Constructing Taxonomy of Numerative Classifiers for Asian Languages

Kiyoaki Shirai
JAIST
kshirai@jaist.ac.jp

Takenobu Tokunaga
Tokyo Inst. of Tech.
take@cl.cs.titech.ac.jp

Chu-Ren Huang
Academia Sinica
churenhuang@gmail.com

Shu-Kai Hsieh
National Taiwan Normal Univ.
shukai@gmail.com

Tzu-Yi Kuo
Academia Sinica
ivykuo@gate.sinica.edu.tw

Virach Sornlertlamvanich
TCL, NICT
virach@tcllab.org

Thatsanee Charoenporn
TCL, NICT
thatsanee@tcllab.org

Abstract

Numerative classifiers are ubiquitous in many Asian languages. This paper proposes a method to construct a taxonomy of numerative classifiers based on a noun-classifier agreement database. The taxonomy defines superordinate-subordinate relation among numerative classifiers and represents the relations in tree structures. The experiments to construct taxonomies were conducted for evaluation by using data from three different languages: Chinese, Japanese and Thai. We found that our method was promising for Chinese and Japanese, but inappropriate for Thai. It confirms that there really is no hierarchy among Thai classifiers.

1 Introduction

Many Asian languages do not mark grammatical numbers (singular/plural) in noun form, but use numerative classifiers together with numerals instead when describing the number of nouns. Numerative classifiers (hereafter "classifiers") are used with a limited group of nouns, in particular material nouns. In English, for example: "three pieces of paper". In Asian languages these classifiers are ubiquitous and used with common nouns. Therefore the number of classifiers is much larger than in Western languages. An agreement between nouns and classifiers is also necessary, i.e., a certain noun specifies possible classifiers. The agreement is determined based on various aspects of a noun, such as its meaning, shape, pragmatic aspect and so on.

This paper proposes a method to automatically construct a taxonomy of numerative classifiers for Asian languages. The taxonomy defines superordinate-subordinate relations between classifiers. For instance, the Japanese classifier "頭 (*tô*)" is used for counting big animals such as elephants and tigers, while "匹 (*hiki*)" is used for all animals. Since "匹" can be considered more general than "頭", "匹" is the superordinate classifier of "頭", represented as "匹" $\succ$ "頭" in this paper. The taxonomy represents such superordinate-subordinate relations between classifiers in the form of a tree structure. A taxonomy of classifiers would be fundamental knowledge for natural language processing. In addition, it will be useful for language learners, because learning usage of classifiers is rather difficult, especially for Western language speakers.

We evaluate the proposed method by using the data of three Asian languages: Chinese, Japanese and Thai.

2 Noun-classifier agreement database

First, let us introduce usages of classifiers in Asian languages. In the following examples, "CL" stands for classifier.

- Chinese: *yi-ju* (CL) *dian-hua* (telephone) $\cdots$ a telephone
- Japanese: *inu* (dog) *2* *hiki* (CL) $\cdots$ 2 dogs
- Thai: *nakrian* (student) *3* *khon* (CL) $\cdots$ 3 students

As mentioned earlier, the agreement between nouns and classifiers is observed. For instance, the Japanese classifier "*hiki*" in the above example agrees with only animals. The agreement is also found in Chinese and Thai.

The proposed method to construct a classifier taxonomy is based on agreement between nouns and classifiers. First we prepare a collection of pairs (n, c) of a noun n and a classifier c which agrees with n for a language. The statistics of our Chinese, Japanese, and Thai database are summarized in Table 1.

Table 1: Noun-classifier agreement database

	Chinese	Japanese	Thai
No. of (n,c) pairs	28,202	9,582	9,618
No. of nouns (type)	10,250	4,624	8,224
No. of CLs (type)	205	331	608

The Japanese database was built by extracting noun-classifier pairs from a dictionary (Iida, 2004) which enumerates nouns and their corresponding classifiers. The Chinese database was derived from a dictionary (Huang et al., 1997). The Thai database consists of a mixture of two kinds of noun-classifier pairs: 8,024 nouns and their corresponding classifiers from a dictionary of a machine translation system (CICC, 1995) and 200 from a corpus. The pairs from the corpus were manually checked for their validity.

