Two-Character Chinese Word Extraction Based on Hybrid of Internal and Contextual Measures

Shengfen Luo, Maosong Sun
National Lab. of Intelligent Tech. & Systems
Tsinghua University, Beijing 100084, China
lkc-dcs@mail.tsinghua.edu.cn

Abstract

Word extraction is one of the important tasks in text information processing. There are mainly two kinds of statistic-based measures for word extraction: the internal measure and the contextual measure. This paper discusses these two kinds of measures for Chinese word extraction. First, nine widely adopted internal measures are tested and compared on individual basis. Then various schemes of combining these measures are tried so as to improve the performance. Finally, the left/right entropy is integrated to see the effect of contextual measures. Genetic algorithm is explored to automatically adjust the weights of combination and thresholds. Experiments focusing on two-character Chinese word extraction show a promising result: the F-measure of mutual information, the most powerful internal measure, is 57.82%, whereas the best combination scheme of internal measures achieves the F-measure of 59.87%. With the integration of the contextual measure, the word extraction achieves the F-measure of 68.48% at last.

1 Introduction

New words are generated quite often with the rapid development of Chinese society, resulting that the lexicon of Chinese cannot well meet the requirement of natural language processing. How to extract word automatically from immense text collection has thus become an important problem.

The task of extracting Chinese words with multi-characters from texts is quite similar to that of extracting phrases (e.g., compound nouns) in English, if we regard Chinese characters as English words. Research in word/phrase extraction has been carried out extensively. Currently the mainstream approach is statistic-based. In general, there are two kinds of statistic-based measures for estimating the soundness of an extracted item being a word/phrase: One is the internal measure, which estimates the soundness by the internal associative strength between constituents of the item. Nine widely adopted internal measures are listed in (Schone et al. 2001), including Frequency, Mutual Information, Selectional Association, Symmetric Conditional Probability, Dice Formula, Log-likelihood, Chi-squared, Z-score, Student's t-score. The other kind is the contextual measure, which estimates the soundness by the dependency of the item on its context, such as the left/right entropy (Sornlertlamvanich et al. 2000), and the left/right context dependency (Chien 1999).

This paper firstly analyzes nine internal measures mentioned above, tests and compares their word extraction performance on individual basis, then tries to improve the performance by properly combining these measures. Furthermore, the contextual measure is integrated with internal measures to acquire more improvement. Throughout the experiments, genetic algorithm is explored to adjust weights of combination and thresholds automatically. We only concern two-character word extraction in this paper, because two-character words reflect the most popular word-formation of Chinese and possess the largest proportion in Chinese lexicon.

2 Internal Measures

Nine internal measures are discussed and compared in this section. These measures tend to estimate the internal associative strength from different perspectives, so it is possible to improve the word extraction performance by properly combining them. This paper will try two combination schemes, i.e., direct combination and interval-based combination.

As stated earlier, the evaluation is based on two-character Chinese word extraction. PDR9596, a raw corpus of People Daily of 1995 and 1996 with about 50.0M characters, is used to train the matrix of Chinese character bigrams throughout the experiments. PDA98J, a manually word-segmented corpus composed of People Daily of January 1998 with about 1.3M characters (developed by Institute of Computational Linguistics of Peking University), is further used to exhaustively generate a list of Chinese character bigrams. The list contains 218,863 distinct bigrams. We randomly divide the list into two parts, 9/10 of it as TS1, the rest 1/10 as TS2.

2.1 Nine Widely Adopted Internal Measures

Table 1 lists nine widely adopted internal measures, as mentioned in (Schone 2000). In the table: xy represents any two-character item, $\bar{x}$ stands for all characters except x, N is the size of training corpus, f_x and p_x are frequency and probability of x respectively, f_{xy} and p_{xy} are frequency and probability of xy respectively, and ξ_{xy} is the frequency expectation of xy suppose x and y are independent.

