Using WordNet to predict numeral classifiers in Chinese and Japanese

Hazel Mok Shu Wen, Gao Huini Eshley and Francis Bond
Linguistics and Multilingual Studies
Nanyang Technological University
haze0004@e.ntu.edu.sg, gaoh0004@e.ntu.edu.sg, bond@ieee.org

Abstract

Most nouns must be modified by a numeral-classifier combination when quantified in classifier languages like Chinese and Japanese. In this paper, we present a method to generate numeral classifiers using Chinese and Japanese WordNets. We assign synsets from WordNet to each classifier by hand and use a modified algorithm to generate sortal classifiers based on semantic hierarchies. We obtained a generation score of 78.80% for Chinese and 89.84% for Japanese.

1 Introduction

Classifiers have long been an interest for many linguists. In many Asian languages like Japanese and Chinese, nouns often require numeral classifiers when they are quantified (Bond & Paik, 2000; Downing, 1996). This contrasts with English where count nouns can be modified by a numeral. In English, it is acceptable to say “two books”. In classifier languages, it is obligatory to use a numeral-classifier pairing, as in (1) and (2).

(1) Japanese:2 冊 の 本
2-satsu-no *hon*
2-CL-ADN book
“2 books”

(2) Chinese: 两 本 书
liǎng *běn* *shū*
2 CL book
“2 books”

Numeral classifier languages typically lack plural markings. Greenberg (1972) has suggested that a classifier in these languages functions like a plural suffix in languages that require plural markings (cited in Downing, 1996). Hundius and Kölver (1983) argue that a classifier establishes immediate reference to individual objects.

The type of classifier used depends on the semantic features of the noun referent (Zhang, 2007). Some of these semantic categories include animacy and shape. For example, 只 *zhǐ* is commonly used as the classifier for counting “animals” in Chinese. However, there are also instances when nouns with similar properties use different classifiers (Guo and Zhong, 2005). For instance, 条 *tiáo*, which is used for long and thin objects like “ropes”, is also used as the classifier for “snakes”.

The Japanese distribution of classifiers is dependent on their referent classes: animates, concrete inanimates and abstract inanimates (Downing, 1996).

Bond and Paik (2000) identified five major types of classifiers that have different properties depending on the context. These are sortal, event, mensural, group and taxonomic.

Sortal classifiers are defined as those which classify the kind of noun they count (e.g. 辆 *liàng*, the classifier for “vehicle”). Mensural classifiers measure some property of the object denoted by the noun they modify (e.g. 米 *mǐ* “metre”). Event classifiers are used to count events (e.g. 次 *cì* “time”). Group classifiers refer to a set of individuals belonging to the type denoted by the noun (e.g. 组 *zǔ* “group”). Taxonomic classifiers exert a generic interpretation of the noun phrase that they modify (e.g. 种 *zhǒng* “kind”).

Understanding the classifier systems of these languages can help us to appreciate how the languages cover the hierarchy of meaning with the use of classifiers. In addition, it will provide us with some insight on how speakers of these languages view the world.

Classifier systems are usually very complex. They are also one of the more difficult aspects of grammar to acquire. Even native speakers may have difficulty using some of them. Furthermore, classifiers are often poorly translated. The wrong classifier may be used or left out altogether. Hence, we hope that this study can help to improve the accuracy and efficiency of machine translation. The results of this study can also benefit learners of Japanese and Chinese by helping them retrieve the appropriate classifier when forming noun phrases.

This paper is structured as follows. Section 2 gives a brief overview of work that has been conducted in this area thus far. It also introduces the resources that we will be using for this study. Section 3 documents the methodology. Then, we present the results in Section 4 and discuss the significance of the results and the limitations faced in Section 5.

2 Background

There has been much more work done on analyzing classifiers than in generating them in natural language processing. One important study investigating the generation of classifiers in Thai was carried out by Sornlertlamvanich, Pantachat & Meknavin (1994). The authors proposed an algorithm for matching an appropriate classifier with a noun. Their study involved obtaining noun-classifier pairs from a tagged, word-segmented corpus. From the pattern of noun-classifier collocations, they determined the best representative classifier for each noun and semantic class. However, they did not include a detailed evaluation of the accuracy of their algorithm.