3 Proposed Method

3.1 Extracting superordinate-subordinate relations of classifiers

We extracted superordinate-subordinate classifier pairs based on inclusive relations of sets of nouns agreeing with those classifiers. Suppose that N_k is a set of nouns that agrees with a classifier c_k. If N_i subsumes N_j ($N_i \supset N_j$), we can estimate that c_i subsumes c_j ($c_i \succ c_j$). For instance, in our Japanese database, the classifier "店 (*ten*)" agrees with shops such as "drug store", "kiosk" and "restaurant", and these nouns also agree with "軒 (*ken*)", since "軒" is a classifier which agrees with any kind of building. Thus, we can estimate the relation "軒" $\succ$ "店".

Given a certain classifier c_j, c_i satisfying the following two conditions (1) and (2) is considered as a superordinate classifier of c_j.

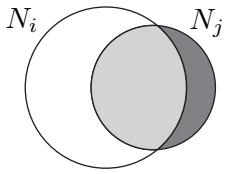

Figure 1: Relation of sets of nouns agreeing with classifiers

$$|N_i| > |N_j| \quad (1)$$

$$\mathrm{IR}(c_i, c_j) \geq T_{ir} \quad (2)$$

$$\text{where } \mathrm{IR}(c_i, c_j) \stackrel{def}{=} \frac{|N_i \cap N_j|}{|N_j|}$$

Condition (1) requires that a superordinate classifier agrees with more nouns than a subordinate classifier. $\mathrm{IR}(c_i, c_j)$ is an inclusion ratio representing to what extent nouns in N_j are also included in N_i (the ratio of the light gray area to the area of the small circle in Figure 1).

Condition (2) means that if $\mathrm{IR}(c_i, c_j)$ is greater than a certain threshold T_{ir}, we estimate a superordinate-subordinate relation between c_i and c_j. The basic idea is that superordinate-subordinate relations are extracted when N_j is a proper subset of N_i, i.e. $\mathrm{IR}(c_i, c_j) = 1$, but this is too strict. In order to extract more relations, we loosen this condition such that relations are extracted when $\mathrm{IR}(c_i, c_j)$ is large enough. If we set T_{ir} lower, more relations can be acquired, but they may be less reliable.

Table 2: Extraction of superordinate-subordinate relations

	Chinese	Japanese	Thai
T_{ir}	0.7	0.6	0.6
No. of extracted relations	251	322	239
No. of CLs not in the extracted relations	36 (18%)	76 (23%)	395 (61%)

Table 2 shows the results of our experiments to extract superordinate-subordinate relations of classifiers. The threshold T_{ir} was determined in an *ad hoc* manner for each language. The numbers of extracted superordinate-subordinate relations are shown in the second row in the table. Manual inspection of the sampled relations revealed that many reasonable relations were extracted. The objective evaluation of these extracted relations will be discussed in 4.2.

The third row in Table 2 indicates the numbers of classifiers which were not included in the extracted superordinate-subordinate relations with its ratio to the total number of classifiers in the database in parentheses. We found that no relation is extracted for a large number of Thai classifiers.

3.2 Constructing structure

The structure of a taxonomy is constructed based on a set of superordinate-subordinate relations between classifiers. Currently we adopt a very naive approach to construct structures, i.e., starting from the most superordinate classifiers as roots, we extend trees downward to less general classifiers by using the extracted superordinate-subordinate relations. Note that since there is more than one classifier that does not have any superordinate classifiers, we will have a set of trees rather than a single tree.