Obviously:

$$\xi_{xy} = p_{xy}N = p_x p_y N = f_x f_y / N$$

Table 1. Nine Widely Adopted Internal Measures

Measure	Marked As	Formula
Frequency	Freq	f_{xy}
Mutual Information	MI	$\log_2 \frac{p_{xy}}{p_x p_y}$
Selectional Association	SA	$\frac{p(x\mid y)MI(xy)}{\sum_z p(z\mid y)MI(zy)}$
Symmetric Conditional Probability	SCP	$\frac{p_{xy}^2}{p_x p_y}$
Dice Formula	Dice	$\frac{2f_{xy}}{f_x + f_y}$
Log-likelihood	LogL	$-2\log \frac{(p_x p_y p_{\bar{x}} p_{\bar{y}})^{f_y}}{(p_{xy} p_{\overline{xy}})^{f_{xy}} (p_{x\bar{y}} p_{\bar{x}y})^{f_{\bar{x}y}}}$
Chi-squared	Chi	$\frac{N(f_{xy} f_{\overline{xy}} - f_{x\bar{y}} f_{\bar{x}y})^2}{(f_{xy} + f_{x\bar{y}})(f_{xy} + f_{\bar{x}y})(f_{\overline{xy}} + f_{x\bar{y}})(f_{\overline{xy}} + f_{\bar{x}y})}$
Z-Score	ZS	$\frac{f_{xy} - \xi_{xy}}{\sqrt{\xi_{xy}(1-\xi_{xy}/N)}}$
Student’s t-Score	TS	$\frac{f_{xy} - \xi_{xy}}{\sqrt{f_{xy}(1-f_{xy}/N)}}$

Table 2. Comparison of Word Extraction Performance of Internal Measures (Open Test on TS1)

Top 17,333	Freq	MI	SA	SCP	Dice	LogL	Chi	ZS	TS
F-measure (%)	26.28	**54.77**	42.98	51.77	49.37	43.13	52.97	53.20	39.12
Comparison of F-measure: MI > ZS > Chi > SCP >Dice > LogL > SA > TS > Freq									

Table 3. Weights for Direct Combination Scheme (Fitness Function is Based on TS1)

Freq	MI	SA	SCP	Dice	LogL	Chi	ZS	TS
0.000598	0.351393	0.000263	0.146348	0.214541	0.002804	0.035930	0.072293	0.17583
Comparison of weights: MI > Dice > TS > SCP > ZS > Chi > LogL > Freq > SA								

When using these nine measures for word extraction, the hypothesis is same: the larger value of measure means the stronger associative strength between *x* and *y*, and thus the more possibility of xy being a word. The criterion of judgment is very simple: xy would be accepted as a word if its internal associative strength is larger than a given threshold.

2.2 Word Extraction Performance of Each Internal Measure

The performance of each internal measure is tested on TS1. TS1 contains 196,977 distinct bigrams, among which 17,333 are two-character words according to PDA98J. The procedure of word extraction is to sort the 196,977 candidate bigrams in descending order in terms of the value of the measure to be tested, and then to select the top 17,333 bigrams as words. In this case, the precision rate, recall rate and F-measure are exactly the same. Table 2 shows the comparison of performances of these nine internal measures. Mutual information achieves the best performance with the F-measure of 54.77%.

2.3 Direct Combination of Internal Measures

The first combination scheme is to directly combine the nine measures with appropriate weights. The internal associative strength of an item xy is estimated by:

$$score(xy) = \sum_{i=1}^{9} (wt_i \times score_i(xy))$$

where $score_i(xy)$ is the value of xy given by the i-th measure, and wt_i is the weight for the i-th measure accordingly (satisfying $\sum_{i=1}^{9} wt_i = 1$).

The determination of weights is not straightforward due to the presence of combinatorial explosion (notice that wt_i is real number). Genetic algorithm (Pan 1998) is explored to adjust the weights automatically, trying to find the optimal one. Let (wt_1, wt_2,…,wt_9) be a possible solution, we set the F-measure of word extraction on TS1 to be the fitness function, and set the size of population to be 25. We simply use the GenocopIII software (Michalewicz) to do the job.

In a PIII650 PC, GenocopIII runs 12 hours, iterates 1,161 generation, and converges to a group of weights (as shown in Table 3). With this group of weights, the F-measure of word extraction on TS1 is 55.44%, improving only 0.67% over the most powerful single measure MI (54.77%). Note this is not a pure open test, because the fitness function of genetic algorithm is based on TS1.

2.4 Interval-based Combination of Internal Measures

The experimental result in section 2.3 shows that it is not so effective to combine the nine measures directly. We try another combination scheme now, i.e., interval-based combination.

2.4.1 The Idea

The idea is as follows: for every measure mentioned above, we first discretize its value range into a number of intervals. Every interval of every measure is then assigned a corresponding probability that indicates the tendency of any item being regarded as a word if its value with respect to this measure falls into this interval. We name this kind of probability ‘the interval probability’. The soundness of an item being a word would be the weighted sum of all of its interval probabilities over nine measures.