Bond and Paik (2000, 2001) presented a modified algorithm based on Sornlertlamvanich et al's (1994) work. This modified algorithm was used for associating classifiers with semantic classes in Japanese and Korean. It is able to handle nouns which belong to more than one semantic class. It does this by organizing the semantic classes according to the noun referent's most frequent use. The general idea is to assign the default classifier of the most typical semantic class to the noun.

Resources

There are 145,000 synsets for different parts of speech (nouns, adjectives, verbs, adverbs) in the Princeton WordNet of English v3.0 (PWN: Fellbaum 1998). The structure of WordNet allows one to see the relationship between words such as hypernyms (superordinates) and hyponyms (subordinates). It is often used for work in natural language processing.

The Japanese Wordnet (JWN: Isahara *et al* 2008), contains about 57,238 synsets based on the same lexical arrangement as PWN. This means that lexical units in the Japanese Wordnet were arranged according to their hierarchical connections among words as well. However, the Japanese and English wordnets are not a direct copy of each other; for instance, there are Japanese synsets that are not found in the English wordnet and vice versa due to the uniqueness of both languages (Isahara et al., 2009). One example is the concept of "rice". Japanese makes a distinction between 米 *kome* "rice" and 御飯 *gohan* "cooked rice". This distinction is not made in English, and therefore the English wordnet does not include a separate entry for the two senses.

The Wordnet used for Chinese is a bilingual Chinese-English Wordnet (CWN: Xu, Gao, Pan, Qu and Huang, 2008). It is a bilingual lexical database, which also uses the semantic hierarchy from WN. This Chinese-English Wordnet has more than 150,000 Chinese words. Each Chinese synset is linked to an English synset, which allows for useful cross-language information retrieval.

We used a 38,000 sentence Japanese-English-Chinese corpus, the NICT Multilingual corpus, (Zhang, Uchimoto, Ma and Isahara, 2005) based on the Kyoto text corpus. The corpus was created using Japanese sentences from Mainichi Newspaper and manually translated into Chinese and English.

3 Methodology

This section documents the steps taken in the study.

3.1 Categorisation of classifiers

We extracted 228 Japanese classifiers and 264 Mandarin Chinese numeral classifiers from the corpus. This was done by extracting anything tagged as classifier for part-of-speech (POS). For Japanese, we pulled out every word that was tagged with *meishi-setsubi-jousuushi* "noun-suffix-classifier". For Chinese, q.* was the POS for classifier.

These were sorted into the following categories: sortal, mensural, date and time, currency and not classifier. Sortal and mensural classifiers were defined as mentioned before. Date and time classifiers measure the span of days and time periods (such as 年 *nián* "year" and 秒 *miǎo* "second"). Currency classifiers are used to refer to a country's currency (such as 美元 *měiyuán* "American dollar"). Lastly, nouns which had been paired with a numeral, but were in fact not classifiers, were removed (such as 三页 *sānyè* “three pages” and 两餐 *liǎngcān* “two meals”).

3.2 Hand annotation of corpus

Distribution pattern	Example
classifier-*no*-noun (NUM)+CL *no* (NOUN)+	2匹の犬 2-*hiki-no-inu* “2 of the dogs”
noun-*no*-classifier (NOUN)+ *no* (NUM)+CL	犬の２匹 *inu-no-2-hiki* “2 dogs”
noun-*ga/wo/mo/wa*-classifier (NOUN)+ *ga*|*wo*|*mo*|*wa* (NUM)+CL	犬が/を/も/は/２匹 *inu-ga/wo/mo/wa-2-hiki* *“*dogs, 2”

Table 1. Distribution pattern for Japanese

Distribution pattern:
(DET)? (NUM)+ CL[1] (NOUN)+ [2]

Table 2. Distribution pattern for Chinese

55 sortal classifiers were identified for Japanese and we extracted sentences containing noun phrases that are modified by those sortal classifiers. We did the same for Chinese. Chinese had more classifiers with 136 sortal classifiers identified in the previous step. Classifiers that appeared more than 100 times had their counts reduced. We used distribution patterns of classifiers and nouns to retrieve sentences with the numeral-classifier combination and the noun phrases. These distribution patterns were identified using language dependent patterns. Tables 1 and 2 show the distribution patterns identified for Japanese and Chinese respectively.

