Mapping and Generating Classifiers using an Open Chinese Ontology

Luis Morgado da Costa,[♠] **Francis Bond**[♠] **Helena Gao**[◇]
[♠]Linguistics and Multilingual Studies [◇]Chinese
Nanyang Technological University
Singapore
<luis.passos.morgado@gmail.com,bond@ieee.org,HELENAGAO@ntu.edu.sg>

Abstract

In languages such as Chinese, classifiers (CLs) play a central role in the quantification of noun-phrases. This can be a problem when generating text from input that does not specify the classifier, as in machine translation (MT) from English to Chinese. Many solutions to this problem rely on dictionaries of noun-CL pairs. However, there is no open large-scale machine-tractable dictionary of noun-CL associations. Many published resources exist, but they tend to focus on how a CL is used (e.g. what kinds of nouns can be used with it, or what features seem to be selected by each CL). In fact, since nouns are open class words, producing an exhaustive definite list of noun-CL associations is not possible, since it would quickly get out of date. Our work tries to address this problem by providing an algorithm for automatic building of a frequency based dictionary of noun-CL pairs, mapped to concepts in the Chinese Open Wordnet (Wang and Bond, 2013), an open machine-tractable dictionary for Chinese. All results will released under an open license.

1 Introduction

Classifiers (CLs) are an important part of the Chinese language. Different scholars treat this class of words very differently. Chao (1965), the traditional and authoritative native Chinese grammar, splits CLs into nine different classes. Cheng and Sybesma (1998) draw a binary distinction between *count-classifiers* and *massifiers*. Erbaugh (2002) splits CLs into three categories (*measure*, *collective* and *sortal classifiers*). Measure classifiers describe quantities (e.g. 'a bottle of', 'a mouthful of'), collective classifiers describe arrangement of objects ('a row of', 'a bunch of'), and sortal classifiers refer to a particular noun category (which can be defined, for example, by shape). Huang et al. (1997) identify four main classes, *individual classifiers*, *mass classifiers*, *kind classifiers*, and *event classifiers*. And Bond and Paik (2000) define five major types of CLs: *sortal* (which classify the kind of the noun phrase they quantify); *event* (which are used to quantify events); *mensural* (which are used to measure the amount of some property); *group* (which refer to a collection of members); and *taxonomic* (which force the noun phrase to be interpreted as a generic kind). This enumeration is far from complete, and Lai (2011) provides a detailed literature review on the most prominent views on Chinese classifiers.

Most languages make use of some of these classes (e.g. most languages have measure CLs, as in *a kilo of coffee*, or group CLs, as in *a school of fish*). What appears to be specific to some languages (e.g. Chinese, Japanese, Thai, etc.) is a class of CLs (**sortal classifiers**: **S-CL**) that depicts a selective association between quantifying morphemes and specific nouns. This association is licensed by a number of features (e.g. physical, functional, etc.) that are shared between CLs and nouns they can quantify, and these morphemes add little (but redundancy) to the semantics of noun-phrase they are quantifying.

Consider the following examples of S-CL usage in Mandarin Chinese:

(1) 两 只 狗
liǎng zhǐ gǒu
2 CL dog
"two dogs"

(2) 两 条 狗
liǎng tiáo gǒu
2 CL dog
"two dogs"

(3) 两 条 路
liǎng tiáo lù
2 CL road

"two roads"

(4) 三 台 电脑
sān tái diànnǎo
3 CL computer

"three computers"

(5) *三 只 电脑
sān zhǐ diànnǎo
3 CL computer

"three computers"

Examples (1) through (4) show how the simple act of counting in Mandarin Chinese involves pairing up nouns with specific classifiers, if incompatible nouns and classifiers are put together then the noun phrase is infelicitous, see (5).

Different S-CLs can be used to quantify the same noun, see (1) and (2), and the same type of S-CL can be used with many different nouns – so long as the semantic features are compatible between the S-CL and the noun, see (2) and (3). Extensive work on these features is provided by Gao (2010) – where more than 800 classifiers (both sortal and non-sortal) are linked in a database according to the nominal features they select, but providing only a few example nouns that can be quantified by each CL. These many-to-one selective associations are hard to keep track of, especially since they depend greatly on context, which often restricts or coerces the sense in which the noun is being used (Huang et al., 1998).

