Distributional Consistency:

As A General Method for Defining A Core Lexicon

Huarui ZHANG*   Churen HUANG**   Shiwen YU*

*Institute of Computational Linguistics, Peking University

Beijing, 100871, P.R.China

{hrzhang, yusw}@pku.edu.cn

**Institute of Linguistics, Academia Sinica,

churen@sinica.edu.tw

Abstract

We propose Distributional Consistency (DC) as a general method for defining a Core Lexicon. The property of DC is investigated theoretically and empirically, showing that it is clearly distinguishable from word frequency and range of distribution. DC is also shown to reflect intuitive interpretations, especially when its value is close to 1. Its immediate application in NLP would include defining a core lexicon in a language and identifying topical words in a document. We also categorize the existent measures of dispersion into 3 groups via ratio of norm or entropy, proposed a simplified measure and a combined kind of measure. These new measures can be used as virtual prototype or medium type for the study and comparison of existent measures in the future.

Keywords: Distributional Consistency; Lexical Usuality; Measure of Dispersion; Square Mean Root (SMR); Modified Frequency; Core Lexicon

1. Introduction and Definition

Defining a core lexicon is a central issue in computational lexicography, psycholinguistics, and language pedagogy. Frequency and Semantic primitives are two most often used criteria. However, these two criteria do not define the same set of lexical items. Neither is there a clear rationale to resolve the discrepancy. In this paper, we propose a measurement that captures the intuitions that previous criteria try to capture. The measure, call Distributional Consistency, is also well-motivated mathematically.

In previous research, the most commonly used words are determined by word occurrence frequency, but frequency is heavily dependent on the corpus selected. If the corpus is not large enough, or not balanced, the result will be not accurate enough. The criteria for judging whether a corpus is balanced are difficult to determine because the purpose of the corpus may vary. The criteria are even more important because building a large corpus would entail expenditure of a large amount of economic and human resources.

However, we have another consideration: if a word is commonly used in a language, it will appear in different parts of a corpus. And if the word is used commonly enough, it will be well-distributed. This constitutes the foundation of **Distributional Consistency (DC)**.

We now propose a metric for distribution of words in a corpus, as follows:

$$DC=((\Sigma f_i^{1/2})/n)^2/((\Sigma f_i)/n)$$

where

DC: the Distributional Consistency of a specific word

f_i: the occurrence frequency of the specified word in the ith part of the corpus

n: the number of equally sized parts into which the corpus is divided

Σ: the sum of

When a corpus is divided into parts with different size, the above formula should be normalized with respect to a size factor. We have had three methods tested, and one of the three is found to be most reasonable and practical so that implemented below.

2. Property of Distributional Consistency

It can be proven that the possible value range of DC is from 0 to 1. The more consistent in distribution, the closer the value is to 1. The minimal value of 0 is unreachable, but can be assigned to the words that do not occur in the corpus at all, as done in the same way for frequency. This is a reasonable extension of the definition that allows differentiation between the unattested words and the infrequent/rare words.

Given the condition that the corpus is divided into *n* equally sized parts:

(1) if a word occurs in only one part, the DC of the word is 1/n;

(2) if a word occurs in every part with the same frequency, the DC of the word is 1;

In pure formal terms, there is a possibility that a word is not so commonly used but simply occurs one time in every corpus partition hence receives the highest DC value of 1. Even though this is a statistical possibility, we would hope that it could hardly be a linguistic possibility. After investigating more than 10 millions tokens involving 150,000 word types, not even a single case was found.

The above cases are boundaries. When a word occurs in *m* parts (1<m<=n) and with different frequencies in different parts, the DC is less than m/n.

3. Distributional Consistency vs. Word Frequency

It is crucial to differentiate DC from frequency. After investigating the DCs computed from the POS-tagged Chinese Corpus of *People's Daily* (Beijing, China, year 1998, about 26M Chinese characters)

Order the DCs of the whole lexicon and give each of them a rank, called DC rank; and order the frequencies of them, as frequency rank. We check the both rank ranges to see if there is a correlation. The following table shows the overlap between the two rank ranges:

Table 1. DC and Frequency Correlation

r-DC	r-freq	n-word
10	10	3
100	100	44
1000	1000	562
10000	10000	7359

where

    r-DC: ranking of words according to DC

    r-freq: ranking of words according to word frequency

    n-word: the number of words which fall in the same DC rank and frequency rank range

We see that, for instance, from rank 100 to rank 1000, about only one half overlaps. There is a significant difference between DC and word frequency ranks.

The basic soundness of DC can be attested by the functional elements that can be reliably predicted to distribute consistently over all parts of a corpus. We look at a) most frequently used punctuation and 2) *de*: the most frequently used word.

