Detecting the Countability of English Compound Nouns Using Web-based Models

Jing Peng
Language Media Laboratory
Hokkaido University
Kita-14, Nishi-9, Kita-ku,
Sapporo, JAPAN
pj@media.eng.hokudai.ac.jp

Kenji Araki
Language Media Laboratory
Hokkaido University
Kita-14, Nishi-9, Kita-ku,
Sapporo, JAPAN
araki@media.eng.hokudai.ac.jp

Abstract

In this paper, we proposed an approach for detecting the countability of English compound nouns treating the web as a large corpus of words. We classified compound nouns into three classes: countable, uncountable, plural only. Our detecting algorithm is based on simple, viable n-gram models, whose parameters can be obtained using the WWW search engine Google. The detecting thresholds are optimized on the small training set. Finally we experimentally showed that our algorithm based on these simple models could perform the promising results with a precision of 89.2% on the total test set.

1 Introduction

In English, a noun can be countable or uncountable. Countable nouns can be "counted", they have a singular and plural form. For example: an apple, two apples, three apples. Uncountable nouns cannot be counted. This means they have only a singular form, such as water, rice, wine. Countability is the semantic property that determines whether a noun can occur in singular and plural forms. We can obtain the information about countability of individual nouns easily from grammar books or dictionaries. Several researchers have explored automatically learning the countability of English nouns (Bond and Vatikiotis-Bateson, 2002; Schwartz, 2002; Baldwin and Bond, 2003). However, all the proposed approaches focused on learning the countability of individual nouns.

A compound noun is a noun that is made up of two or more words. Most compound nouns in English are formed by nouns modified by other nouns or adjectives. In this paper, we concentrate solely on compound nouns made up of only two words, as they account for the vast majority of compound nouns. There are three forms of compound words: the closed form, in which the words are melded together, such as "*songwriter*", "*softball*", "*scoreboard*"; the hyphenated form, such as "*daughter-in-law*", "*master-at-arms*"; and the open form, such as "*post office*", "*real estate*", "*middle class*".

Compound words create special problems when we need to know their countability. According to "*Guide to English Grammar and Writing*", the base element within the compound noun will generally function as a regular noun for the countability, such as "*Bedrooms*". However this rule is highly irregular. Some uncountable nouns occur in their plural forms within compound nouns, such as "*mineral waters*" (water is usually considered as uncountable noun). The countability of some words changes when occur in different compound nouns. "*Rag*" is countable noun, while "*kentish rag*" is uncountable; "*glad rags*" is plural only. "Wages" is plural only, but "*absolute wage*" and "*standard wage*" are countable. So it is obvious that determining countability of a compound noun should take all its elements into account, not consider solely on the base word.

The number of compound nouns is so large that it is impossible to collect all of them in one

dictionary, which also need to be updated frequently, for newcoined words are being created continuously, and most of them are compound nouns, such as "*leisure sickness*", "*Green famine*".

Knowledge of countability of compound nouns is very important in English text generation. The research is motivated by our project: post-edit translation candidates in machine translation. In Baldwin and Bond (2003), they also mentioned that many languages, such as Chinese and Japanese, do not mark countability, so how to determine the appropriate form of translation candidates is depend on the knowledge of countability. For example, the correct translation for "发育性痛[1]" is "growing pains", not "growing pain".

In this paper, we learn the countability of English compound nouns using WWW as a large corpus. For many compound nouns, especially the relatively new words, such as *genetic pollution*, have not yet reached any dictionaries. we believe that using the web-scale data can be a viable alternative to avoid the sparseness problem from smaller corpora. We classified compound nouns into three classes: countable (eg., bedroom), uncountable (eg,. cash money), plural only (eg,. crocodile tears). To detect which class a compound noun is, we proposed some simple, viable n-gram models, such as freq(N) (the frequency of the singular form of the noun) whose parameters' values (web hits of literal queries) can be obtained with the help of WWW search engine Google. The detecting thresholds (a noun whose value of parameter is above the threshold is considered as plural only) are estimated on the small countability-tagged training set. Finally we evaluated our detecting approach on a test set and showed that our algorithm based on the simple models performed the promising results.