When constructing structures, redundant relations are ignored in order to make the structures as concise as possible. A relation is considered redundant if the relation can be inferred by using other relations and transitivity of the relations. The formal definition of redundant relations is given below:

$$c_a \succ c_b \text{ is redundant } \textit{iff } \exists c_m : c_a \succ c_m, c_m \succ c_b$$

Statistics of constructed structures for each language are shown in Table 3. More than 50 isolated structures (trees) were obtained for Chinese and Japanese, while more than 100 for Thai. We obtained several large structures, the largest containing 45, 85 and 23 classifiers for Chinese, Japanese and Thai, respectively. As indicated in the fifth row in Table 3, however, many structures consisting of only 2 classifiers were also constructed.

Table 3: Construction of structures

	Chinese	Japanese	Thai
No. of structures	52	54	102
No. of CLs in a structure			
Average	4.9	6.3	3.3
Maximum	45	85	23
Max. depth of structures	4	3	3
No. of structures with 2 CLs	18	24	54

4 Discussion

In this section, we will discuss the results of our experiments. First 4.1 discusses appropriateness of our method for the three languages. Then we evaluate our method in more detail. The evaluation of extracted superordinate-subordinate relations is described in 4.2, and the evaluation of structures in 4.3.

4.1 Comparison of different languages

According to the results of our experiments, the proposed method seems promising for Chinese and Japanese, but not for Thai. From the Thai data, no relation was obtained for about 60% of classifiers (Table 2), and many small fragmented structures were created (Table 3).

This is because of the characteristic that nouns and classifiers are strongly coupled in Thai, i.e., many classifiers agree with only one noun. In our Thai database, 252 (41.5%) classifiers agree with only one noun. This means that the overlap between two noun sets N_i and N_j can be quite small, making the inclusion ratio $\mathrm{IR}(c_i, c_j)$ very small. Out basic idea is that we can extract superordinate-subordinate relations between two classifiers when the overlap of their corresponding noun sets is large. However, this assumption does not hold in Thai classifiers. The above facts suggest that there seems to be no hierarchical taxonomy of classifiers in Thai.

4.2 Evaluation of extracted relations

4.2.1 Analysis of Nouns in $N_j \setminus N_i$

As explained in 3.1, our method extracts a relation $c_i \succ c_j$ even when N_i does not completely subsume N_j. We analysed nouns in the relative complement of N_i in N_j ($N_j \setminus N_i$), i.e., the dark gray area in Figure 1. The relation $c_i \succ c_j$ implies that all nouns which are countable with a subordinate classifier c_j are also countable with its superordinate classifier c_i, but there is no guarantee of this for nouns in $N_j \setminus N_i$, since we loosened the condition as in (2) by introducing a threshold.

To see to what extent nouns in $N_j \setminus N_i$ agree with c_i as well, we manually verified the agreement of nouns in $N_j \setminus N_i$ and c_i for all extracted relations $c_i \succ c_j$. The verification was done by native speakers of each language. Results of the validation are summarized in Table 4. For Japanese and Chinese, multiple judges verified the results. When judgments conflicted, we decided the final decision by a discussion of two judges for Japanese, and by majority voting for Chinese. The 4th and 5th rows

in Table 4 show the agreement of judgments. The "Agreement ratio" is the ratio of cases that judgments agree. Since three judges verified nouns for Chinese, we show the average of the agreement ratios for two judges out of the three. The agreement ratio and Cohen's κ is relatively high for Japanese, but not for Chinese. We found many uncertain cases for Chinese nouns. For example, "位 (*wei*)" is a classifier used when counting people with honorific perspective. However, judgement if "位" can modify nouns such as "political prisoner" or "local villain" is rather uncertain.