The distribution patterns that we identified were able to retrieve many correct matches with the numeral-classifier combination and target noun phrase. However, there were some instances when the noun phrase identified was incomplete, as shown in (3) below.

(3)

三千七百	名	会员
sānqiānqībǎi	*míng*	*huìyuán*
3700	CL	member

几	普通市民
jí	*pǔtōngshìmín*
and	citizen

“3700 members of the party and citizens.”

In (3), the target noun phrase picked out by the regular expression was “members of the party”. However, the entire target noun phrase should be “members of the party and citizens”. In such instances, we had to redefine the correct boundaries for the full noun phrase.

After retrieving the matches made by the regular expression, we tagged the classifiers to the target noun phrases by hand and marked the boundaries for both classifier and target noun phrase. We also marked the type of relationship they had: sortal, mensural, event, group, anaphoric, non-classifier and other. In cases where the target noun phrase was present in the sentence but was syntactically distant from the numeral classifier, as in (4), the relationship was marked as anaphoric.

(4)

苦情	は	毎月
kujou$_T$	*wa*	*mai-getsu*
complaints	wa	every-month

平均	600件	に	上る
heikin	600-*ken$_1$*	*ni*	*noboru,*
average	600-CL	ni	add up

前年	より	約
zen-nen	*yori*	*yaku*
last-year	than	about

2000件	増	の
2000-ken	*zou*	*no*
2000-CL	increase	no

18320件	を	摘発した
18320-ken$_2$	*wo*	*tekihatsushi-ta*
18,320-CL	wo	expose-PST

“The number of complaints is as many as 600 per month on average and the police wrote tickets for 18, 320 cases this year, up about 2,000 from last year.”

In the example, the target of the first 件 *ken*,a classifier used for things like “cases” or

[1] There can only be one classifier in an expression.

[2] DET: Determiner, NUM: Numeral, CL: Classifier, ?: 0 or 1, +: 1 or more

“matters”, is 苦情 *kujou* “complaint”. Therefore, the classifier was tagged to 苦情 *kujou* “complaint” with a sortal relationship. However, for the second *ken,* although the referent 苦情 *kujou “*complaint” is still in the sentence, it is not in the same clause as the classifier, therefore, for this classifier, it was tagged as anaphoric.

Anaphoric targets share a sortal relationship with the classifier that modifies the noun phrase. When the target is anaphoric, and the relationship between the target and classifier is under ‘other’, the set will be tagged as other instead of anaphoric.

Any instance of synecdoche was tagged as ‘other’. One instance of synecdoche was found for 名 *míng* (one of the classifiers for “people”), as shown in (5) below:

(5) 六 名 自民党
liù *míng* *zìmíndǎng*
6 CL Liberal Democratic Party
“6 members of the Liberal Democratic Party”

In this example, the numeral-classifier pair counts the number of party members and not the number of political parties.

Numeral-classifier combinations that were being used in an ordinal sense were tagged as ‘not’. In addition, noun phrases with very abstract referents like hope and courage were also tagged as ‘not’ as they were considered uncountable.

3.3 Assignment of synsets to classifiers

Then, we associated synsets from Japanese and Chinese Wordnet to each of these classifiers by hand. We looked up the semantic class for each target noun phrase and checked if it was suitable for the classifier. We also checked how high in the semantic hierarchy the use of the classifier could extend to. Table 3 illustrates how we assigned synsets to 个 *gè* (general classifier) and 只 *zhǐ* (animal classifier) based on the semantic hierarchy shown in Figure 1.