We describe the idea in a more formal way. Suppose $score_i(xy)$ is the internal associative strength of any item xy with respect to the i-th measure, $v_i(xy)$ is its corresponding interval determined by the value of $score_i(xy)$, $pv_i(xy)$ is the interval probability of $v_i(xy)$, then the soundness of xy being word, $pv(xy)$, will be given by:

$$pv(xy) = \sum_{i=1}^{9}(wt_i \times pv_i(xy))$$

where wt_i is the weight for the i-th measure.

If $pv(xy)$ is larger than a threshold, xy would be extracted out as a word.

2.4.2 The Related Issues

Three related issues need to be clarified.

(1) How to discretize the range of a measure with continuous values?

We use D-2, an entropy-based top-down algorithm (Catlett 1991) to discretize the value range by supervised learning. It adopts the information gain as the criterion to decide whether a given training set should be further partitioned or not. Given a set of examples *S*, the information gain caused by a cut point t will be:

$$IG(t,S) = Ent(S) - \frac{|S_1|}{|S|}Ent(S_1) - \frac{|S_2|}{|S|}Ent(S_2)$$

where *Ent(S)* is the entropy of *S*, and S_1 and S_2 are two subsets of *S* partitioned by the cut point t.

It has been proved that the information gain obtains optimal discretization only on boundary points (Fayyad et al. 1992, Elomaa et al. 2000). So only boundary points need to be examined as potential cut points. Suppose *T* is the set of boundary points, the D-2 algorithm for discretizing set S is:

ALGORITHM DISCRETE (*S*,*T*)

```
BEGIN
    Step1. For each $t$ in $T$, calculate $IG(t,S)$
    Step2. Select $t_0 = \arg\max_t(IG(t,S))$, $S$ is
               partitioned into two subsets: $S_1, S_2$
    Step3. If stopping criteria are satisfied,
    Step4.    then DO NOT partition $S$, Return $\Phi$.
                  // $\Phi$ is an empty set
    Step5.    else    $P1$ = DISCRETE($S_1$, $T_1$)
                      $P2$ = DISCRETE($S_2$, $T_2$)
                      $P = P1 + \{t_0\} + P2$, Return $P$.
                  //$P$ is the set of cut points for
                    discretizing $S$
END
```

This algorithm only considers two stopping criteria. One is the minimal number of samples in an interval, the other is the minimum information gain. With this algorithm, we finally get a set of cut points each measure, which discretize the value range to variable-length intervals.

(2) How to assign the interval probability to each interval?

After discretization, training examples (i.e., a list of Chinese character bigrams) will be distributed to a certain interval according to its value of a given measure. Let v_{ij} represent the j-th interval of the i-th measure, then $pv(v_{ij})$, the interval probability of v_{ij}, is defined as:

$$pv(v_{ij}) = \frac{\#\ \ of\ \ bigrams\ being\ words\ in\ \ v_{ij}}{\#\ \ of\ \ bigrams\ in\ \ v_{ij}}$$

(3) How to set the weights for combining all $pv_i(xy)$ in the process of word extraction?

Genetic algorithm is again invoked to adjust the weight wt_i. The configuration of GenecopIII is the same as that in session 2.3.

2.4.3 Effect of the Stopping Criteria and Discretization Strategy on Combination

The stopping criteria and discretization strategy take effect on the word extraction performance of interval-based combination.

First, have a look at the effect of stopping criteria. In DISCRETE, two stopping criteria are needed to set. We fix the minimal number of samples in one interval on 50 arbitrarily. And, we change the setting of the minimum information gain. In Table 4, performances under five different minimum information gains are compared, marked as D1, D2 ,…, D5 respectively. It can be seen that, the smaller the minimal information gain, the finer the granularity of discretization, and the better the performance of word extraction. But if the discretization is too grainy, it may cause over-fitting problem. Compared with D4 and D5, D3 achieves nearly the same performance but has a much rough discretization. So, we set the minimum information gain to be 0.0001 (D3) in the following experiments.

Second, observe the effect of the discretization strategy. The equal-length discretization is compared to the variable-length discretization. We divide the value range of each measure into equal-length intervals, and let the number of intervals be identical to that in D3 accordingly. As shown in Table 4, the equal-length discretization only achieves the F-measure of 55.56%, which is much less than D3 (57.45%). This means the entropy-based discretization is more reasonable than equal-length discretization, and the discretization strategy has significant impact on the performance of interval-based combination.

