A New Psychometric-inspired Evaluation Metric for Chinese Word Segmentation

Peng Qian Xipeng Qiu* Xuanjing Huang
Shanghai Key Laboratory of Intelligent Information Processing, Fudan University
School of Computer Science, Fudan University
825 Zhangheng Road, Shanghai, China
{pqian11, xpqiu, xjhuang}@fudan.edu.cn

Abstract

Word segmentation is a fundamental task for Chinese language processing. However, with the successive improvements, the standard metric is becoming hard to distinguish state-of-the-art word segmentation systems. In this paper, we propose a new psychometric-inspired evaluation metric for Chinese word segmentation, which addresses to balance the very skewed word distribution at different levels of difficulty[1]. The performance on a real evaluation shows that the proposed metric gives more reasonable and distinguishable scores and correlates well with human judgement. In addition, the proposed metric can be easily extended to evaluate other sequence labelling based NLP tasks.

1 Introduction

Word segmentation is a fundamental task for Chinese language processing. In recent years, Chinese word segmentation (CWS) has undergone great development, which is, to some degree, driven by evaluation conferences of CWS, such as SIGHAN Bakeoffs (Emerson, 2005; Levow, 2006; Jin and Chen, 2008; Zhao and Liu, 2010). The current state-of-the-art methods regard word segmentation as a sequence labeling problem (Xue, 2003; Peng et al., 2004). The goal of sequence labeling is to assign labels to all elements in a sequence, which can be handled with supervised learning algorithms, such as maximum entropy (ME) (Berger et al., 1996), conditional random fields (CRF) (Lafferty et al., 2001) and Perceptron (Collins, 2002).

Benefiting from the public datasets and feature engineering, Chinese word segmentation achieves quite high precision after years of intensive research. To evaluate a word segmenter, the standard metric consists of **precision** p, **recall** r, and an evenly-weighted **F-score** f_1.

However, with the successive improvement of performance, state-of-the-art segmenters are hard to be distinguished under the standard metric. Therefore, researchers also report results with some other measures, such as out-of-vocabulary (OOV) recall, to show their strengths besides p, r and f_1.

Furthermore, although state-of-the-art methods have achieved high performances on p, r and f_1, there exists inconsistence between the evaluation ranking and the intuitive feelings towards the segmentation results of these methods. The inconsistence is caused by two reasons:

(1) The high performance is due to the fact that the distribution of difficulties of words is unbalanced. The proportion of trivial cases is very high, such as '的 ('s)', '我们 (we)', which results in that the non-trivial cases are relatively despised. Therefore, a good measure should have a capability to balance the skewed distribution by weighting the test cases.

(2) Human judgement depends on difficulties of segmentations. A segmenter can earn extra credits when correctly segmenting a difficult word than an easy word. Conversely, a segmenter can take extra penalties when wrongly segmenting an easy word than a difficult word.

Taking a sentence and two predicted segmentations as an example:
S : 白藜芦醇 是 一 种 酚类 物质
(Trans: Resveratrol is a kind of phenols material.)
P1: 白 藜芦 醇 是 一种 酚类 物质
P2: 白藜 芦醇 是 一 种 酚类物 质

We can see that the two segmentations have the

*Corresponding author.

[1]We release the word difficulty of the popular word segmentation datasets at http://nlp.fudan.edu.cn/data/ .

Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, pages 2185–2194,
Berlin, Germany, August 7-12, 2016.

same scores in p, r and f_1. But intuitively, P1 should be better than P2, since P2 is worse even on the trivial cases, such as '酚类 (phenols)' and '物质 (material)'.

Therefore, we think that an appropriate evaluation metric should not only provide an all-around quantitative analysis of system performances, but also explicitly reveal the strengths and potential weaknesses of a model.

Inspired by psychometrics, we propose a new evaluation metric for Chinese word segmentation in this paper. Given a labeled dataset, not all words have the same contribution to judge the performance of a segmenter. Based on psychometric research (Lord et al., 1968), we assign a difficulty value to each word. The difficulty of a word is automatically rated by a committee of segmenters, which are diversified by training on different datasets and features. We design a balanced precision, recall to pay different attentions to words according to their difficulties.

We also give detailed analysis on a real evaluation of Chinese word segmentation with our proposed metric. The analysis result shows that the new metric gives a more balanced evaluation result towards the human intuition of the segmentation quality. We will release the weighted datasets focused this paper to the academic community.

Although our proposed metric is applied to Chinese word segmentation for a case study, it can be easily extended to other sequence labelling based NLP tasks.