Querying in WWW adds noise to the data, we certainly lose some precision compared to supervised statistical models, but we assume that the size of the WWW will compensate the rough queries. Keller and Lapata (2003) also showed the evidence of the reliability of the web counts for natural language processing. In (Lapata and Keller, 2005), they also investigated the countability leaning task for nouns. However they only distinguish between countable and uncountable for individual nouns. The best model is the determiner-noun model, which achieves 88.62% on countable and 91.53% on uncountable nouns.

In section 2 of the paper, we describe The main approach used in the paper. The preparation of the training and test data is introduced in section 3. The details of the experiments and results are presented in section 4. Finally, in section 5 we list our conclusions.

2 Our approach

We classified compound nouns into three classes, countable, uncountable and plural only. In Baldwin and Bond (2003), they classified individual nouns into four possible classes. Besides the classes mentioned above, they also considered *bipartite* nouns. These words can only be plural when they head a noun phrase (trousers), but singular when used as a modifier (trouser leg). We did not take this class into account in the paper, for the *bipartite* words is very few in compound nouns.

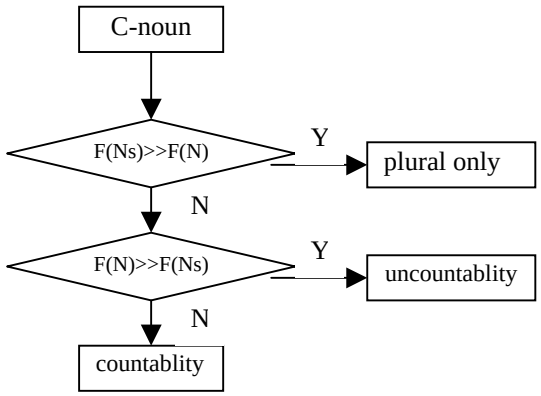

Figure 1. Detecting processing flow

For plural only compound noun, we assume that the frequency of the word occurrence in the plural form is much larger than that in the singular form, while for the uncountable noun, the frequency in the singular form is much larger than that in the plural form. The main processing flow is shown in Figure 1. In the figure, "C-noun" and "*Ns*" mean compound noun and the plural form of the word respectively. "*F(Ns)>>F(N)*" means that the frequency of the plural form of the noun is much larger than that of the singular form.

[1] "发育性痛"(fa yu xing tong) which is Chinese compound noun means "growing pains".

Our approach for detecting countability is based on some simple unsupervised models.

$$\frac{f(Ns)}{f(N)} \geq \theta \qquad (1)$$

In (1), we use the frequency of a word in the plural form against that in the singular form. θ is the detecting threshold above which the word can be considered as a plural only.

$$\frac{f(much, N)}{f(many, Ns)} \geq \theta \quad (2)$$

In (2), we use the frequency of a word in the singular form co-occurring with the determiner "*much*" against the frequency of the word in the plural form with *many*, if above θ, the word can be considered as uncountable word. (2) is used to distinguish between countable and uncountable compound nouns.

$$\frac{f(Ns, are)}{f(N, is)} \geq \theta \qquad (3)$$

The model 3 that compares the frequencies of noun-be pairs (eg,. f("account books are"), f("account book is") is used to distinguish plural only and countable compound nouns.

With the help of WWW search engine Google, the frequencies (web hits) in the models can be obtained using quoted n-gram queries ("soft surroundings"). Although in Keller and Lapata (2002), they experimentally showed that web-based approach can overcome data sparseness for bigrams, but the problem still exists in our experiments. When the number of pages found is zero, we smooth zero hits by adding them to 0.01.

Countable compound nouns create some problems when we need to pluralize them. For no real rules exist for how to pluralize all the words, we summarized from "*Guide to English Grammar and Writing*" for some trends. We processed our experimental data following the rules below.

1. Pluralize the last word of the compound noun. Eg,. *bedrooms, film stars*.
2. When "*woman*" or "*man*" are the modifiers in the compound noun, pluralize both of the words. Eg,. *Women-drivers*.
3. When the compound noun is made up as "noun + preposition (or prep. phrase)", pluralize the noun. Eg,. *fathers-in-law*.
4. When the compound noun is made up as "verb (or past participle) + adverb", pluralize the last word. Eg,. *grown-ups, stand-bys.*

Although the rules cannot adapt for each compound noun, in our experimental data, all the countable compound nouns follow the rules. We are sure that the rules are viable for most countable compound nouns.