2.4.4 Reduction of Measures for Combination

To improve the performance of word extraction through combination, the premise is that there must be enough mutual supplements among those measures. However, if the combination involves too many measures, interference may become obvious. We try to reduce the number of measures for combination.

The reduction procedure is recursive: It first compares the performance after removing any of the n measures, then reduces the one that can bring the most improvement of performance if it is removed. Repeat this reduction procedure in the left n-1 measures, until the performance cannot improve anymore.

Table 5 shows the reduction procedure of the nine internal measures. It indicates that, excluding SA and SCP, the interval-based combination of other seven measures could achieve the best F-measure of 57.77%, with the weights in Table 6. That result is 3.00% higher than that of the most powerful internal measure MI (54.77%).

Note again that all tests in section 2.4 are not pure open, because all the related parameters such as granularity of discretization, reduction of measures and adjustment of combination weights, are based on TS1.

Table 4. Effect of the Stopping Criteria and Discretization Strategy (Based on TS1)

Entropy-based Discretization		F-measure (%)	Number of Partitions								
	Min Gain		Freq	MI	SA	SCP	Dice	LogL	Chi	ZS	TS
D1	0.001	55.92	80	117	88	63	122	126	237	189	66
D2	0.0005	56.75	104	341	230	96	255	314	380	343	625
D3	0.0001	57.45	340	1449	641	109	471	1390	949	1411	1234
D4	0.00005	57.67	385	1660	693	113	543	1777	1316	1555	1808
D5	0.00001	57.69	423	2204	754	120	581	2522	2375	2233	2304
Equal-length Discretization		55.56	340	1449	641	109	471	1390	949	1411	1234

Table 5. The Reduction Procedure of the Nine Measures (Based on TS1)

N	F-measure (%) after Removing									Action
	Freq	MI	SA	SCP	Dice	LogL	Chi	Zs	Ts	
9	57.62	55.09	57.63	**57.69**	57.48	57.65	57.64	57.50	57.37	Reduce SCP
8	57.71	55.19	**57.77**		57.63	57.60	57.40	57.36	57.49	Reduce SA
7	57.67	55.25			57.66	57.71	57.74	57.64	57.48	No Reduction

Table 6. Weights for Interval-based Combination (Based on TS1)

Freq	MI	SA	SCP	Dice	LogL	Chi	ZS	TS
0.00034	0.47238	0	0	0.00238	0.00125	0.09339	0.25636	0.17390
Comparison of weights: MI > ZS > TS > Chi > Dice > Freq								

Table 7. Open Test for Effect of Internal Measures, the Contextual Measures and the Hybrid (on TS2)

	Precision(%)	Recall(%)	F-measure(%)	Setting t_1 and t_2 for Left/Right Entropy
MI	56.72	58.97	57.82	N.A.
Comb	60.41	59.35	59.87	N.A.
MI+Le/Re	83.53	54.88	66.24	MI-tuned threshold
Comb+Le/Re*	85.69	55.76	67.56	MI-tuned threshold
Comb+Le/Re	85.71	57.02	68.48	Comb-tuned threshold

3 The Contextual Measure

This section turns to discuss how to make use of contextual measures. The most commonly used contextual measure is the left/right entropy:

$$Le(xy) = -\sum_{\forall a \in A} p(axy \mid xy) \cdot \log_2 p(axy \mid xy)$$

$$\mathrm{Re}(xy) = -\sum_{\forall b \in A} p(xyb \mid xy) \cdot \log_2 p(xyb \mid xy)$$

where: xy is the candidate item, *a*, *b* are Chinese characters belonging to *A*, the set of Chinese characters.

In the sight of entropy, the larger the value of *Le(xy)* and *Re(xy)*, the more various the characters coming after/before *xy*, and thus the more possible *xy* to be a word.

4 The Hybrid of Internal and Contextual Measures

Combining the contextual measure with internal measures, the word extraction process would become like this: First, any candidate item xy not satisfying the contextual condition is rejected. The contextual condition is, *Le(xy)*>t_1 and *Re(xy)*>t_2. Second, those residual candidates will be extracted out as words if their internal measure or combination of internal measures is high than a given threshold t_3. In this paper, we try two alternatives of hybrid for comparison: one is the contextual measure with mutual information, the best single internal measure; another one is the contextual measure with Comb, the best result of interval-based combination of seven internal measures (Freq, MI, Dice, LogL, Chi, ZS, TS).