2 Standard Evaluation Metric

The standard evaluation usually uses three measures: precision, recall and balanced F-score.

Precision p is defined as the number of correctly segmented words divided by the total number of words in the automatically segmented corpus.

Recall r is defined as the number of correctly segmented words divided by the total number of words in the gold standard, which is the manually annotated corpus.

F-score f_1 the harmonic mean of precision and recall.

Given a sentence, the gold-standard segmentation of a sentence is $w_1, \cdots, W_N$, N is the number of words. The predicted segmentation is $w'_1, \cdots, w'_{N'}$, N' is the number of words. Among that, the number of words correctly identified by the predicted segmentation is c, and the number of incorrectly predicted words is e.

p, r and f_1 are defined as follows:

$$p = \frac{c}{N'}, \tag{1}$$

$$r = \frac{c}{N}, \tag{2}$$

$$f_1 = \frac{2 \times p \times r}{p + r}. \tag{3}$$

As a complement to these metrics, researchers also use the recall of out-of-vocabulary (OOV) words to measure the segmenter's performance in detecting unknown words.

3 A New Psychometric-inspired Evaluation Metric

We involve the basic idea from psychometrics and improve the evaluation metric by assigning weights to test cases.

3.1 Background Theory

This work is inspired by the test theory in psychometrics (Lord et al., 1968). Psychologists, as well as educators, have studied the way of analyzing items in a psychological test, such as IQ test. The general idea is that test cases should be given different weights, which reflects the effectiveness of a certain item to a certain measuring object.

Similarly, we consider an evaluation task as a kind of special psychological test. The psychological traits, or the ability of the model, is not an explicit characteristics. We propose that the test cases for NLP task should also be assigned a real value to account for the credits that the tagger earned from answering the test case.

In analogy to the way of computing difficulty in psychometrics, the difficulty of a target word w_i is defined as the error rate of a committee in the case of word segmentation.

Given a committee of K base segmenters, we can get K segmentations for sentence $w_1, \cdots, W_N$. We use a mark $m_i^k \in \{0, 1\}$ to indicate whether word w_i is correctly segmented by the k-th segmenter.

The number of words c^k correctly identified by the k-th segmenter is

$$c^k = \sum_{i=1}^{N} m_i^k. \tag{4}$$

Thus, we can calculate the degree of difficulty of each word w_i.

$$d_i = \frac{1}{K}\sum_{k=1}^{K}(1 - m_i^k). \quad (5)$$

This methodology of measuring test item difficulty is also widely applied in assessing standardized exams such as TOEFL (Service, 2000).

3.2 Psychometric-Inspired Evaluation Metric

Since the distribution of the difficulties of words is very skew, we design a new metric to balance the weights of different words according to their difficulties. In addition, we also should keep strictly a fair rule for rewards and punishments.

Intuitively, if the difficulty of a word is high, a correct segmentation should be given an extra rewards; otherwise, if the difficulty of a word is low, it is reasonable to give an extra punishment to a wrong segmentation.

Our new metric of precision, recall and balanced F-score is designed as follows.

Balanced Recall Given a new predicted segmentation, the mark $m_i \in \{0, 1\}$ indicates whether word w_i is correctly segmented. d_i is the degree of difficulty of word w_i.

According to the difficulties of each word, we can calculated the reward recall r_{reward} which biased for the difficult cases.

$$r_{reward} = \frac{\sum_{i=1}^{N} d_i \times m_i}{\sum_{i=1}^{N} d_i}, \quad (6)$$

where $r'_{reward} \in [0, 1]$ is biased recall, which places more attention on the difficult cases and less attention on the easy cases.

Conversely, we can calculated another punishment recall $r_{punishment}$ which biased for the easy cases.

$$r_{punishment} = \frac{\sum_{i=1}^{N} (1 - d_i) \times m_i}{\sum_{i=1}^{N} (1 - d_i)}, \quad (7)$$

where $r_{punishment} \in [0, 1]$ is biased recall, which places more attention on the easy cases and less attention on the difficult cases.

$r_{punishment}$ can be interpreted as a punishment as bellows.

$$r_{punishment} = \frac{\sum_{i=1}^{N} (1 - d_i) \times m_i}{\sum_{i=1}^{N} (1 - d_i)}, \quad (8)$$

$$= 1 - \frac{\sum_{i=1}^{N} (1 - d_i) \times (1 - m_i)}{\sum_{i=1}^{N} (1 - d_i)}. \quad (9)$$

From Eq (9), we can see that an extra punishment is given to wrong segmentation for low difficult word. In detailed, for a word w_i that is easy to segment, its weights $(1 - d_i)$ is relative higher. When its segmentation is wrong, $m_i = 0$. Therefore, $(1 - d_i) \times (1 - m_i) = (1 - d_i)$ will be larger, which results to a smaller final score.

