Recall is the Proper Evaluation Metric for Word Segmentation

Yan Shao and **Christian Hardmeier** and **Joakim Nivre**
Department of Linguistics and Philology, Uppsala University
{yan.shao, christian.hardmeier, joakim.nivre}@lingfil.uu.se

Abstract

We extensively analyse the correlations and drawbacks of conventionally employed evaluation metrics for word segmentation. Unlike in standard information retrieval, precision favours under-splitting systems and therefore can be misleading in word segmentation. Overall, based on both theoretical and experimental analysis, we propose that precision should be excluded from the standard evaluation metrics and that the evaluation score obtained by using only recall is sufficient and better correlated with the performance of word segmentation systems.

1 Introduction

Word segmentation (WS) or tokenisation can be viewed as correctly identifying valid boundaries between characters (Goldwater et al., 2007). It is the initial step for most higher level natural language processing tasks, such as part-of-speech tagging, syntactic analysis, information retrieval and machine translation. Thus, correct segmentation is crucial as segmentation errors propagate to higher level tasks.

Because only correctly segmented words are meaningful to higher level tasks, word level precision, recall and their evenly-weighted average F1-score that are customised from information retrieval (IR) (Kent et al., 1955) are conventionally used as the standard evaluation metrics for WS (Sproat and Emerson, 2003; Qiu et al., 2015).

In this paper, we thoroughly investigate precision and recall in addition to true negative rate in the scope of WS, with a special focus on the drawbacks of precision. Precision and F1-score can be misleading as an under-splitting system may obtain higher precision despite having fewer correctly segmented words. Additionally, we conduct word segmentation experiments to investigate the connections between precision and recall as well as their correlations with actual performance of segmenters. Overall, we propose that precision should be excluded and that using recall as the sole evaluation metric is more adequate.

2 Evaluation Metrics for WS

2.1 Precision and Recall

By employing word-level precision and recall, the adequacy of a word segmenter is measured via comparing to the annotated reference. The correctly segmented words are regarded as true positives (*TP*). To obtain precision, *TP* is normalised by the prediction positives (*PP*), which is equal to total number of words returned by the system. For recall, we divide *TP* by the real positives (*RP*), the total number of words in the reference. The complement of *RP* is referred to as real negatives (*RN*).

In the evaluation setup for standard IR tasks, there is no entanglement between *RP* and *RN*. For any instances i_p and i_n in *RP* and *RN*, they can be in the same output set I of an IR system as:

$$\forall i_p \in RP, \forall i_n \in RN, \exists I, \{i_p, i_n\} \subset I$$

Precision and recall are thus not directly correlated. For IR, system performance is well measured only if both precision and recall are used as it is trivial to optimise with respect to either precision or recall, but difficult to improve both. This is not the case for WS. In contrast to the situation in IR, the characters as basic elements are fixed in WS. We only predict the boundaries whereas the characters can be neither added nor deleted, which makes positives and negatives correlated.

In Table 1, the source Chinese sentence and its English translation in the form of character strings

Proceedings of the The 8th International Joint Conference on Natural Language Processing, pages 86–90,
Taipei, Taiwan, November 27 – December 1, 2017

	Source sentences: 约翰喜欢玛丽 John likes Mary Reference Segmentations: 约翰 /喜欢 /玛丽 John / likes / Mary				
Segmenters	T1	T2	S1	S2	S3
Output	约 /翰 /喜 /欢 /玛 /丽	约翰喜欢玛丽	约翰 /喜欢玛丽	约翰 /喜 /欢 /玛 /丽	约翰 /喜欢 /玛 /丽
TP	0	0	1	1	2
P	0	0	1/2 = 0.5	1/5 = 0.2	2/4 = 0.5
R	0	0	1/3 = 0.33	1/3 = 0.33	2/3 = 0.67
F	0	0	0.40	0.25	0.57
TNR	1-6/18 = 0.67	1-1/18 = 0.94	1-1/18 = 0.94	1-4/18 = 0.77	1-2/18 = 0.89
Output	J/o/h/n/l/i/k/e/s/M/a/r/y	John likes Mary	John / likes Mary	John /l/i/k/e/s/M/a/r/y	John /likes /M/a/r/y
TP	0	0	1	1	2
P	0	0	1/2 = 0.5	1/10 = 0.1	2/6 = 0.33
R	0	0	1/3 = 0.33	1/3 = 0.33	2/3 = 0.67
F	0	0	0.40	0.15	0.44
TNR	1-13/88 = 0.85	1-1/88 = 0.99	1-1/88 = 0.99	1-9/88 = 0.90	1-4/88 = 0.95