(6) 一 个 木头
yī ge mùtou
1 CL log (of wood) / blockhead

"a log / blockhead"

(7) 一 位 木头
yī wèi mùtou
1 CL blockhead

"a blockhead"

(8) 一 根 木头
yī gēn mùtou
1 CL log (of wood)

"a log"

Examples (6–8) show how the use of different CLs with ambiguous senses can help resolve this ambiguity. In (6), we can see that with the use of 个 *ge*, the most general S-CL in Mandarin Chinese, mu4tou is ambiguous because it does not restrict the noun's semantic features. With the use of 位 *wèi* (7), an honorific S-CL used almost exclusively with people, it can only be interpreted as "blockhead". And the reverse happens when using 根 *gēn* (8), a S-CL for long, slender, inanimate objects: the sense of *log (of wood)* of 木头 *mùtou* is selected.

Even though written resources concerning CLs are abundant, they are not machine tractable, and their usage is limited by copyright. Natural Language Processing (NLP) tasks depend heavily on open, machine tractable resources. Wordnets (WN) are a good example on the joint efforts to develop machine tractable dictionaries, linked in rich hierarchies. Resources like WNs play a central role in many NLP tasks (e.g. Word Sense Disambiguation, Question Answering, etc.).

Huang et al. (1998) argue that the integration between corpora and knowledge rich resources, like dictionaries, can offer good insights and generalizations on linguistic knowledge. In this paper, we follow the same line of thought by integrating both a large collection of Chinese corpora and a knowledge rich resource (the Chinese Open Wordnet: COW (Wang and Bond, 2013)). COW is a large open, machine tractable, Chinese semantic ontology, but it lacks information on noun-CL associations. We believe that enriching this resource with concept-CL links will increase the domain of it's applicability. Information about CLs could be used to generate CLs in MT tasks, or even to improve on Chinese Word Sense Disambiguation.

The remainder of this paper is structured as follows: Section 2 presents related work, followed by a description of the resources used in Section 3; Section 4 describes the algorithms applied, and Section 5 presents and discusses our results; Section 6 describes ongoing and future work; and Section 7 presents our conclusion.

2 Related Work

Mapping CLs to semantic ontologies has been attempted in the past (Sornlertlamvanich et al., 1994; Bond and Paik, 2000; Paik and Bond, 2001; Mok et al., 2012). Sornlertlamvanich et al. (1994) is the first description of leveraging hierarchical

semantic classes to generalize noun-CL pairs (in Thai). Still, their contribution was mainly theoretical, as it failed to report on the performance of their algorithm. Bond and Paik (2000) and Paik and Bond (2001) further develop these ideas to develop similar works for Japanese and Korean. In their work, CLs are assigned to semantic classes by hand, and achieve up to 81% of generation accuracy, propagating CLs down semantic classes of Goi-Taikei (Ikehara et al., 1997). Mok et al. (2012) develop a similar approach using the Japanese Wordnet (Isahara et al., 2008) and the Chinese Bilingual Wordnet (Huang et al., 2004), and report a generation score of 78.8% and 89.8% for Chinese and Japanese, respectively, on a small news corpus.

As it is common in dictionary building, all works mentioned made use of corpora to identify and extract CLs. Nevertheless, extracting noun-CL associations from corpora is not a straightforward task. Quantifier phrases are often used without a noun, resorting to anaphoric or deictic references to what is being quantified (Bond and Paik, 2000). Similarly, synecdoches also generate noise when pattern matching (Mok et al., 2012).

3 Resources

Our corpus joins data from three sources: the latest dump of the Chinese Wikipedia, the second version of Chinese Gigaword (Graff et al., 2005) and the UM-Corpus (Tian et al., 2014). This data was cleaned, sentence delimited and converted to simplified Chinese script. It was further preprocessed using the Stanford Segmentor and POS tagger (Chang et al., 2008; Tseng et al., 2005; Toutanova et al., 2003). The final version of this corpus has over 30 million sentences (950 million words). For comparison, the largest reported corpora from previous studies contained 38,000 sentences (Mok et al., 2012). In addition, we also used the latest version (2012) of the Google Ngram corpus for Chinese (Michel et al., 2011).

There are some differences between the usage of classifiers in different dialects and variations of Chinese in these different corpora, but our current goal focused on collecting generalizations. Future work could be done to single out differences across dialects and variants.