To balance the reward and punishment, a balanced recall r_b is used, which is the harmonic mean of r_{reward} and $r_{punishment}$.

$$r_b = \frac{2 \times r_{punishment} \times r_{reward}}{r_{punishment} + r_{reward}} \quad (10)$$

Balanced Precision Given a new predicted segmentation, the mark $m'_i \in \{0, 1\}$ to indicate whether segment s'_i is correctly segmented. d'_i is the degree of difficulty of segment s'_i, which is an average difficulty of the corresponding gold words.

Similar to balanced recall, we use the same way to calculate balance precision p_b. Here N' is the number of words in the predicted segmentation. d'_i is the weight for the predicted segmentation unit w'_i. It equals to the word difficulty of the corresponding word w that cover the right boundary of w'_i in the gold segmentation.

$$p_{reward} = \frac{\sum_{i=1}^{N} (1 - d_i) \times m_i}{\sum_{i=1}^{N'} (1 - d'_i)}, \quad (11)$$

$$p_{punishment} = \frac{\sum_{i=1}^{N} (1 - d_i) \times m_i}{\sum_{i=1}^{N'} (1 - d'_i)}, \quad (12)$$

$$p_b = \frac{2 \times p_{reward} \times p_{punishment}}{p_{reward} + p_{punishment}}. \quad (13)$$

$$(14)$$

Balanced F-score The final balanced F-score is

$$f_b = \frac{2 \times p_{balanced} \times r_{balanced}}{p_{balanced} + r_{balanced}}. \quad (15)$$

4 Committee of Segmenters

It is infeasible to manually judge the difficulty of each word in a dataset. Therefore, an empirical method is needed to rate each word. Since the difficulty is also not derivable from the observation of the surface forms of the text, we use a committee of automatic segmenters instead. To keep fairness

F1	$C_iT_0, (i = -1, 0, 1)$
	$C_{i:i+1}T_0, (i = -1, 0)$
	$T_{-1,0}$
F2	$C_iT_0, (i = -2, -1, 0, 1, 2)$
	$C_{i:i+1}T_0, (i = -2, -1, 0, 1)$
	$T_{-1,0}$
F3	$C_iT_0, (i = -2, -1, 0, 1, 2)$
	$C_{i:i+1}T_0, (i = -2, -1, 0, 1)$
	$C_{i:i+2}T_0, (i = -2, -1, 0)$
	$T_{-1,0}$

Table 1: Feature templates. C represents a Chinese character, and T represents the character-based tag in set {B, M, E, S}. The subscript indicates its position relative to the current character, whose subscript is 0. $C_{i:j}$ represents the subsequence of characters form relative position i to j.

and justice of the committee, we need a large number of diversified committee members.

Thus, the grading result of committee is fair and accurate, avoiding the laborious human annotation and the deviation caused by the subjective factor of the artificial judgement.

4.1 Building the Committee

Base Segmenters The committee is composed of a series of base segmenters, which are based on discriminative character-based sequence labeling method. Each character is labeled as one of {B, M, E, S} to indicate the segmentation. 'B' indicates the beginning character of a word. 'M' indicates the middle character of a word. 'E' indicates the end character of a word. 'S' indicates that the word consists of only a single character.

Diversity of Committee To objectively assess the difficulty of a word, we need to maintain a large enough committee with diversity.

To encourage diversity among committee members, we train them with different datasets and features. Specifically, each base segmenter adopts one of three types of feature templates (shown in Table 1), and are trained on randomly sampling training sets. To keep a large diversity, we set sampling ratio to be 10%, 20% and 30%. In short, each base segmenter is constructed with a random combination of the candidate feature template and the sampling ratio for training dataset.

Size of Committee To obtain a valid and reliable assessment for a word, we need to choose the

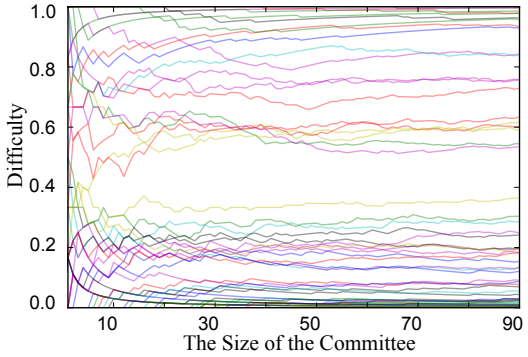


Figure 1: Judgement of difficulty against the committee size. Each line represents a sampled word.

appropriate size of committee. For a given test case, the judgement of its difficulty should be relatively stable. We analyze how the judgement of its difficulty changes as the size of committee increases.