We need to determine three thresholds in above process, Threshold t_3 are set as the value to select the top 17,333 candidates from TS1: according to experiments in section 2, MI will choose t_3=4.0, while Comb will choose t_3=0.26. To set appropriate thresholds t_1 and t_2, we still employ genetic algorithm. We let a group of threshold (t_1, t_2) be a possible solution, and let the F-measure of word extraction on TS1 be fitness. Two groups of thresholds can be thus obtained:

(1) MI-tuned thresholds: t_1=2.2, t_2=1.4.

(2) Comb-tuned thresholds: t_1=1.8, t_2=1.2.

To further investigate the effect of internal measures, the contextual measure and the hybrid, we conduct a series of open tests on TS2, as demonstrated in Table 7. Since the left/right entropy would become less reliable in cases that the occurrences of contexts are not sufficient, we drop out those candidates whose frequencies are no more than 5 in TS2. After dropping, TS2 contains 14,867 candidates, out of which 1,589 are words according to PDA98J.

In the first two rows of Table 7, the best single internal measure, MI, and our best combination of internal measures Comb are open tested. The successive three rows show the effect of contextual measures. The row of 'MI+Le/Re' selects MI as the internal measure, and use the MI-tuned thresholds as t_1 and t_2. The rows of 'Comb+Le/Re*' and 'Comb+Le/Re' both select Comb as the internal measure, but use different t_1 and t_2: The former uses MI-tuned thresholds, while the latter uses Comb-tuned thresholds.

From Table 7, we can draw several conclusions: (1) With open test, the F-measure of MI, the best single internal measure, is 57.82%, whereas the F-measure of our interval-based combination is 59.87%; (2) The integration of the commonly used contextual measure, the left/right entropy with internal measures, can bring a large improvement of about 8%~9%; (3) There is only a modest difference between the performances of

‘Comb+Le/Re*’ and ‘Comb+Le/Re’, and two group of thresholds adjusted by different internal measures have small difference as well.

5 Conclusion

This paper focuses on the research of pure statistic-based measures for automatic extraction of two-character Chinese words. Two kinds of statistic-based measures are discussed: internal measures and contextual measures. Nine internal measures are tested and compared. Two schemes are tried to improve the performance by properly combining these nine measures. Experimental results in open tests show that, the best combination scheme, interval-based combination, achieves the F-measure of 59.87%, improving 2.05% over the best single internal measure mutual information. On the other hand, the left/right entropy, a kind of contextual measure, is integrated to acquire further improvement in word extraction. With the left/right entropy and interval-based combination of internal measures, the F-measure ultimately achieves 68.48%. Another point of this paper is that, weights for combination and thresholds for left/right entropy are adjusted automatically by genetic algorithm, rather than manually.

Future work will extend the proposed method to automatic extraction of multi-character Chinese words. Other useful information, such as lexicon and semantic resource, are expected to be included for consideration so as to further improve the performance.

Acknowledgements

This research is supported by the National ‘973’ Plan of China under grant no G1998030507 and NSFC under grant no 60083005.

References

Catlett. (1991) On changing continuous attributes into odered discrete attributes. In Proceedings of the European Working Session on Learning, Berlin, Germany. pp. 164-178

Chien, L.F. (1999) Pat-tree-based adaptive keyphrase extraction for intelligent Chinese information retrieval. Information Processing and Management vol.35 pp.501-521

Elomaa T., Rousu J., (2000) Generalizing boundary points. In Proceedings of the 17th National Conference on Artificial Intelligence, Menlo Park, CA.

Fayyad, U., Irani, K., (1992) On the handling of continuous-valued attributes in decision tree generation. Machine Learning. Vol.(8) pp.87-102

Pan, Z.J., (1998) Evolution Computing. Tsinghua University Press, Beijing.

Sornlertlamvanich V., Potipiti T., Charoenporn T. (2000) Automatic corpus-based Thai word extraction with the C4.5 learning algorithm. In Proceedings of COLING 2000.

Schone, P., Jurafsky D. (2001) Is knowledge-free induction of multiword unit dictionary headwords a solved problem? In proceedings of EMNLP 2001.

Michalewicz, Z., Genocop III, available at:

http://www.coe.uncc.edu/~gnazhiya/gchome.html