We used COW (Wang and Bond, 2013) as our lexical ontology, which shares the structure of the Princeton Wordnet (PWN) (Fellbaum, 1998). To minimize coverage issues, we enriched it with data from the Bilingual Ontological Wordnet (BOW) (Huang et al., 2004), the Southeast University Wordnet (SEW) (Xu et al., 2008), and automatically collected data from Wiktionary and CLDR, made available by the Extended OMW (Bond and Foster, 2013). The final version of this resource had information for over 261k nominal lemmas, from which over 184k were unambiguous (i.e. have only a single sense).

We filtered all CLs against a list of 204 S-CLs provided by Huang et al. (1997). Following Lai (2011), we treated both Huang's *individual classifiers* and *event classifiers* as S-CLs.

4 Our Algorithm

Our algorithm produces two CL dictionaries with frequency information: a lemma based dictionary, and a concept based dictionary, using COW's extended ontology. We tested both dictionaries with a generation task, automatically validated against a held out portion the corpus.

4.1 Extracting Classifier-Noun Pairs

Extracting CL-noun pairs is done by matching POS patterns against the training section of our corpus. To avoid, as much as possible, noise in the extracted data, we choose to take advantage of our large corpus to apply restrictive pattern variations of the basic form: (determiner or numeral) + (CL) + (noun) + (end of sentence punctuation/select conjunctions). Our patterns assure that no long dependencies exist after the CL, and try to maximally reduce the noise introduced by anaphoric, deictic or synecdochic uses of classifiers (Mok et al., 2012). Variations of this pattern were also included to cover for different segmentations produced by the preprocessing tools.

If an extracted CL matches the list of S-CLs, we include this noun-CL pair in the lemma based dictionary. The frequency with which a specific noun-CL pair is seen in the corpus is also stored, showing the strength of the association.

Extracting noun-CL pairs from the Chinese Google NGram corpus required a special treatment. We used the available 4 gram version of this corpus to match a similar pattern (and variations) to the one mentioned above: (determiner or numeral) + (CL) + (X) + (end of sentence punctuation/select conjunctions). Given we had no POS information available for the NGram corpus, we

used regular expression matching, listing common determiners, numerals, punctuation, and our list of 204 S-CLs. We did not restrict the third gram. We also transferred the frequency information provided for matched ngrams to our lemma based dictionary.

Our training set included 80% of the text portion of the corpus, from which we extracted over 435k tokens of noun-CL associations, along with the full Chinese Google NGram corpus, from which we extracted 13.5 million tokens of noun-CL associations.

This lemma based dictionary contained, for example, 59 pairs of noun-CL containing the lemma 类别 *lèibié* "category". It occurred 58 times with the CL 个 *ge*, and once with the CL 项 *xiàng*. Despite the large difference in frequencies, both CLs can be used with this lemma. Another example, where the relevance of the frequency becomes evident, is the word 养鸡场 *yǎngjīchǎng* "chicken farm", which was seen in our corpus 12 times: 6 times with the CL 个 *ge*, 3 times with the CL 家 *jiā*, twice with the CL 只 *zhǐ*, and once with the CL 座 *zuò*. Chinese native speaker judgments identified that three out of the 4 CLs identified were correct (个 *ge*, 家 *jiā* and 座 *zuò*). In addition, two other classifiers would also be possible: 间 *jiān* and 所 *suǒ*. This second example shows that while the automatic matching process is still somewhat noisy, and incomplete, the frequency information can help to filter out ungrammatical examples. When used to generate a classifier, our lemma based dictionary can use the frequency information stored for each identified CL for a particular lemma, and choose the most frequent CL. This process will likely increase the likelihood of it being a valid CL. Also, by setting a minimum frequency threshold for which noun-CLs pair would have to be seen before being added to the dictionary, we can exchange precision for coverage.

4.2 Concept Based Dictionary

The concept based dictionary is created by mapping and expanding the lemma based dictionary onto COW's expanded concept hierarchy. Since ambiguous lemmas can, in principle, use different CLs depending on their sense, we map only unambiguous lemmas (i.e. that belong to a single concept). This way, each unambiguous entry from the lemma based dictionary matching to COW contributes information to a single concept. Frequency information and possible CLs are collected for each matched sense. The resulting concept-based mapping, for each concept, is the union of CLs for each unambiguous lemma along with sum of frequencies.