Figure 2 show PKU data from SIGHAN 2005 (Emerson, 2005) the difficulty is stable when the sample size is large enough.

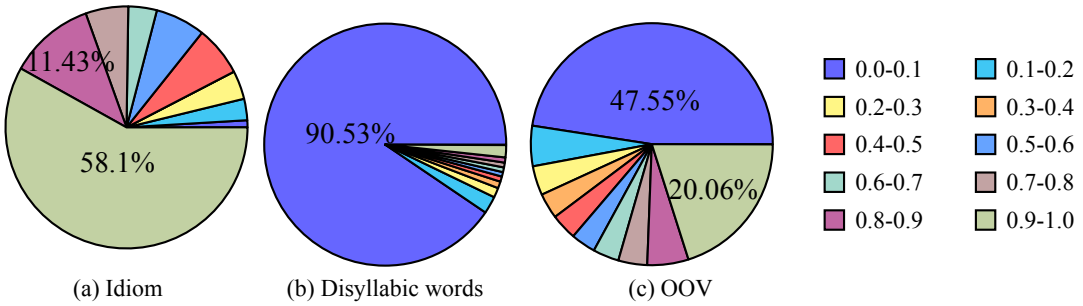
4.2 Interpreting Difficulty with Linguistic Features

Since we get the difficulty for each word empirically, we naturally want to know whether the difficulty is explainable, as what TOEFL researchers have done (Freedle and Kostin, 1993; Kostin, 2004). We would like to know whether the variation of word difficulty can be partially expalined by a series of traceable linguistic features.

Based on the knowledge about the characteristics of Chinese grammar and the practical experiences of corpus annotation, we consider the following surface linguistic features. In order to explicitly display the relationship between the linguistic predictors and the distribution of the word difficulty at a micro level, we divide the difficulty scale into ten discrete intervals and calculate the distributions of these linguistic features on different ranges of difficulty.

Here, we interpret the difficulties of the words from the perspective of three important linguistic features:

Idiom In Chinese, the 4-character idioms have special linguistic structures. These structure usually form a different pattern that is hard for the machine algorithm to understand. Therefore,



(a) Idiom (b) Disyllabic words (c) OOV

Figure 2: Difficulty distribution of (a) idioms, (b) dysyllabic words and (c) Out-of-vocabulary words from PKU dataset. Similar pattern has also been found in other datasets.

it is reasonable to hypothesize that the an idiom phrase is more likely to be a difficult word for word segmentation task. We can see from Figure 2a that 58.1% of idioms have a difficulty at (0.9,1]. The proportion does increase with the degree of difficulty, which corresponds with the human intuition.

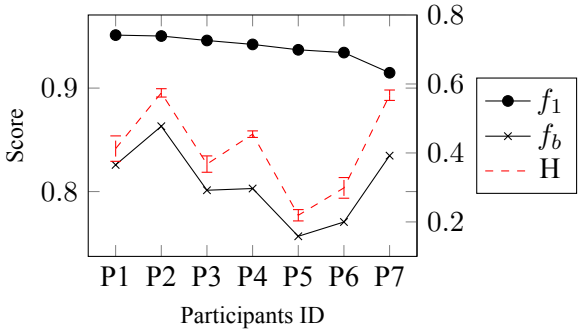
Dissyllabic Word Disyllabic word is a word formed by two consecutive Chinese characters. We can see from Figure 2b that the frequency of disyllabic words has a negative correlations with the degree of difficulty. This is an interesting result. It means that a two-syllable word pattern is easy for a machine algorithm to recognize. This is consistent with the lexical statistics (Yip, 2000), which shows that dissyllabic words account for 64% of the common words in Chinese.

Out-of-vocabulary Word Processing out-of-vocabulary (OOV) word is regarded as one of the key factors to the improvement of model performance. Since these words never occur in the training dataset, it is for sure that the word segmentation system will find it hard to correctly recognize these words from the contexts. We can see from Figure 2c that OOV generally has high difficulty. However, a lot of OOV is relatively easy for segmenters.