In Chinese, comma is the most frequently used punctuation, and its distribution is also quite consistent. This is partly due to the fact that the use of period to end a sentence is not very conventionalized. When punctuations are included in the computing the lexical DCs of Chinese, comma has one of the highest DC. This is expected with a straightforward interpretation.

On the other hand, the frequency of the most frequent Chinese word *de* is so dominant that it is usually at least 2 or 3 times more frequent than the word ranked second. However, when our DC rank is given, *de* loses its dominance. It is still among the highest dozen but has no obvious superiority. This also clearly indicates the different implications and interpretations of DC and word frequency.

4. Distributional Properties of DC

Another important characteristic of DC is its inter-dependency with the number of partition of a corpus. We can define the **Range** of a word as the number of partitions of a corpus that it occurs in. In other words, a word with a higher number of range is distributed more widely. When a word's distribution is totally balanced and it occurs in equal number of times in each part, there is a direct correlation between range and DC. In this case, as shown in Section 2, if number of parts that it occurs is m, then the DC of this word is m/n. This serves as the upper limit of DC for a word with range m. If the occurrence number in different parts is not all equal, then the DC is less than m/n. This is the theoretical relation between DC and range.

On the other hand, we could also test the theoretical lower bound of the DC of word by assuming that the distribution of a word is unbalanced to the extreme, i.e. that it occurs only once in all but one of the parts. In other words, when it occurs once only in m-1 parts, and occurs f-m+1 times in one of the parts, where f is its frequency in the corpus. When f is a big enough number, then the contribution of the m-1 parts can be minimal and the DC of that word will approach 1/n, where n is the number of parts. In other words, for all words, regardless of it range, its minimal value will be 1/n.

Now that we have shown that the higher bound of the DC of a word is dependent on its range, while the lower bound is not, it is interesting to see the actual distribution. The following is the empirical result when we divided the whole year of People's Daily corpus into 12 parts by the 12 months of the year. (The value has been adjusted by the size of each month's corpus). The range of each word is from 1 to 12. And the DC, as predicted, is roughly from 1/12 to 1. Although this appears to indicate a linear relation between range and DC, the appearance is actually misleading. There are plenty of instances where the range of a word is high, but the DC is rather low. The reason is that the distribution in different part is too diverse and not balanced. For those words whose range are 12(the highest one, which means that they occur in every month of the year), the DC vary from near 1 to below 0.5. It is interesting to show that this actual lower bound is significantly higher than the theoretical lower bound of nearly 0.08 (about 1/12). The empirical lower limit is not only higher than the absolute uniform lower limit, it also varies with the increase of the range number. The tendency is positive related, which means that when the number increase, the real lower limit is also relative high.

The following figures show the relation between DC and range. Only those whose range is above 3 are shown.
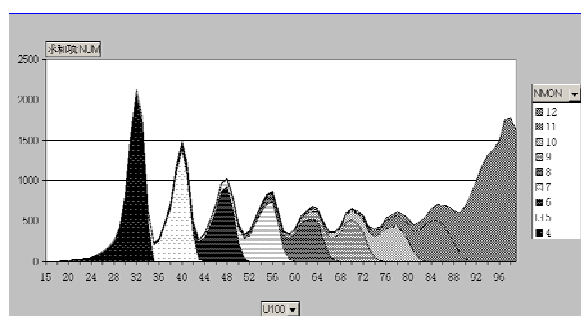

Figure 1: Decomposition of Range Components
(x: percentage of DC;
y: number of words whose DC is x in percentage (integer);
pattern: to show the range, from left to right, 4 to 12.)

Figure 2 below shows that the range peaks have more elongated and graduated left tail extending and approaching the shared absolute minimal value. This is in contrast with what is shown in figure 3 for the right foot of the range peaks. The right feet have steep slopes and reach zero abruptly, while the left tail stretches over a wide value range before ending.
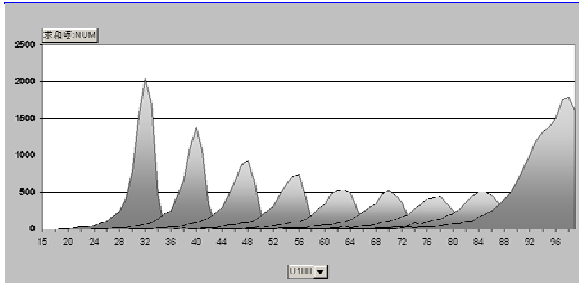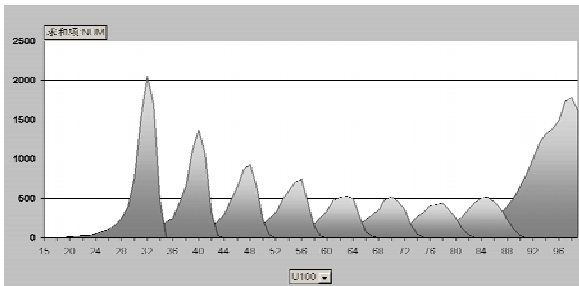
Figure 2: Left Sides of Range Peaks