Table 4: Analysis of nouns in $N_j \setminus N_i$

	Chinese	Japanese	Thai
No. of nouns in $N_j \setminus N_i$	1,650	579	43
No. of nouns countable	1,195	241	24
with c_i as well	72%	42%	56%
No. of judges	3	2	1
Agreement ratio	0.677	0.936	–
Cohen's κ	0.484	0.868	–

Table 4 reveals that a considerable number of nouns in $N_j \setminus N_i$ are actually countable with c_i, meaning that our databases do not include noun-classifier agreement exhaustively.

4.2.2 Reliability of relations "$\succ$"

Based on the analysis in 4.2.1, we evaluate extracted superordinate-subordinate relations. We define the reliability R of the relation $c_i \succ c_j$ as

$$\mathrm{R}(c_i \succ c_j) \quad = \quad \frac{|N_i \cap N_j| + |NC_{j,i}|}{|N_j|}, \qquad (3)$$

where, $NC_{j,i}$ is a subset of $N_j \setminus N_i$ consisting of nouns which are manually judged to agree with c_i. We can consider that the more strictly this statement holds, the more reliable the extracted relations will be.

Figure 2 shows the relations between the threshold T_{ir} and both the number of extracted relations and their reliability. The horizontal axis indicates the threshold T_{ir} in (2). The bar charts indicate the number of extracted relations, while the line graphs indicate the averages of reliability of all extracted relations. Of course, if we set T_{ir} lower, we can extract more relations at the cost of their reliability. However, even when T_{ir} is set to the lowest value, the averages of reliability are relatively high, i.e. 0.98 (Chinese), 0.91 (Japanese) and 0.99 (Thai). Thus we can conclude that the extracted superordinate-subordinate relations are reliable enough.

4.3 Evaluation of structures

As in ordinary ontologies, we will assume that properties of superordinate classifiers can be inherited to their subordinate classifiers. In other words, a classifier taxonomy suggests transitivity of agreement with nouns over superordinate-subordinate relations as

$$c_1 \succ c_2 \ \wedge \ c_2 \succ c_3 \quad \Rightarrow \quad c_1 \succ c_3.$$

In order to evaluate the structures of our taxonomy, we verify the validity of transitivity.

First, we extracted all pairs of classifiers having an ancestor-descendant relation from our classifier taxonomy. Hereafter we denote ancestor-descendant pairs of classifiers as (c_a, c_d), where c_a is an ancestor and c_d an descendant. The path from c_a to c_d on the taxonomy can be represented as

$$c_0(= c_a) \succ c_1 \succ ... \succ c_n(= c_d). \qquad (4)$$

We denote a superordinate-subordinate relation derived by transitivity as $\overset{*}{\succ}$, such as $c_0 \overset{*}{\succ} c_n$. Among all ancestor-descendant relations, we extracted ones with a path length of more than one, or $n > 1$ in (4). Then we compare $\mathrm{R}(c_a \overset{*}{\succ} c_d)$, the reliability of a relation derived by transitivity, with $\mathrm{R}(c_i \succ c_{i+1})$ $(0 \leq i < n)$, the reliability of direct relations in the path from c_a to c_d. If these are comparable, we can conclude that transitivity in the taxonomy is valid.

Table 5 shows the results of the analysis of transitivity. As indicated in the column "all" in Table 5, 78 and 86 ancestor-descendant pairs (c_a, c_d) were extracted from the Chinese and Japanese classifier taxonomy, respectively. In contrast, only 6 pairs were extracted from the Thai taxonomy, since each structure of the Thai taxonomy is rather small as we already discussed with Table 3. Thus we have omitted further analysis of Thai. The extracted ancestor-descendant pairs of classifiers are then classified into three cases, (A), (B) and (C). Their numbers are shown in the last three rows in Table 5, where $\min_i$ and $\max_i$ denote the minimum and maximum of reliability among all direct relations $\mathrm{R}(c_i \succ c_{i+1})$ in the path from c_a to c_d.