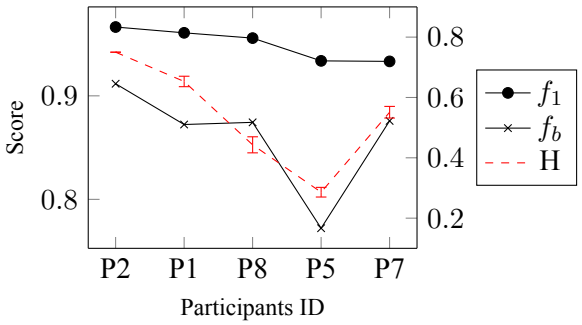
All the linguistic predictors above prove that the degree of difficulty, namely the weight for each word, is not only rooted in the foundation of test theory, but also correlated with linguistic intuition.

5 Evaluation with New Metric

Here we demonstrate the effectiveness of the proposed method in a real evaluation by re-analyzing the submission results from NLPCC



(a) Closed Track



(b) Open Track

Figure 3: Comparisons of standard metric and our new metric for the closed track and the open track of NLPCC 2015 Weibo Text Word Segmentation Shared Task. The black lines for f_1 and f_b are plotted against the left y-axis. The red lines for human judgement scores are plotted against the right y-axis.

2015 Shared Task[2] of Chinese word segmentation. The dataset of this shared task is collected from micro-blog text. For convenience, we use WB to represent this dataset in the following discussions.

We select the submissions of all 7 participants from the closed track and the submissions of all

[2]Conference on Natural Language Processing and Chinese Computing. http://tcci.ccf.org.cn/conference/2015/

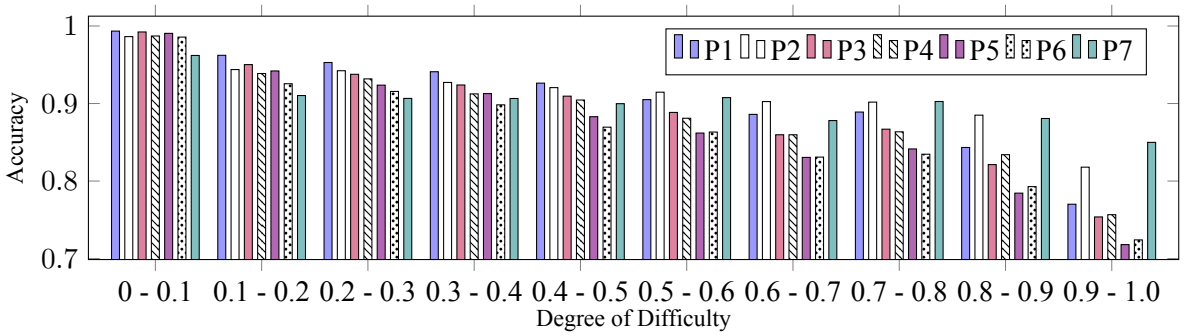


Figure 4: Accuracies of different participants in Closed Track by different difficulties on WB dataset.

5 participants from the open track. In the closed track, participants could only use information found in the provided training data. In the open track, participants could use the information which should be public and be easily obtained.

We compare the standard precision, recall and F-score with our new metric. The result is displayed in Figure 3. Considering the related privacy issues, we will refer to the participants as P1, P2, etc. The order of these participants in the sub-figures is sorted according to the original ranking given by the standard metric in each track. The same ID number refers to the same participants.

It is interesting to see that the proposed metric gives out significantly different rankings for the participants, compared to the original rankings. Based on the standard metric, Participant 1 (P1) ranks the top in closed track while P7 is ranked as the worst in both tracks. However, P2 ranks first under the evaluation of the new metric in the Closed track. P7 also get higher ranking than its original one.

5.1 Correlation with Human Judgement

To tell whether the standard metric or the proposed metric is more reasonable, we asked three experts to evaluate the quality of the submissions from the participants. We randomly selected 50 test sentences from the WB dataset. For each test sentence, we present all the submitted candidate segmentation results to the human judges in random order. Then, the judges are asked to choose the best candidate(s) with the highest segmentation quality as well as the second-best candidate(s) among all the submissions. Human judges had no access to the source of the sentences.

