	1.2	770:paper
-teki	(滴)	liquid	21	0.8	652:tear
-dai	(台)	mechanic items/ furniture	18	0.7	962:machinery
-hiki	(匹)	animals	12	0.6	537:beast
Other		38 classifiers	91	3.4	

Table 1: Japanese Numeral Classifiers and associated Semantic Classes

CLASSIFIER		Referents classified	No.	%	Sample Semantic Class
None		Uncountable referents	799	29.5	3:agent
-kae	(개)	abstract/general objects	737	27.1	2:concrete
-hyoi	(회)	events	707	26.1	1699:visit
-myong	(명)	people	296	10.9	5:person
-bangul	(방울)	liquid	26	1.0	652:tear
-jang	(장)	flat objects	24	0.9	770:paper
-dae	(대)	mechanic items/ furniture	20	0.7	962:machinery
-keun	(건)	incidents	14	0.5	1717:contract
-mari	(마리)	animals	14	0.5	537:beast
Other		26 classifiers	73	2.7	

Table 2: Korean Numeral Classifiers and associated Semantic Classes

CLASSIFIER		Referents classified	No.	%	Sample Semantic Class
None		Uncountable referents	765	28.2	3:agent
-ci4	(次)	events	692	25.5	1699:visit
-ge4	(个)	general object/people	655	24.1	2:concrete
-wei4	(位)	people (honored)	68	2.5	228:doctor
-quai4	(块)	big objects	61	2.2	461:land
-ren2	(人)	people	39	1.4	92:descendants
-tiao2	(条)	long thin objects	33	1.2	417:traffic route
-pian4	(片)	parts/pieces	25	0.9	2578:flake
-zhang1	(张)	big flat objects	23	0.8	773:board
-ming2	(名)	people (respected)	22	0.8	351:expert
-di1	(滴)	liquid	20	0.7	652:tear
-jian4	(件)	incidents	19	0.7	1717:contract
Other		70 classifiers	293	10.8	

Table 3: Chinese Numeral Classifiers and associated Semantic Classes

some difference in both ranking and the kinds of numeral classifiers. First of all, as we can see, the most frequent classifier is *-kae*. This is because Korean has only one residual classifier, unlike Japanese which has *-tsu* and *-ko*.

Also, many nouns are counted with the Japanese shape classifiers 本 *-hon* "long-thin object" and 枚 *-mai* "flat object", whereas the residual classifier 개 *-kae* is frequently used in Korean. Chinese classifiers also specify much detail with respect to shape and size. For example, 块 *-quai4* and 片 *-pian4* can be used for chunks of object, but 块 is used for far bigger objects than 片. 张 *-zhang1* and 枚 *-mei2* are used for flat objects while 滴 *-di1* and 条 *-tiao3* are used for tear shape objects and long thin objects. This partly explains why Chinese has so many 'other' type classifiers: 70 as opposed to 38 (Japanese other type classifiers) and 26 (Korean other type classifiers).

Overall, Chinese has far more different classifiers. As we can see from the Table 3, the classifier for the general object 个 *-ge* accounts for 655 semantic classes which is far fewer than Japanese and Korean. In particular, after excuding the cases (185) of counting people from the 655 semantic classes, 个 *-ge*, counting general/concrete objects, accounts for only 470 semantic classes. The reason for this is that Chinese has more specific classifiers. For example, Chinese has many different classifiers for different types of animals. Therefore horses are counted with 匹, snakes with 条, bugs with 只 and pigs with 头, where they would all typically be counted with the same classifier in Japanese and Korean.

Further, Japanese does not use the classifer 张 *-zhang1* for flat objects, whereas Chinese and Korean do. The classifier 枚 *-mei2* is used for flat objects in Japanese, without considering what the objects are made of and how big it is. Both of the classifiers can count flat objects in Chinese and Korean, but 张 *zhang1* is used for relatively big objects and its coverage is wider than 枚 *-mei2*, for which the usage is limited to small and very thin objects. One interesting area for further analysis would be to investigate the historical backgrounds behind the variation among the three languages.

4 Conclusion

In this paper we presented an algorithm to generate Japanese, Korean and Chinese numeral classifiers using a common ontology. Mapping the classifiers to the ontology show interesting differences in usage among the three languages.

For the further work, we plan to evaluate the accuracy of generation for Chinese. In addition, we want to investigate more syntactic and semantic features related to the usage of classifiers. This will leads to a greater understanding of language-specific characteristics and consequently to improve the quality of NLP processing for classifier languages.

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.

Pamela Downing. *Numeral Classifier Systems, the case of Japanese*. John Benjamins, Amsterdam, 1996.

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.

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`).