Following one of the examples above, the lemma 类别 *lèibié*, was unambiguously mapped to the concept ID 05838765-n – defined as "a general concept that marks divisions or coordinations in a conceptual scheme". This concept provides two other synonyms: 范畴 *fànchóu* and 种类 *zhǒnglèi*. In the concept based dictionary, the concept ID 05838765-n will aggregate the information provided by all its unambiguous senses. This results in a frequency count of 132 for the CL 个 *ge*, and of 2 for 项 *xiàng* (both valid uses).

As has been shown in previous works, semantic ontologies should, in principle, be able to simulate the taxanomic features hierarchy that link nouns and CLs. We use this to further expand the concept based dictionary of CLs.

For each concept that didn't receive a classifier, we collect information concerning ten levels of hypernymy and hyponymy around it. If any pair of hypernym-hyponym was associated with the same CL, we assign this CL to the current concept. Since we're interested in the task of generating the best (or most common) CL, we rank CLs inside these expanded concepts by summing the frequencies of all hypernyms and hyponyms that shared the same CL. If more than one CL can be assigned this way, we do so.

Figure 1 exemplifies this expansion. While concepts A, B and C did not get classifiers directly assigned to them, they are still assigned one or more classifiers based on their place in the concept hierarchy. For every concept that didn't receive any CL information, if it has at least a hypernym and a hyponym sharing a CL (within a distance of 10 jumps), then it will inherit this CL and the sum of their frequencies. Assuming a full concept hierarchy is represented in Figure 1, concept A would inherit two classifiers, and concept B and C would inherit one each.

This expansion provides extra coverage to the concept based dictionary. But we differ from previous works in the sense that we do not blindly assign CLs down the concept hierarchy, making it depend on previously extracted information for both hypernyms and hyponyms. By following a

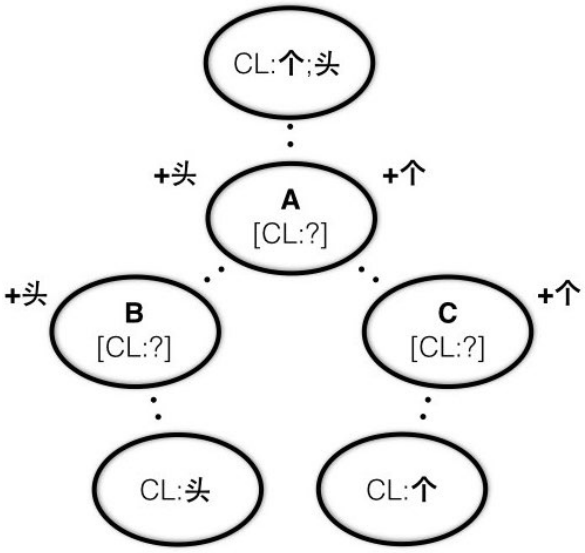

Figure 1: Classifier Expansion

stricter approach, we hope to provide results of better quality.

4.3 Automatic Evaluation

We evaluated both lemma and concept based dictionaries with two tasks: predicting the validity of and generating CLs. We used roughly 10% of held out data (dev-set), from which we extracted about 37,4k tokens of noun-CL pairs, as described in 4.1. We used this data to evaluate the prediction and generation capabilities of both dictionaries in the following ways: predicting the validity of a CL was measured by comparing every noun-CL pair extracted from the dev-set to the data contained in the dictionary for that particular lemma (i.e. if that particular classifier was already predicted by the dictionary); generation was measured by selecting the best likely classifier, based on the cumulative frequencies of noun-CL pairs in the dictionary (i.e. if the classifier seen in the example matched the most frequent classifier). This was done separately for both dictionaries.

When no other classifier had been assigned, we used 个 *ge*, the most frequent CL on the corpus, as the default classifier. And a baseline was established by assigning 个 *ge* as the only CL for every entry.

The dev-set was used to experiment with different thresholds (τ) of the minimum frequency, from one to five, for which noun-CL pairs would have to be seen in the train-set in order to be considered into the dictionaries. These different minimum frequency thresholds were compared between both tasks.