Table 1: Sample sentences along with the output of two trivial segmenters (T1, T2) and three other segmenters (S1, S2, S3). True Positives (TP), Precision (P), recall (R), F1-score (F) and true negative rate (TNR) are calculated respectively.

are presented along with the outputs of five handcrafted segmenters. In WS, a *TP* simultaneously rejects the associated true negatives (*TN*). For the English sentence in Table 1, the positive segment *John* never appears simultaneously with its associated negatives *Joh*, *Jo* or *ohn* in the output. This positively correlates precision and recall, because if we modify a boundary that optimises recall, the precision will also improve. In WS, 100% recall guarantees 100% precision and it is non-trivial to optimise one without the other.

In the most trivial case, a segmenter either predicts and returns all the possible word boundaries (T1, extremely over-splitting) or fails to identify any boundaries at all (T2, extremely under-splitting). In the example, both strategies yield zero scores for both precision and recall as both fail to return any *TP*.

Despite not being completely trivial, S1 is heavily under-splitting while S2 is the opposite. Both return one correctly segmented word for the sentences in both languages. Their corresponding recalls are therefore equal as *TP* is normalised by *RP*, which is hard-constrained by the references. However, adopting precision as the metric, S1 yields substantially higher scores as it returns much fewer *PP*. Referring to the trivial examples as well as the fact that only *TP* are meaningful to higher-level applications, S1 and S2 perform equally poorly, which is consistent with recall but not precision. Furthermore, a segmenter with less *TP* may achieve higher precision if it is drastically under-segmenting, as demonstrated by the comparison between S1 and S3.

2.2 True Negative Rate

Neither recall nor precision measure how well the system rejects the negatives. True negative rate (*TNR*) is therefore proposed by Powers (2011) as the complement. Jiang et al. (2011) show that segmenters measured by *TNR* are better correlated than precision and recall with their actual performances within IR systems. For WS, it is not straightforward to compute *TNR* by directly normalising the true negatives (*TN*) by the real negatives (*RN*). However, it can be indirectly computed via *TP*, *PP*, *RP* and the total number of possible output *TW* given a sentence. Regarding the input characters as a string, *TW* is equal to the number of substrings as $\frac{(1+N)N}{2}$, where *N* is the number of the characters. *RN* can then be computed by subtracting *RP*, the number of words in reference. The false negatives (*FN*) generated by the segmenter can be obtained by subtracting *TP* from *PP*, total number of words return by the segmenter. To put everything together:

$$TNR = \frac{TN}{RN} = 1 - \frac{FN}{RN} = 1 - \frac{PP - TP}{TW - RP} \quad (1)$$

When *PP* equals *TP*, we will have a *TNR* of 1, indicating that a WS system correctly rejects all *TN* if and only if all the *PP* are *TP*. Since *TW* is bounded by the input sentence length and *RP* is bounded by the reference, *TNR* is negatively correlated to *PP* as longer segmented word eliminates more *TN* and generates less *FN* in general. As shown in Table 1, *TNR* heavily favours under-splitting systems. T2 obtains the highest *TNR* in the table despite being trivial. S1 also ob-

tains higher scores than S3, despite having lower *TP*. Overall, *TNR* is very insensitive and not always well-correlated to actual performances of segmenters.

2.3 Boundary-Based Evaluation

Instead of directly evaluating the performance in terms of *TP* at word-level, an alternative is to use boundary-based evaluation (Palmer and Burger, 1997). The drawback is that incorrectly segmented words that are not interesting to higher-level applications still contribute to the scores as long as one of the two associated boundaries is correctly detected.