Classifier	**Usage**	**Synsets**
个 *gè*	general	+00001740-n -00015388-n
只 *zhǐ*	animal	+00015388-n -02374249-n -02512053-n -01726692-n
条 *tiáo*	fish snake	+02512053-n +01726692-n
匹 *pǐ*	equine	+02374149-n

Table 3. Assignment of synsets to classifiers

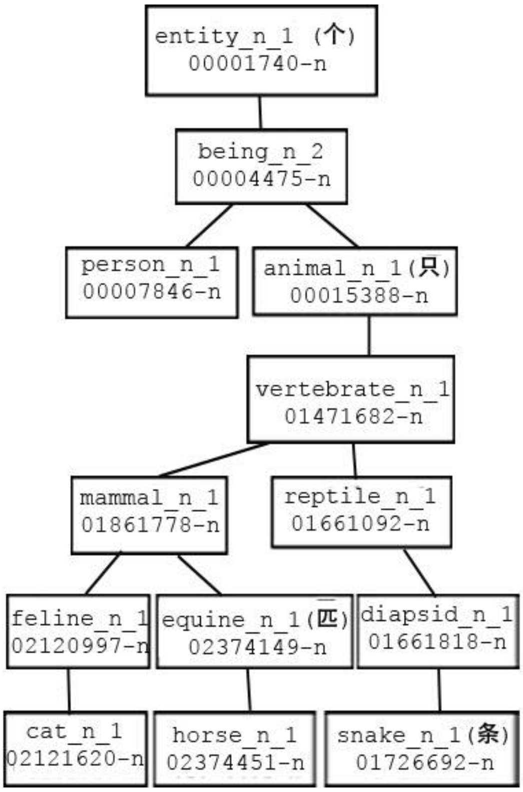

Figure 1. Semantic hierarchy in Chinese Wordnet

In both Chinese and Japanese, there is a general classifier that can be used for any entity when there is no specific classifier (个 *gè* for Chinese and 個 *ko* for Japanese). While 个 *gè* in Chinese can be used to count both “humans” and “objects”, 個 *ko* in Japanese cannot be used for “humans”.

Since 个 *gè* is considered a general classifier, we assigned the semantic class of `entity_n_1` to it. A + symbol added in front of the synset signifies that all synsets below `entity_n_1` will share the same classifier, 个 *gè*. We also removed `animal_n_1` from it because animals are usually not counted with this classifier. The − symbol added in front of the synset signifies that the synset `animal_n_1` does not share the same classifier.

只 *zhǐ* is the classifier used to count “animal”. We assigned `animal_n_1` to it. We also removed `fish_n_1`, `snake_n_1`, `equine_n_1` from it. These animals are not counted with the default animal classifier in Chinese. “Fish” and “snakes” are counted with 条

tiáo and “horses” and “mules” are counted with 匹 *pǐ*. There were quite a number of nouns that did not have an entry in both the Chinese and Japanese Wordnets. Thus, we had to add these nouns into the Wordnets. Due to time constraints, we only added nouns which occurred very frequently in the extracted noun phrases.

3.4 Generation of classifiers

We ran the annotated data through a program such that it picked out the head noun from a noun phrase (6). This is achieved by having the program go through the noun phrase and match it to a synset in the Wordnet, with the assumption that noun phrases are right-headed.

Based on example (6), the head noun that is retrieved from the noun phrase is オオカミ *ookam*i “wolf”. The Chinese example in (7) shows that the head noun extracted and matched with a synset is 飞机 *fēijī* “aircraft”.

(6) 灰色 オオカミ
hai-iro *ookami*
gray-colour wolf
“gray wolf”

(7) 轻型 飞机
qīngxíng *fēijī*
light aircraft
“light aircraft”

We used the modified algorithm by Bond and Paik (2000) to generate the classifier. Based on this algorithm, the classifier most closely associated to the head noun in terms of semantic class was generated (Figure 2). All hyponyms under that synset will be counted with the same classifier unless a specific classifier has been marked for it.

Based on intuition, the algorithm will select the classifier marked on the closest possible hypernym. For example, the synset `antelope_n_1` is compatible with 只 *zh*ǐ, which is the classifier for “animals” and 头 *tóu*, which is the classifier for certain types of animals like “pigs”, “cattle”, “elephants” and “livestock”. Since 头 *tóu* is associated with `bovid_n_1`, which is the hypernym of `antelope_n_1`, we select 头 *tóu* as the classifier for `antelope_n_1`.

```
For a noun phrase:
  (a) find its synset
  (b) find all hypernyms and see if any are selected for by classifiers
  (c) select the classifier from the synset with the highest similarity
```

Figure 2. Algorithm to generate a classifier

To analyse noun phrases with more than one sense, we used an algorithm shown in Figure 3. This algorithm identifies the semantic class of the noun based on the classifier used.