	$\tau=1$	$\tau=3$	$\tau=5$	*Test*
baseline	44.2	44.2	44.2	40.4
All lemmas				
lem-all	92.7	88.5	86.2	93.6
lem-all-mfcl	75.1	73.8	72.8	78.9
lem-all-no-info	4.7	9.2	12.1	4.1
Unamb. lemmas				
lem-unamb	93.2	88.2	85.5	94.5
wn-unamb	**95.1**	90.9	88.3	**95.9**
lem-unamb-mfcl	**77.0**	75.5	74.1	**77.9**
wn-unamb-mfcl	72.3	71.6	70.7	73.5
lem-unamb-no-info	3.4	9.5	13.6	2.8
wn-unamb-no-info	1.7	5.3	8.3	1.5
Coverage				
lemmas-w/cl	32.4k	10.4k	7.0k	
wn-concepts-w/cl	22.7k	15.0k	12.3k	

Table 1: Automatic Evaluation Results

The best performing τ was then tested in a a second held-out set of data (test-set), also containing roughly 10% of the size of the text corpus, roughly 39.9k tokens of noun-CL pairs. The test-set is used to report our final results.

The results are presented in Table 1, and are discussed in the following section.

5 Discussion and Results

In Table 1 we can start to note that the baseline, of consistently assigning 个 *ge* to every entry in the dictionary is fairly high, of roughly 40%.

In order to allow a fair comparison, since we decided that the concept based dictionary would contain only unambiguous lemmas, we only use unambiguous lemmas to compare the performance across dictionaries. All results can be compared across the different thresholds discussed in 4.3. $\tau = 1, 3$ and 5 present the results obtained in the automatic evaluation, using minimum frequencies of one, three and five, respectively.

The first three reported results report exclusively about the lemma dictionary (including both ambiguous and unambiguous lemmas). *lem-all* reports the results of the prediction task, *lem-all-mfcl* reports the results of the generation task, and *lem-all-no-info* reports the relative frequency of lemmas for which there was no previous infor-

mation in the dictionary, and which could have boosted both task's performance by falling back on the default CL $\uparrow$ *ge*.

These initial results show that it was easy to perform better than baseline, and that $\tau = 1$ achieved the best results on both predicting noun-CL pairs, and generating CLs that matched the data.

Comparing different τs shows that, even considering the over-generation reduction that imposing minimum frequencies brings (validated but not presented here), the best generation performance is achieved by not filtering the training data. And this will be consistent across the remainder of the results.

When comparing both dictionaries, we look only at unambiguous lemmas. Similar to what was explained above, *lem-unamb* and *wn-unamb* report the results of the prediction task for the lemma based and concept based dictionary, respectively. The labels *lem-unamb-mfcl* and *wn-unamb-mfcl* report the results for the generation task. And the *lem-unamb-no-info* and *wn-unamb-no-info* report about the lack of informed coverage (where backing-off to the default CL might have help the performance).

Between the lemma and the concept based dictionaries, this automatic evaluation shows that while the concept based dictionary is better at predicting if a noun-CL pair was valid, the lemma based dictionary outperforms the former in the generation task.

The final results of this automatic evaluation are shown in column *Test*, where we re-evaluated the dictionary produced by $\tau = 1$ on the test-set. *Test* shows slightly better results, perhaps because the random sample was easier than the dev-set, but the same tendencies as reported above.

Considering that the concept based dictionary should be able to provide CL information to some lemmas that have not been seen in the training data (either by expansion or by leveraging on a single lemma to provide information about synonyms), we expected the concept based dictionary to present the best results.

Many different reasons could be influencing these results, such as errors in the ontology, the fact that Chinese CLs relate better to specific senses than to concepts (i.e. different lemmas inside a concept prefer different CLs), or noise introduced by the test and dev-set (since we don't have a hand curated golden test-set). For this reason, we decided to hand validate a sample of each dictionary.

Based on a random sample of 100 concepts and 100 lemmas extracted from each dictionary, a Chinese native speaker checked if the top ranked CL (i.e. with highest frequency), that would be used to generate a CL for each of the randomly selected entries, was in fact a valid CL for that lemma or concept. This human validation showed the concept based dictionary outperforming the lemma based dictionary by a wide margin: 87% versus 76% valid guesses. This inversion of performance, when compared to the automatic evaluation, was confirmed to be mainly due to noisy data in the test-set caused by the automatic segmentation and POS tagging.

We then looked at a bigger sample of 200 lemmas and found roughly 7.5% of invalid lemmas in the lemma based dictionary. Conversely, the concept based dictionary assigns CLs by 'bags of lemmas' (i.e. synsets). This allows the noise introduced by a few senses to be attenuated by the 'bag' nature of the concept. More importantly, most of the nominal lemmas included in the extended version of COW are human validated, so the quality of the concept based dictionary was confirmed to be better – since most lemmas included in it are attested to be valid.