Figure 3: Right Sides of Range Peaks

Figure 3 also shows that the slope of the right feet become smaller when the range increases. Which means, that if a word does not occur in every part, then the local balance of distribution even becomes a unnatural case when the number of range increases because of its paradox between global and local distributions. In other words, when there is a small number of parts for the word to appear in (say 2 or 3), it is easier to have a balanced distribution among the parts. But when there are many more parts to appear in, it become harder to have even distribution.
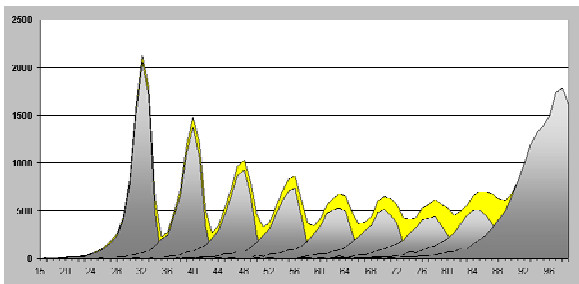

Figure 4: Comparison

Figure 4 shows that the intersection between adjacent peaks increases with the range, both in relative and absolute value. However, it gradually decreases down to zero where the right foot of the range-11 peak reaches zero. After that point, it is totally the contribution of the range-12 peak (i.e. total range peak).

There is also an important fact that should be obvious from all four figures. That is, except for the total range peak, peak size diminishes as range number increases. Indeed, the reason why range peaks 1 to 3 are not shown is for the practical reason that their sizes would have make the presentation of the lowers peaks difficult.

This distributional fact can be predicted with intuitive interpretation of DC. We can interpret words with range m as the words that stratify conditions (such as topicality, or temporal related activity) to appear in m parts of the corpus. In this interpretation, it is easier to satisfy a smaller number of sets of conditions. That is, the interpretation predicts that more words will appear in an i-1 range than in an i range. This prediction is largely born out except for the relation between the total peak and its next peak. In other words, the distribution of the total peak cannot be predicted with the above interpretation.

Our proposal is that this anomaly is exactly the consequence of a core lexicon. In Huang et al. (2004), we try to define a core lexicon as the maximal necessary lexicon in a language. In other words, a core lexicon is the (largest possible) set of words that can be expected to be used regardless of the environments. If each language has a core lexicon, then they can be expected to appear in the total peak. Intuitively, this means that the core lexicon is consisted of words that are less constrained by the conditions of use predicted by the partition of corpus. In this interpretation, the main part to the right of range-11 peak (in this case, for those words whose DC is above 92%), is of special significance for selection of most common words in a language. The number of words thus selected in this study is about ten thousand

5. Formal Properties of DC: comparative studies and future developments

From the 1960 on, there had been several studies on the dispersion measure of words in a corpus partitioned to a number of divisions. These studies can be classified into the following 3 groups:

(1) n_2/n_1 related
   e.g. Juilland (1964) $\{ (n-1)(1-D)^2=(n_2/n_1)^2-1 \}$

(2) S/S_{max} related
   e.g. Carroll (1970), Kromer (2003) (in the sense of using sum of reciprocal as discrete natural logarithm)

(3) $n_{1/2}/n_1$ related
   e.g. Rosengren (1971), (Yin(1994)'s t-degree frequency is also related, in fact, which is just (1/t)-norm.)

Where
   n_1 is 1-norm
   n_2 is 2-norm
   $n_{1/2}$ is 1/2-norm
      While p-norm is defined as
      $(\Sigma(f_i^p)/n)^{1/p}$
   S is Shannon Entropy
   S_{max} is the maximum Entropy whatever the relative size of the divisions
      While Shannon Entropy was defined as
      $[-]\Sigma (p_i log(p_i))$.

Two notes can be made regarding the above classification. First, general speaking, the division by n (number of divisions) is not considered in these studies. It is in our study and does not affect the relative value of different words in the same corpus. Second, for entropy, a negative sign is added to return a positive value. However, whether there is such a negative sign does not matter because the relevant final value is the ratio of two samples calculated by this definition.

As we see, most of these measures are used to get modified frequency (or "corrected frequency", "adjusted frequency") in order to assign better ranking significance regarding the usage/importance of the words. Various dictionaries or word frequency lists have been produced via these modified frequencies.

Our proposed DC falls in the 3rd group. However, what we proposed DC not to adjust frequency, but to obtain a relatively independent index (or quantity) comparable with frequency. This is somewhat like the discrete range of appearances, yet it is continuous and takes the distribution in different subsets into consideration. We showed that there are real mismatches between DC and range. We have shown, for example, that words have the highest range of 12 (namely, 100%) do have a DC as low as 0.5.