From the Japanese Wordnet, シイル *shiiru* “seal” has at least two senses:

a. `seal_n_9` marine mammal (subset of mammal)

b. `seal_n_5` a stamp affixed to a document (subset of stamp)

Since the classifier 匹 *hiki* is tagged to the semantic class of animal, the sense of シイル *shiiru* “seal” implied in the example is that of the synset `seal_n_9`, a marine mammal which is a hyponym of animal.

For a noun phrase:
(a) Identify the semantic class of the head noun based on the classifier present

e.g. シイル ２匹 買いました
shiiru *2-hiki* *kaimashita*
seal 2-CL buy-PST
“I have two seals”

Figure 3. Algorithm to analyse a noun phrase

In this generation stage, we only looked at nouns with unique synsets that had been identified in the analysis stage. Lastly, we tested the predictions made using Wordnet to check the accuracy of the predictions.

4 Results

Japanese Classifier		Usage	Count
人	*nin*	“person”	122
件	*ken*	“(abstract) matters/cases”	69
台	*dai*	“vehicles”; “machines”	58
社	*sha*	“companies”; “shrines”	47

本 *hon*	long, thin objects e.g. “roads/ties/pencils”	45
枚 *mai*	thin, flat objects e.g. “papers/photographs/plates”	40
個 *ko*	general measure word; “military units”	30
点 *ten*	“pieces of a set”; “goods/items”	21
棟 *tou*	“buildings/apartments”	15
戸 *ko*	“houses”	15

Table 4. Ten most frequent Japanese classifiers

Chinese Classifier	**Usage**	**Count**
家 *jiā*	“families/businesses”	112
名 *míng*	“people”	107
个 *gè*	“people/objects”	111
场 *chǎng*	“events e.g. exams/sporting events”	95
件 *jiàn*	“things/clothes”	93
位 *wèi*	“people” (honorific)	87
条 *tiáo*	“long and thin things e.g. snakes/rivers/ropes” “lives”	82
辆 *liàng*	“vehicles”	80
张 *zhāng*	“flat objects/things with flat surfaces e.g. beds, paper” “votes”	74
句 *jù*	“phrases/lines of verse/sayings”	70

Table 5. Ten most frequent Chinese classifiers

Tables 4 and 5 show the ten most frequent classifiers in Japanese and Chinese respectively.

Classifier type	**Japanese (J)**	**Chinese (C)**
Sortal	592	1906
Anaphoric	133	113
Event	61	26
Group	7	41
Other	407	267
Non-classifier	142	921
Total	1400	3274

Table 6. Results of hand annotation

Table 6 summarises the results of the hand annotation for Japanese and Chinese. The total number of classifier phrases in Chinese is much more than in Japanese. This is because Chinese classifiers can also appear with determinatives like 这 *zhè* “this”, not just with numerals.

As shown in Table 6, the number of sortal classifiers in Japanese that share a sortal relationship with the target is less than half. This is much lower than we expected.

For Chinese, there were 1,906 noun phrases modified by a sortal classifier. This was slightly more than half of the total number of extracted sentences. Classifiers tagged as ‘other’ and ‘not’ were mostly being used in an ordinal sense or had uncountable abstract noun referents.

Scores	**% (J)**	**Total (J)**	**% (C)**	**Total (C)**
Correctly analysed	76.33	405	79.37	1312
Total	100	528	100	1653
Correctly generated	89.84	116	78.80	223
Total	100	129	100	283

Table 7. Analysis and generation scores

Table 7 presents the results of our evaluation.

The analysis score tells us how often we match a noun’s semantic class and the generation score tells us how often we correctly generate a classifier. A generated classifier is judged to be correct if it exactly matched the original classifier used in the annotated corpus. For Japanese, the analysis score is 76.33% and the generation score is higher at 89.84%. The analysis score is slightly lower by 3.04% than that for Chinese. However, the generation score is higher by 11.04% as compared to Chinese.

5 Discussion

For this study, we only considered classifiers that share a sortal relationship with the noun phrase they modify. Noun phrases that are modified by classifiers that share a group, event, mensural relationship were not included in the evaluation. Similarly, noun phrases in which the target is anaphoric were also not included.