Comparing the size of both dictionaries in Table 1, even though the $\tau 1$ lemma based dictionary is considerably larger (32.4k compared to 22.5k entries of the concept based dictionary), we have shown that noise is a problem for the lemma based approach. Also, since the extended COW has, on average, 2.25 senses per concept, the concept based dictionary provides CL information for over 50.6k lemmas. When comparing the size of both dictionaries across τs, we can also effectively verify the potential of the expansion step possible only for the concept based dictionary. As τ increases, the size of the concept based dictionary increases relatively to the lemma based. When applied to other tasks, where noise reduction would play a more important role (which can be done by raising τ), the concept based dictionary is able to produce more informed decisions with less data.

Lastly, coverage was also tested against data from a human curated database of noun-CL associations (Gao, 2014), by replicating the automatic evaluation generation task described in 4.3. This dictionary contains information about more than

800 CLs and provides a few hand-selected examples for each CL – and hence it is not designed with the same mindset. Testing the best performing dictionaries (τ1) against the data provided for S-CLs, we achieved only 43.9% and 28.3% for prediction and generation, respectively, using the lemma based dictionary; compared to 49.8% and 22.4% using the concept based dictionary.

The same trends in prediction and generation are observed, where the concept based dictionary is able to predict better than the lemma base, but it is outperformed by the later in the generation task. Ultimately, these weak results show that even though we used a very large quantity of data, our restrictive matching patterns in conjunction with infrequent noun-CLs pairs still leaves a long tail of difficult predictions.

6 Ongoing and Future Work

Since our method is mostly language independent, we would like to replicate it with other calssifier languages for which there are open linked WN resources (such as Japanese, Indonesian and Thai). This would require access to large amounts of text segmented, POS tagged text, and adapting the matching expressions for extracting noun-CL pairs.

More training data would not only help improving overall performance on open data, by minimizing unseen data, but would also allow us to make better use of frequency threshold filters for noise reduction. Lack of training data as our biggest drawback on performance, we would like to repeat this experiment with more data – including, for example, a very large web-crawled corpus in our experiments.

In addition, we would also like to perform WSD on the training set, using UKB (Agirre and Soroa, 2009) for example. This would allow an informed mapping of ambiguous senses onto the semantic ontology and, arguably, comparable performance on generating CLs for ambiguous lemmas. We will also investigate further how to deal with words not in COW: first looking them up in the lemma dictionary, and then associating CLs to the head (character / noun) of unseen noun-phrases, as proposed in Bond and Paik (2000).

Even though this work was mainly focused on producing an external resource linked to COW, we are also investigating adding a new set of sortal classifiers concepts to COW. The absence of this class of words in COW currently prevents us from using the internal ontology structure to link nouns and classifiers. Once classifiers are represented as concepts in this lexical ontology, we will make use of this work to link nominal concepts and corresponding valid classifiers.

7 Conclusions

Our work shows that it is possible to create a high quality dictionary of noun-CLs, with generation capabilities, by extracting frequency information from large corpora. We compared both a lemma based approach and a concept based approach, and our best results report a human validated performance of 87% on generation of classifiers using a concept based dictionary. This is roughly a 9% improvement against the only other known work done on Chinese CL generation using wordnet (Mok et al., 2012).

Finally, we will merge all three data sets and, from them, produce a release of this data. We commit to make both lemma and WN mappings available under an open license, release along with the Chinese Open Wordnet at `http://compling.hss.ntu.edu.sg/cow/`.

8 Acknowledgments

This research was supported in part by the MOE Tier 2 grant *That's what you meant: a Rich Representation for Manipulation of Meaning* (MOE ARC41/13).

References

Eneko Agirre and Aitor Soroa. 2009. Personalizing pagerank for word sense disambiguation. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, pages 33–41. Association for Computational Linguistics.

Francis Bond and Ryan Foster. 2013. Linking and extending an open multilingual wordnet. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 1352–1362, Sofia, Bulgaria. Association for Computational Linguistics.

Francis Bond and Kyonghee Paik. 2000. Reusing an ontology to generate numeral classifiers. In *COLING 2000 Volume 1: The 18th International Conference on Computational Linguistics*, pages 90–96.

Pi-Chuan Chang, Michel Galley, and Christopher D. Manning. 2008. Optimizing Chinese word segmentation for machine translation performance. In

Proceedings of the Third Workshop on Statistical Machine Translation, StatMT '08, pages 224–232, Stroudsburg, PA, USA. Association for Computational Linguistics.