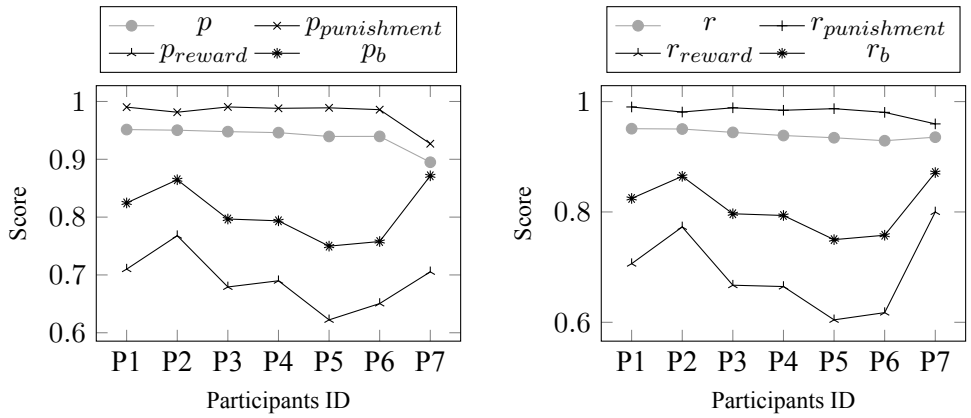
Once we collect the human judgement of the segmentation quality, we can compute the score for each participants. If a candidate segmentation result from a certain participant is ranked first for n times, then this participants earned n point. If second for m times, then this participants earned $\frac{m}{2}$ points. Then we can get the probability of a participants being ranked the best or sub-best by computing $\frac{n+\frac{m}{2}}{50}$. Finally, we get the human-intuition-based gold ranking of the participants through the means of scores from all the human judges.

It is worth noticing that the ranking result of our proposed metric correlates with the human judgements better than that of the standard metric, as is shown in Figure 3. The Pearson correlation between our proposed metric and human judgements are 0.9056 ($p = 0.004$) for closed session and 0.8799 ($p = 0.04$) for open session while the Pearson correlation between standard metric and human judgements are only 0.096 ($p = 0.836$) for closed session and 0.670 ($p = 0.216$). This evidence strongly supports that the proposed method is a good approximate of human judgements.

5.2 Detailed Analysis

Since we have empirically got the degree of difficulty for each word in the test dataset, we can compute the distribution of the difficulty for words that have been correctly segmented. We divided the whole range of difficulty into 10 intervals. Then, we count the ratio of the correct segmented units for each difficulty interval. In this way, we can quantitatively measure to what extent the segmentation system performs on difficult test cases and easy test cases.

As is shown in Figure 4, P7 works better on difficult cases than other systems, but the worst on easy cases. This explains why P7 gets good rank based on the new evaluation metric. Besides,



(a) Standard and Weighted Precision

(b) Standard and Weighted Recall

Figure 5: Comparisons of standard and weighted precision and recall on NLPCC Closed Track.

if we compare P1 and P2, we will notice that P2 performs just slightly worse than P1 on easy cases, but much better than P1 on difficult cases. Therefore, conventional evaluation metric rank P1 as the top system because the P1 gets a lot of credits from a large portion of easy cases. Unlike conventional metric, our new metric achieves balance between hard cases and easy cases and ranks P2 as the top system.

The experiment result indicates that the new metric can reveal the implicit difference and improvement of the model, while standard metric cannot provide us with such a fine-grained result.

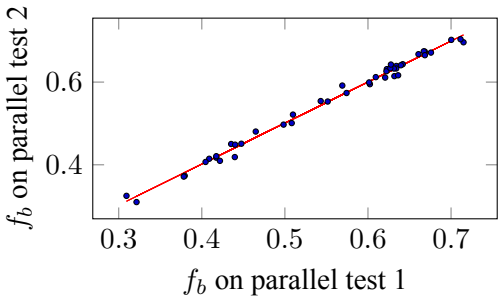


Figure 6: Correlation between the evaluation results f_b of two parallel testsets with the proposed metrics on a collection of models. The Pearson correlation is 0.9961, $p = 0.000$.

5.3 Validity and Reliability

Jones (1994) concluded some important criteria for the evaluation metrics of NLP system. It is very important to check the validity and reliability of a new metric.

Previous section has displayed the validity of the proposed evaluation metric by comparing the evaluation results with human judgements. The evaluation results with our new metric correlated with human intuition well.

Regarding reliability, we perform the parallel-test experiment. We randomly split the test dataset into two halves. These two halves have similar difficulty distribution and, therefore, can be considered as a parallel test. Then different models, including those used in the first experiment, are evaluated on the first half and the second half. The results in Figure 6 shows that the performances of different models with our proposed evaluation metric are significantly correlated in two parallel tests.