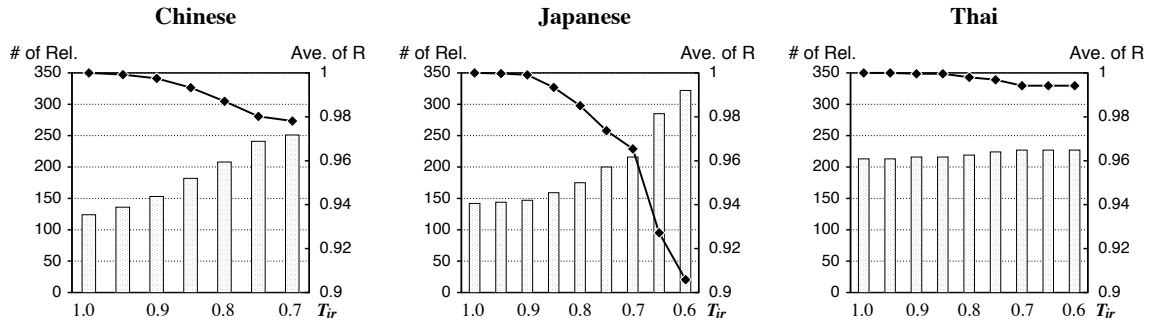

Figure 2: Reliability of extracted superordinate-subordinate relations

Table 5: Verification of transitivity

	Chinese			Japanese		
	all	direct	indirect	all	direct	indirect
No. of (c_a, c_d)	78	58	20	86	55	31
Average of $\mathrm{R}(c_a \overset{*}{\succ} c_d)$	0.88	0.98	0.61	0.77	0.93	0.48
(A) $\min_i > \mathrm{R}(c_a \overset{*}{\succ} c_d)$	16 (21%)	4 (7%)	12 (60%)	24 (28%)	3 (5%)	21 (68%)
(B) $\min_i \leq \mathrm{R}(c_a \overset{*}{\succ} c_d) < \max_i$	39 (50%)	34 (59%)	5 (25%)	27 (31%)	24 (44%)	3 (9%)
(C) $\max_i \leq \mathrm{R}(c_a \overset{*}{\succ} c_d)$	23 (29%)	20 (34%)	3 (15%)	35 (41%)	28 (51%)	7 (23%)

In case (A), reliability of a relation derived by transitivity, $\mathrm{R}(c_a \overset{*}{\succ} c_d)$, is less than that of any direct relations, $\mathrm{R}(c_i \succ c_{i+1})$. In case (B), reliability of a transitive relation is comparable with that of direct relations, i.e. $\mathrm{R}(c_a \overset{*}{\succ} c_d)$ is greater or equal to min_i and less than max_i. In case (C), the transitive relation is more reliable than direct relations.

The average of the reliability of $c_a \overset{*}{\succ} c_d$ is relatively high, 0.88 for Chinese and 0.77 for Japanese. We also found that more than 70% of derived relations (case (B) and case (C)) are comparable to or greater than direct relations. The above facts indicate transitivity on our structural taxonomy is valid to some degree.

From a different point of view, we divided pairs of (c_a, c_d) into two other cases, "direct" and "indirect" as shown in the columns of Table 5. The "direct" case includes the relations which are also extracted by our method. Note that such relations are discarded as redundant ones. On the other hand, the "indirect" case includes the relations which can not be extracted from the database but only inferred by using transitivity on the taxonomy. That is, they are truly new relations. In order to calculate reliability of "indirect" cases, we performed additional manual validation of nouns in $N_d \backslash N_a$.

However, the average of $\mathrm{R}(c_a \overset{*}{\succ} c_d)$ in "indirect" cases is not so high for both Chinese and Japanese, as a large amount of pairs are classified into case (A). Thus it is not effective to infer new superordinate-subordinate relations by transitivity. Since we currently only adopted a very naive method to construct a classifier taxonomy, more sophisticated methods should be explored in order to prevent inferring irrelevant relations.