Although we used Google as our search engine, we did not use Google Web API service for programme realization, for Google limits to 1000 automated queries per day. As we just need web hits returned for each search query, we extracted the numbers of hits from the web pages found directly.

3 Experimental Data

The main experimental data is from Webster's New International Dictionary (Second Edition). The list of compound words of the dictionary is available in the Internet[2]. We selected the compound words randomly from the list and keep the nouns, for the word list also mixes compound verbs and adjectives with nouns together. We repeated the process several times until got our experimental data. We collected 3000 words for the training which is prepared for optimizing the detecting thresholds, and 500 words for the test set which is used to evaluate our approach. In the sets we added 180 newcoined compound nouns (150 for training; 30 for test). These relatively new words that were created over the past seven years have not yet reached any dictionaries[3].

Countability	Training set	Test set
Plural only	80	21
Countable	2154	352
Uncountable	766	127
Total	3000	500

Table 1. The make-up of the experimental data

We manually annotated the countability of these compound nouns, plural only, countable, uncountable. An English teacher who is a native speaker has checked and corrected the annotations. The make-up of the experimental data is listed in Table 1.

[2] The compound word list is available from http://www.puzzlers.org/wordlists/dictinfo.php.

[3] The new words used in the paper can be found in http://www.worldwidewords.org/genindex-pz.htm

4 Experiments and Results

4.1 Detecting plural only compound nouns

Plural only compound nouns that have not singular forms always occur in plural forms. The frequency of their singular forms should be zero. Considering the noise data introduced by search engine, we used model (1) and (3) in turn to detect plural noun. We detected plural only compound nouns with the following algorithm (Figure 2), which is used to distinguish between plural only and non-plural only compound.

if $(\frac{f(Ns)}{f(N)} \geq \theta 1)$
then plural only;
else if $(\frac{f(Ns, are)}{f(N, is)} \geq \theta 2)$
then plural only;
else
countable or uncountable;

Figure 2. Detecting algorithm for plural only

The problem is how to decide the two thresholds. We preformed exhaustive search to adjust $\theta 1, \theta 2$ optimized on the training set. With $0 \leq \theta 1, \theta 2 \leq 20$, all possible pair values are tried with the stepsize of 1.

$$\mathrm{Re}\,call = \frac{A}{AB} \quad (4)$$

$$\mathrm{Pr}\,ecision = \frac{A}{AC} \quad (5)$$

$$F - score = \frac{2 \times \mathrm{Pr}\,ecision \times \mathrm{Re}\,call}{\mathrm{Pr}\,ecision + \mathrm{Re}\,call} \quad (6)$$

We use Recall and Precision to evaluate the performance with the different threshold pairs. The fundamental Recall/Precision definition is adapted to IE system evaluation. We borrowed the measures using the following definition for our evaluation. For one experiment with a certain threshold pair, *A* stands for the number of plural found correctly; *AB* stands for the total number of plural only compound nouns in training set (80 words); *AC* stands for the total number of compound nouns found. The Recall and Precision are defined in (4) and (5). We also introduced F-score when we need consider the Recall and Precision at the same time, and in the paper, F-score is calculated according to (6).

Figure 3 shows the performance evaluated by the three measures when $\theta 1$=8 and $0 \leq \theta 2 \leq 10$ with a stepsize of 1. We set $\theta 2$ to 5 for the test later, and accordingly the values of Recall, Precision and F-score are 91.25%, 82.95% and 87.40% respectively.