In order to combine and compare the three different approaches towards lexical computing and ranking, We adopt the simplest one of the 1st group of definition:

$$\mathbf{n_1/n_2} \quad \text{or} \quad (\mathbf{n_1/n_2})^2$$

We consider that the property of this measure will reflect the characteristic feature of the 1st group (for example, Juilland measure) because they share the same (and the only) kernel factor (namely, n_2/n_1) and can be taken as simplified type for the study.

We also adopt a combined kind:

$$\mathbf{n_{1/2}/n_2} \text{ related,}$$

possible instances:

$$(\mathbf{n_{1/2}/n_2})^{1/2} \quad \text{or} \quad 1\text{-}(1\text{-}\mathbf{n_{1/2}/n_2})^{1/2}$$

We consider that the property of this measure will lie between the 1st group and the 3rd group of measure. Thus there will be a graduation from the 1st group, through this combined kind, to the 3rd group. Future studies on this direction will focus both on their formal properties and their interpretations.

6. Possible applications

The Distributional Consistency of words can be used for
(1) word selection in language teaching and textbook writing and dictionary compilation,
(2) topic words detection in a document by investigating the distributional consistency inside that document,
(3) author signature, works signature(specific, the age of, the area of, etc.), and
(4) empirical linguistics, such as in lexicostatistics.

In fact, we have made the first attempt to apply DC to the prediction and verification of the basic lexicon, represented by the Swadesh list in previous studies (Huang et al. 2004). It is shown that unlike word frequency, the words clustered by DC are often conceptually driven.

7. Conclusion

Distributional Consistency is a general measurement of word distribution in a corpus and it is distinguishable from word frequency and occurrence range. We also categorize the existent measures of dispersion into 3 groups by a uniform expression of the kernel factor via ratio of norm or entropy, proposed a simplified measure of the 1st group and a combined kind between the 1st group and the 3rd group of measure. These new measures can be used as virtual prototype or middle type for the study and comparison of dispersion measures in the future.

Last, DC can be applied to any partitionable data set. We suggest to refer to the properties captured by DC as the **usuality** of an element. This borrows the concept, though not the formal definition, of usuality in Zadeh (1985), as the 'usual' value of a variable or an event. We have used a corpus to calculate the **lexical usuality** based on months. Other types of usuality as well as usuality based on different partition criteria will present challenging and rewarding future studies.

References (in chronological order)

Juilland A and Chang-Rodriguez E (1964). Frequency dictionary of Spanish words. The Hague, Mouton.

Juilland A, Brodin D and Davidovitch C (1970). Frequency dictionary of French words. Paris, Mouton.

Carroll, J.B. (1970). An alternative to Juilland's usage coefficient for lexical frequencies and a proposal for a Standard Frequency Index (SFI). Computer Studies in the Humanities and Verbal Behavior, 3, 61–65.

Carroll J B, Davies P and Richman B (1971). The American Heritage word frequency book. Boston, Houghton Mifflin

Rosengren, I. (1971). The quantitative concept of language and its relation to the structure of frequency dictionaries, Etudes de Linguistique Applique, 1:103-127.

Francis, W. Nelson, and Henry Kucera (1982). Frequency Analysis of English Usage: Lexicon and Grammar, Boston: Houghton Mifflin Company.

Zadeh, L.A. (1985). Syllogistic Reasoning in Fuzzy Logic and its Application to Usuality and Reasoning with Dispositions. IEEE Transactions on Systems, Man, and Cybernetics, SMC-15, 6, 754-763.

Yin, Binyong, and Shizeng Fang, (1994). Word frequency counting: A new concept and a new approach, Applied Linguistics (Yuyan Wenzi Yingyong), No. 2, pp. 69-75

Brown, Nicholas J. (1996). Russian Learners' Dictionary: 10,000 words in frequency order. London and New York: Routledge. pp. 429.

Yu Shiwen et al, (2002). The Grammatical Knowledge-base of Contemporary Chinese ---A Complete Specification (The 2nd Version), Tsinghua University Press,

Savický, Petr and Jaroslava Hlavácová, (2002). Measures of Word Commonness, Journal of Quantitative Linguistics, Vol.9, No.3, pp. 215-231

Kromer, Victor (2003). An Usage Measure Based on Psychophysical Relations, Journal of Quantitative Linguistics, Volume 10 Number 2

Huang, Chu-Ren, Huarui Zhang, and Shiwen Yu. (2004). On Predicting and Verifying a Basic Lexicon: Proposals inspired by Distributional Consistency. [In Chinese]. To appear in D.-A. Ho Ed. Papers from POLA2003. Taipei: Language and Linguistics. Academia Sinica.