5.4 Visualization of the Weight

As is known, there might be some annotation inconsistency in the dataset. We find that most of the cases with high weight are really valuable difficult test cases, such as the visualized sentences from WB dataset in Figure 7. In the first sentence, the word ‘BMW 族’ (NOUN.People who take bus, metro and then walk to the destination) is an OOV word and contains English characters. The weight of this word, as expected, is very high. In the second sentence, the word ‘素不相识’ (VERB.not familiar with each other) is a 4-character Chinese idiom. the conjunction word ‘就算’ (CONJ.even if) has structural ambiguity. It can also be decomposed into a two-word phrase ‘就’ (ADV.just) and ‘算’ (VERB.count). From the visualization of the weight, we can see that these difficult words are all given high weights.

Data	Corpus Size	p	r	$f1$	p_b	r_b	f_b
PKU	20%	90.04	89.90	89.97	45.22	43.37	44.28
	50%	92.87	91.58	92.22	54.24	49.12	51.55
	80%	94.07	92.21	93.13	61.80	54.74	58.05
	100%	94.03	92.91	93.47	64.22	59.16	61.59
MSR	20%	92.93	92.58	92.76	45.76	44.13	44.93
	50%	95.22	95.18	95.20	63.00	62.22	62.60
	80%	95.68	95.74	95.71	67.26	66.96	67.11
	100%	96.19	96.02	96.11	70.80	69.45	70.12
NCC	20%	87.32	86.37	86.84	42.16	40.23	41.17
	50%	89.34	89.03	89.19	50.31	49.26	49.78
	80%	91.42	91.10	91.26	60.48	59.25	59.86
	100%	92.00	91.77	91.89	63.72	62.70	63.20
SXU	20%	89.70	89.31	89.50	43.53	42.35	42.93
	50%	93.04	92.42	92.73	56.21	54.27	55.23
	80%	94.45	93.94	94.19	64.55	62.50	63.51
	100%	94.89	94.61	94.75	68.10	66.63	67.36

Table 2: Model evaluation with standard metric and our new metric. Models vary in the amount of training data and feature types.

6 Comparisons on SIGHAN datasets

In this section, we give comparisons on SIGHAN datasets. We use four simplified Chinese datasets: PKU and MSR (SIGHAN 2005) as well as NCC and SXU (SIGHAN 2008).

For each dataset, we train four segmenters with varying abilities, based on 20%, 50%, 80% and 100% of training data respectively. The used feature template is F2 in Table 1.

Table 2 shows the different evaluation results with standard metric and our balanced metric. We can see that the proposed evaluation metric generally gives lower and more distinguishable score, compared with the standard metric.

7 Related work

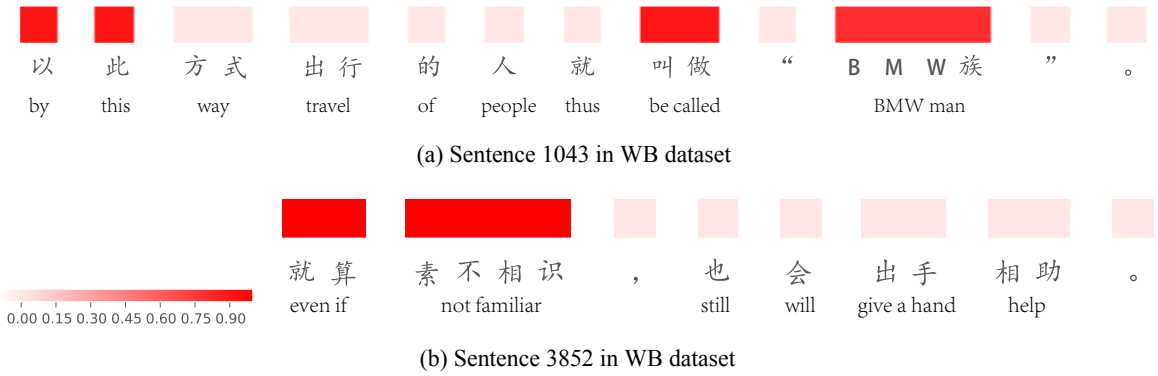
Evaluation metrics has been a focused topic for a long time. Researchers have been trying to evaluate various NLP tasks towards human intuition (Papineni et al., 2002; Graham, 2015a; Graham, 2015b). Previous work (Fournier and Inkpen, 2012; Fournier, 2013; Pevzner and Hearst, 2002) mainly deal with the near-miss error case on the context of text segmentation. Much attention has been given to different penalization for the error. These work criticize that traditional metrics such as precision, recall and F-score, consider all the error similar. In this sense, some studies aimed at assigning different penalization to the word. We think that these explorations can be regarded as the foreshadowing of our evaluation metric that balances reward and punishment.