Y.R. Chao. 1965. *A Grammar of Spoken Chinese*. University of California Press.

Lisa Lai-Shen Cheng and Rint Sybesma. 1998. Yi-wan tang, yi-ge tang: Classifiers and massifiers. *Tsing Hua journal of Chinese studies*, 28(3):385–412.

Mary S Erbaugh. 2002. Classifiers are for specification: Complementary functions for sortal and general classifiers in Cantonese and Mandarin. *Cahiers de linguistique-Asie orientale*, 31(1):33–69.

Christine Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. MIT Press.

Helena Gao. 2010. Computational lexicography: A feature-based approach in designing an e-dictionary of Chinese classifiers. In *Proceedings of the 2nd Workshop on Cognitive Aspects of the Lexicon*, pages 56–65. Coling 2010.

Helena Gao. 2014. Database design of an online e-learning tool of Chinese classifiers. In *Proceedings of the 4th Workshop on Cognitive Aspects of the Lexicon (CogALex)*, pages 126–137.

David Graff, Ke Chen, Junbo Kong, and Kazuaki Maeda. 2005. *Chinese Gigaword Second Edition LDC2005T14*. Web Download. Linguistic Data Consortium.

Chu-Ren Huang, Keh-Jiann Chen, and Ching-Hsiung Lai, editors. 1997. *Mandarin Daily Dictionary of Chinese Classifiers*. Mandarin Daily Press, Taipei.

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. *Quantitative and Computational Studies of Chinese Linguistics*, pages 339–352.

Chu-Ren Huang, Ru-Yng Chang, and Hshiang-Pin Lee. 2004. Sinica BOW (Bilingual Ontological Wordnet): Integration of bilingual wordnet and sumo. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, pages 825–826. European Language Resources Association (ELRA).

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

Hitoshi Isahara, Francis Bond, Kiyotaka Uchimoto, Masao Utiyama, and Kyoko Kanzaki. 2008. Development of the Japanese WordNet. In *Sixth International conference on Language Resources and Evaluation (LREC 2008)*, Marrakech.

Wan-chun Lai. 2011. Identifying True Classifiers in Mandarin Chinese. Master's thesis, National Chengchi University, Taiwan.

Jean-Baptiste Michel, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K Gray, Joseph P Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, Jon Orwant, et al. 2011. Quantitative analysis of culture using millions of digitized books. *Science 14 January 2011*, 331(6014):176–182.

Hazel Mok, Eshley Gao, and Francis Bond. 2012. Generating numeral classifiers in Chinese and Japanese. In *Proceedings of the 6th Global WordNet Conference (GWC 2012)*, Matsue. 211-218.

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

Virach Sornlertlamvanich, Wantanee Pantachat, and Surapant Meknavin. 1994. Classifier assignment by corpus-based approach. In *Proceedings of the 15th conference on Computational linguistics-Volume 1*, pages 556–561. Association for Computational Linguistics.

Liang Tian, Derek F. Wong, Lidia S. Chao, Paulo Quaresma, Francisco Oliveira, and Lu Yi. 2014. UM-Corpus: A large English-Chinese parallel corpus for statistical machine translation. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)*. European Language Resources Association (ELRA).

Kristina Toutanova, Dan Klein, Christopher D. Manning, and Yoram Singer. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the NAACL HLT 2003 2003 - Volume 1*, NAACL '03, pages 173–180, Stroudsburg, PA, USA. Association for Computational Linguistics.

Huihsin Tseng, Pichuan Chang, Galen Andrew, Daniel Jurafsky, and Christopher Manning. 2005. A conditional random field word segmenter for sighan bakeoff 2005. In *Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing*, pages 168–171.

Shan Wang and Francis Bond. 2013. Building the Chinese Open Wordnet (COW): Starting from core synsets. In *Proceedings of the 11th Workshop on Asian Language Resources, a Workshop at IJCNLP-2013*, pages 10–18, Nagoya.

Renjie Xu, Zhiqiang Gao, Yingji Pan, Yuzhong Qu, and Zhisheng Huang. 2008. An integrated approach for automatic construction of bilingual Chinese-English wordnet. In John Domingue and Chutiporn Anutariya, editors, *The Semantic Web*, volume 5367 of *Lecture Notes in Computer Science*, pages 302–314. Springer Berlin Heidelberg.