Based on the results of the hand annotation, most sortal classifiers often have an anaphoric use. The anaphoric target can either be in the same sentence but different clause or in a different sentence.

Overall, the evaluation of both algorithms is satisfactory. For Chinese, we were able to analyse correctly 79.37% or 1312 noun phrases. By using the default classifier assigned to each semantic class, we were able to generate correctly 78.80% or 223 classifiers.

For Japanese, we were able to correctly analyse 76.33% or 405 noun phrases and generate 89.84% or 108 classifiers.

One of the issues we faced in this study is the problem of dealing with synecdoche, particularly for Chinese. In example (5), given in Section 3.2, we saw the classifier 名 *míng* being used to count the number of members of the political party and not the number of political parties.

Based on our mapping of classifiers to semantic class, the "Liberal Democratic Party" would belong to the semantic class of organization which uses a different classifier 个 *gè*. However, this type of synecdoche will not be captured using the current method of analysis.

For Japanese, a large number of classifiers that were tagged with having "other" relationship with the targets were in fact functioning as ordinal classifiers. These were often preceded by an ordinal prefix (8) or followed by the ordinal suffix 目 *me* (9).

(8) 第一棟　の　旅館
dai-1-dou no　ryokan
ORD-1-CL　no　hotel
"the first hotel"

(9) 二回目　の　優勝
ni-kai-me　no　yuushou
2-CL-ORD　no　victory
"the second victory"

One of the limitations of this study is the coverage of Wordnet. During our assignment of synsets to classifiers, we found that approximately 20% of our target noun phrases were not represented in CWN. For instance, 球队 *qiú duì* "a team for ball sports" was not included. Although it had synsets for 棒球队 *bàngqiúduì* "baseball team" (`baseball_team_n_1`) and 篮球队 *lánqiúduì* "basketball team", (`basketball_team_n_1`), it did not have a generic term to refer to a team that played ball sports. For the Japanese study too, there are some nouns that are not yet represented in JWN, for example 大手 *oote* "major company" or チッシュウ *tisshuu* "tissues". In addition, the lexicon does not include proper nouns like names of companies like 三菱電機 *mitsubishi denki* "Mitsubishi Electric Corporation". In a similar manner, a noun may be present in Wordnet but is missing the correct sense. One example is 白紙 *hakushi* "blank paper". Although this noun is represented in JWN, the sense given is that of "fresh start". The lack of a corresponding noun or sense in the lexicon may have affected the evaluation scores.

Similarly, some nouns which were represented in CWN were also missing other senses. For instance, 车 *chē* has two senses. The first sense is "car" and the second sense is "rook", a type of chess piece. When we looked up 车 *chē* "car" CWN presented us with `rook_n_2`. This sense is the subset of corvine bird.

In this case, the noun is being represented by the wrong synset and one of its sense "car" was also missing. Although we assigned the synset for `vehicle_n_1` to the classifier 辆 *liàng* (the classifier for "vehicles"), it was not able to generate this classifier when it encountered 车 *chē* "car" in the test sentences.

In addition, since the Chinese sentences were translated from Japanese sentences from the Mainichi Newspaper, there were also a few Japanese loanwords which were not represented in Chinese Wordnet. Some of these include 榻榻米 *tàtàmǐ* "tatami" and 横岗 *hénggāng* "yokozuna". This could also account for the lower generation scores obtained for Chinese as compared to Japanese. Although CWN covers more synsets, its coverage of common senses appears to be slightly worse than JWN.

Some nouns can also be used with more than one classifier. For instance, in Japanese, 住宅 *juutaku* "residence" can be used with classifiers 軒 *ken*, 個 *ko*, or 棟 *tou*, all three being classifiers for houses. In the corpus, there were instances of all three classifiers being used to quantify 住宅 *juutaku* "residence". The choice of which classifier to use is up to the individual's personal preference. Hence, it is difficult to predict the correct classifier for cases like this.

Another issue that may have contributed to the generation error is the problem of fixed expressions, particularly for Chinese. This was often seen with the classifier 口 *kǒu*, which is used for counting things with mouths. It is commonly used to count the number of people in a family or household, as shown in (10).