5 Related Work

Bond (2000) proposed a method to choose an appropriate classifier for a noun by referring its semantic class. This method is implemented in a sentence generation module of a machine translation system. Similar attempts to generate both Japanese and Korean classifiers were also reported (Paik and Bond, 2001). Bender and Siegel (2004) implemented a HPSG that handles several intricate structures including Japanese classifiers. Matsumoto (1993) reported his close analysis of Japanese classifiers based on prototype semantics. Sornlertlamvanich (1994) presented an algorithm for selecting an adequate classifier for a noun by using a corpus. Their research can be regarded as a method to construct a noun-classifier agreement database au-

tomatically from corpora. We used databases derived from dictionaries except for a small number of noun-classifier pairs in Thai, because we believe dictionaries provide more reliable and stable information than corpora, and in addition they were available and on hand. Note that we are not concerned with frequencies of noun-classifier coocurrence in this study. Huang (1998) proposed a method to construct a noun taxonomy based on noun-classifier agreement that is very similar to ours, but aims at developing a taxonomy for nouns rather than one for classifiers. There has not been very much work on building resources concerning noun-classifier agreement. To our knowledge, this is the first attempt to construct a classifier taxonomy.

6 Conclusion

This paper proposed a method to construct a taxonomy of numerative classifiers based on a noun-classifier agreement database. First, superordinate-subordinate relations of two classifiers are extracted by measuring the overlap of two sets of nouns agreeing with each classifier. Then these relations are used as building blocks to build a taxonomy of tree structures. We conducted experiments to build classifier taxonomies for three languages: Chinese, Japanese and Thai. The effectiveness of our method was evaluated by measuring reliability of extracted relations, and verifying validity of transitivity in the taxonomy. We found that extracted relations are reliable, and the transitivity in the taxonomy relatively valid. Relations inferred by transitivity, however, are less reliable than those directly derived from noun-classifier agreement.

Future work includes investigating a way to enlarge classifier taxonomies. Currently, not all classifiers are included in our taxonomy, and it consists of a set of fragmented structures. A more sophisticated method to build a large taxonomy including more classifiers should be examined. Our method should also be refined in order to make superordinate-subordinate relations inferred by the transitivity more reliable. We are now investigating a stepwise method to construct taxonomies that prefers more reliable relations, i.e. an initial taxonomy is built with a small number of highly reliable relations, and is then expanded with less reliable ones.

Acknowledgment

This research was carried out through financial support provided under the NEDO International Joint Research Grant Program (NEDO Grant).

References

Emily M. Bender and Melanie Siegel. 2004. Implementing the syntax of Japanese numeral classifiers. In *Proceedings of the the First International Joint Conference on Natural Language Processing*, pages 398–405.

Francis Bond and Kyonghee Paik. 2000. Reusing an ontology to generate numeral classifiers. In *Proceedings of the COLING*, pages 90–96.

CICC. 1995. CICC Thai basic dictionary. (developed by Center of the International Cooperation for Computerization).

Chu-Ren Huang, Keh-Jian Chen, and Chin-Hsiung Lai, editors. 1997. *Mandarin Daily News Dictionary of Measure Words*. Mandarin Daily News Publisher.

Chu-Ren Huang, Keh-jiann Chen, and Zhao-ming Gao. 1998. Noun class extraction from a corpus-based collocation dictionary: An integration of computational and qualitative approaches. In *Quantitative and Computational Studies of Chinese Linguistics*, pages 339–352.

Asako Iida. 2004. *Kazoekata no Ziten (Dictionary for counting things)*. Shôgakukan. (in Japanese).

Yo Matsumoto. 1993. The Japanese numeral classifiers: A study of semantic categories and lexical organization. *Linguistics*, 31:667–713.

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

Virach Sornlertlamvanich, Wantanee Pantachat, and Surapant Meknavin. 1994. Classifier assignment by corpus-based approach. In *Proceedings of the COLING*, pages 556–561.