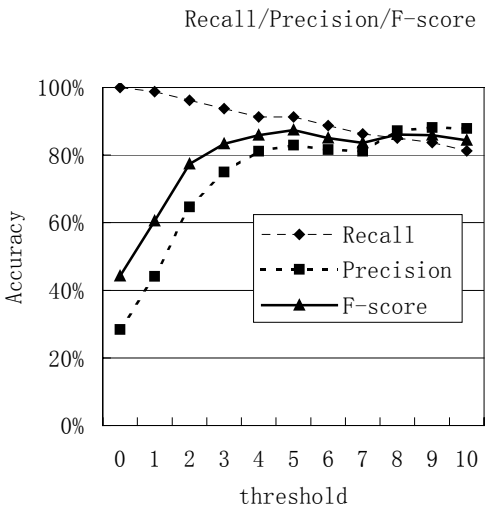

Figure 3. The Recall/Precision/F-score graph ($\theta 1$=8 and $0 \leq \theta 2 \leq 10$)

4.2 Detecting uncountable compound nouns

Uncountable compound nouns that have not plural form always occur in singular form.

if (N is not plural only)
then if $(\frac{f(N)}{f(Ns)} \geq \theta 3)$
then uncountable;
else if $(\frac{f(much, N)}{f(many, Ns)} \geq \theta 4)$
then uncountable;
else
countable;

Figure 4. Detecting algorithm for uncountable compound nouns

The algorithm detecting uncountable compound nouns is shown in Figure 4. Using model (1) and (2), we attempted to fully make use of the characteristic of uncountable compound nouns, that is the frequencies of their occurrence in the singular forms are much larger than that in the plural forms.

The method to obtain the optimal threshold $\theta 3$ and $\theta 4$ is the same to 4.1. We set $\theta 3$ to 24, $\theta 4$ to 2, and the values of Recall, Precision and F-score are 88.38%, 80.27% and 84.13% respectively.

4.3 Performance on the test suite

We evaluated our complete algorithm with the four thresholds ($\theta 1$=8, $\theta 2$=5, $\theta 3$=24, $\theta 4$=2) on the test set, and the detecting results are summarized in Table 2. There are 352 countable compound nouns in our test set, then when classify all the test words as countable, we can at least get the accuracy of 70.4%. We used it as our baseline. The accuracy on the total test date is 89.2% that significantly outperforms the baseline. For the 30 newcoined compound nouns, the detecting accuracy is 100%. This can be explained by their infrequence. Newcoined words are not prone to produce noise data than others just because they are not occurring regularly.

	Correct	Incorrect	Recall	Precision	F-score
Plural only	18	4	85.71%	81.81%	83.71%
Countable	320	22	90.90%	93.57%	92.22%
Uncountable	108	28	85.04%	79.41%	82.15%
Total	446	54	89.2%	89.2%	89.2%

Table 2. The accuracy on the test suit

5 Conclusion

From the results, we show that simple unsupervised web-based models can achieve the promising results on the test data. For we roughly adjusted the threshold with stepsize of 1, better performance is expected with stepsize of such as 0.1.

It is unreasonable to compare the detecting results of individual and compound nouns with each other since using web-based models, compound nouns made up of two or more words are more likely to be affected by data sparseness, while individual nouns are prone to produce more noise data because of their high occurrence frequencies.

Anyway using WWW is an exciting direction for NLP, how to eliminate noise data is the key to improve web-based methods. Our next step is aiming at evaluating the internet resource, distinguishing the useful and noise data.

References

Baldwin, Timothy and Francis Bond.2003. Learning the countability of English nouns from corpus data. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics.* Sapporo, Japan, 463-470.

Francis Bond and Caitlin Vatikiotis-Basteson. 202. Using an ontology to determine English countability. In *Proceeding of the 19th International conference on computational Linguistics (COLING 2002)*, Taipei, Taiwan.

Keller, F, Lapata, M. and Ourioupina, O. 2002. Using the web to overcome data sparseness. In Proceedings of the Conference on Empirical Methods in Natural Language Processing. Philadelphia. 230-237.

Keller, F and Lapata, M. 2003. Using the web to obtain frequencies for unseen *bigrams. Computational Linguistics* 29, 3, 459-484.

Lapata, M and Keller, F. 2004. The web as a baseline: Evaluating the performance of unsupervised web-based models for a range of NLP tasks. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*. Boston,

Lapata, M and Keller, F. Web-based Models for Natural Language Processing. To appear in 2005 *ACM Transactions on Speech and Language Processing*.

Lane O.B. Schwartz. 2002. Corpus-based acquisition of head noun countability features. Master's thesis, Cambridge University, Cambridge, UK.

Guide to English Grammar and Writing. http:// cctc. commnet.edu/grammar/