Our paper differs from previous research in that we take the difficulty of the test case into consideration, while previous works only focus on the variation of error types and penalisation. We involve the basic idea from psychometrics and improve the evaluation with a balance between difficult cases and easy cases, reward and punishment.

We would like to emphasize that our weighted evaluation metric is not a replacement of the traditional precision, recall, and F-score. Instead, our new weighted metrics can reveal more details that traditional evaluation may not be able to present.

8 Conclusion

In this paper, we put forward a new psychometric-inspired method for Chinese word segmentation evaluation by weighting all the words in test dataset based on the methodology applied to psychological tests and standardized exams. We empirically analyze the validity and reliability of the new metric on a real evaluation dataset. Experiment results reveal that our weighted evaluation metrics gives more reasonable and distinguishable scores and



(a) Sentence 1043 in WB dataset

(b) Sentence 3852 in WB dataset

Figure 7: Visualising the word weight of WB dataset.

correlates well with human judgement. We will release the weighted datasets to the academic community.

Additionally, the proposed evaluation metric can be easily extended to word segmentation task for other languages (e.g. Japanese) and other sequence labelling-based NLP tasks, with just tiny changes. Our metric also points out a promising direction for the researchers to take into the account of the biased distribution of test case difficulty and focus on tackling the hard bones of natural language processing.

Acknowledgments

We would like to thank the anonymous reviewers for their valuable comments. This work was partially funded by National Natural Science Foundation of China (No. 61532011, 61473092, and 61472088), the National High Technology Research and Development Program of China (No. 2015AA015408).

References

A.L. Berger, V.J. Della Pietra, and S.A. Della Pietra. 1996. A maximum entropy approach to natural language processing. *Computational Linguistics*, 22(1):39–71.

Michael Collins. 2002. Discriminative training methods for hidden markov models: Theory and experiments with perceptron algorithms. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing*.

T. Emerson. 2005. The second international Chinese word segmentation bakeoff. In *Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing*, pages 123–133. Jeju Island, Korea.

Chris Fournier and Diana Inkpen. 2012. Segmentation similarity and agreement. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 152–161. Association for Computational Linguistics.

Chris Fournier. 2013. Evaluating text segmentation using boundary edit distance. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 1702–1712.

Roy Freedle and Irene Kostin. 1993. The prediction of toefl reading item difficulty: Implications for construct validity. *Language Testing*, 10(2):133–170.

Yvette Graham. 2015a. Improving evaluation of machine translation quality estimation. In *Proceedings of the 53th Annual Meeting on Association for Computational Linguistics*.

Yvette Graham. 2015b. Re-evaluating automatic summarization with bleu and 192 shades of rouge. In *Proceedings of EMNLP*.

C. Jin and X. Chen. 2008. The fourth international Chinese language processing bakeoff: Chinese word segmentation, named entity recognition and Chinese pos tagging. In *Sixth SIGHAN Workshop on Chinese Language Processing*, page 69.

Karen Sparck Jones. 1994. Towards better nlp system evaluation. In *Proceedings of the workshop on Human Language Technology*, pages 102–107. Association for Computational Linguistics.

Irene Kostin. 2004. Exploring item characteristics that are related to the difficulty of toefl dialogue items. *ETS Research Report Series*, 2004(1):i–59.

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*.

Gina-Anne Levow. 2006. The third international chinese language processing bakeoff: Word segmentation and named entity recognition. In *Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing*, pages 108–117, Sydney, Australia, July.

Frederic M Lord, Melvin R Novick, and Allan Birnbaum. 1968. *Statistical Theories of Mental Test Scores.* Addison-Wesley.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.

F. Peng, F. Feng, and A. McCallum. 2004. Chinese segmentation and new word detection using conditional random fields. *Proceedings of the 20th international conference on Computational Linguistics*.

Lev Pevzner and Marti A Hearst. 2002. A critique and improvement of an evaluation metric for text segmentation. *Computational Linguistics*, 28(1):19–36.

Educational Testing Service. 2000. *Computer-Based TOEFL Score User Guide*. Princeton, NJ.

N. Xue. 2003. Chinese word segmentation as character tagging. *Computational Linguistics and Chinese Language Processing*, 8(1):29–48.

Po-Ching Yip. 2000. *The Chinese Lexicon: A Comprehensive Survey*. Psychology Press.

H. Zhao and Q. Liu. 2010. The cips-sighan clp 2010 chinese word segmentation bakeoff. In *Proceedings of the First CPS-SIGHAN Joint Conference on Chinese Language Processing*.