(10) 一家 五 口 人
yījiā *wǔ* *kǒu* *rén*
a family5 CL person
“a family of five.”

Although (10) shows that 口 *kǒu* can be used to count “person”, this classifier is generally used this way only when it follows 家 *jiā* “family”. 个 *gè* is the more common classifier to use when counting “person”. Such fixed expressions cannot be properly analysed with our current methods of analysis and will require special processing.

Shape, size and animacy are some factors that play a part in selecting the correct classifier (Allan, 1977). For instance, 张 *zhāng* is used to count flat objects or things with a flat surface. Some examples include tables, stools, papers, newspapers and beds. However, the wordnets do not contain such information about shape or size of the nouns. Hence, some world knowledge is still required in order to predict the right classifier for a target noun phrase.

In order to further research on Japanese and Chinese classifiers, we release the following data for both languages under the creative-common attribution license (CC-by): (i) the table of classifier phrase + antecedent noun phrase pairs with their disambiguated synset (ii) the table of which synsets are classified by which classifiers (Figure 1). It is available from the Japanese Wordnet page: http://nlpwww.nict.go.jp/wn-ja/. We are also feeding back information on missing senses to the respective Wordnet projects.

6 Conclusion

In this paper, we presented an algorithm to generate numeral classifiers based on semantic hierarchies present in wordnets. For Chinese, it was shown to select the correct sortal classifier 78.80% of the time. We believe that this score can be raised with improvements to Chinese Wordnet. For Japanese, it was shown to select the correct sortal classifier 89.84% of the time. At the present moment, the wordnets do not provide a full coverage of all the nouns in the world. In addition, there are factors that may guide the choice of selection, making a purely taxonomic hierarchy inadequate. This study has shown that the selection of a classifier based only on a taxonomic hierarchy may not be accurate all the time because semantic attributes of the noun are also important. Future studies can work to improve on the coverage of wordnet and also perhaps expand the wordnet in terms of linking semantic attributes. World knowledge is also required in order to select the most suitable classifier.

Acknowledgements

We wish to acknowledge the funding support for this project from Nanyang Technological University under the Undergraduate Research Experience on Campus (URECA) programme. We thank NICT for granting permission to use the NICT multi-lingual corpus.

References

Allan, K. Classifiers. *Language*, 53:285-311, 1977.

Bond, F. and Paik, K. Re-using an ontology to generate numeral classifiers. In *18th International Conference on Computational Linguistics: COLING-2000*, pp. 90-96, Saarbrucken, 2000.

Downing, P. *Numeral classifier systems: The case of Japanese*. John Benjamins, Philadelphia: 1996.

Fellbaum C. *WordNet*, MIT Press, 1998

Guo, H., and Zhong, H. Chinese classifier assignment using SVMs. In *4th Sighan Workshop on Chinese Language Processing*, pp. 25-31, Jeju Island, 2005.

Hundius, H. and Kölver, U. Syntax and semantics of numeral classifiers in Thai. *Studies in Language*, 7(2), pp. 165-214, 1983.

Isahara H., Bond F., Uchimoto K., Utiyama M. and Kanzaki K. Development of Japanese WordNet. In *LREC-2008*, Marrakech. 2008

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

Sornlertlamvanich, V., Pantachat, W. and Meknavin, S. Classifier assignment by corpus based approach. In *15th International Conference on Computational Linguistics: COLING-94*, pp. 556-561, Kyoto, August 1994.

Xu, R., Gao, Z., Pan, Y., Qu, Y., and Huang, Z. An Integrated Approach for Automatic Construction of Bilingual Chinese-English WordNet. In *Proceedings of ASWC2008*, pp. 302-314, Bangkok, Thailand, 2008.

Zhang, H. Numeral classifiers in Mandarin Chinese. *Journal of East Asian Linguistics*, 16:43-59, 2007.

Zhang, Y., Uchimoto, K., Ma, Q. and Isahara, H. Building an Annotated Japanese-Chinese Parallel Corpus – A Part of NICT Multilingual Corpora. In *10th Machine Translation Summit Proceedings*, pp. 71-78, Phuket, 2005.