3 Experiments

To further investigate the correlations and drawbacks of the metrics discussed in the previous section experimentally, we employ a neural-based word segmenter to see how they measure the segmentation performance in a real scenario. The segmenter is a simplified version of the joint segmentation and POS tagger introduced in Shao et al. (2017). It is fully character-based. The vector representations of input characters are passed to the prevalent bidirectional recurrent neural network equipped with gated recurrent unit (GRU) (Cho et al., 2014) as the basic cell. A time-wise softmax layer is added as the inference for the recurrent layers to obtain probability distribution of binary tags that indicate the boundaries of the words. Cross-entropy with respect to time step is applied as the loss function. We train the segmenter for 30 epochs and pick the weights of the best epoch that minimises the loss on the development set.

The Chinese and English sections of Universal Dependencies v2.0 are employed as the experimental data sets. We follow the conventional splits of the data sets. For Chinese, the concatenated trigram model in Shao et al. (2017) is applied. Table 2 shows the experimental results on the test sets in terms of different metrics using the standard argmax function to obtain the final output. The segmenter is relatively under-splitting for Chinese as it yields higher recall than precision, which is opposite to English. The segmenter nonetheless achieves state-of-the-art performance on both languages.[1]

[1] http://universaldependencies.org/conll17/results-words.html

	P	R	F	TNR
Chinese	92.85	93.46	93.16	99.81
English	99.33	99.09	99.21	99.99

Table 2: Evaluation scores on the test sets in precision (P), recall (R), F1-score (F) and true negative rate (TNR).

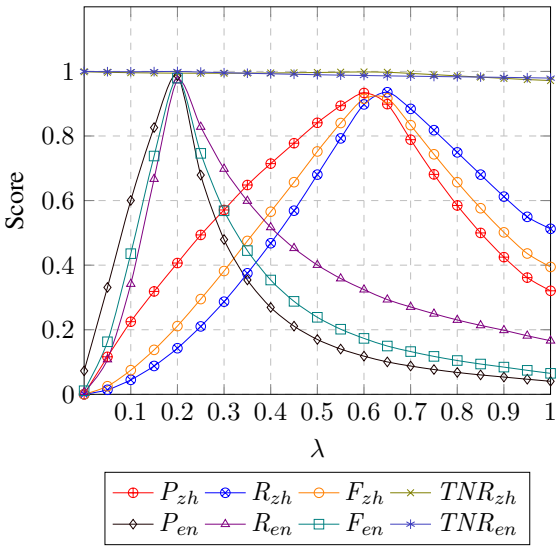

Figure 1: Evaluation scores on Chinese (zh) and English (en) in precision (P), recall (P), F1-score (F) and true negative rate (TNR) with different ratios of most probable boundaries λ.

To get a more fine-grained picture, instead of using argmax when decoding, we manually set a threshold to determine the word boundaries with respect to the scores returned by the inference layer of the neural network. All the possible output tags are ranked according to their scores of being a word boundary. For each test experiment, we accept the $\lambda * 100$ percent most probable word boundaries and regard the rest as non-word boundaries. The segmenter therefore tends towards under-splitting when λ is closer to 0 and over-splitting when λ is closer to 1. The segmenter becomes trivial when λ is equal to 0 or 1, corresponding to the extreme under-splitting and over-splitting segmenters T1 and T2 introduced in Table 1 respectively.

Figure 1 presents the evaluation scores according to the metrics under consideration with respect to different λ in the interval of 0.05. With the optimal λ_F^*, the segmenter achieves comparable F1-scores to those reported in Table 2. For Chinese, λ_F^* is around 0.6, indicating there are roughly 60% true boundaries out of all the possible segmenta-

tion points between consecutive characters. For English, λ_F^* is 0.2 as the fact that English words are relatively more coarse-grained and composed of more characters on average. In general, precision and recall are positively correlated. When λ is close to its the optimal, the values of both precision and recall increase. However, when λ is far away from both the optima and 0, precision and recall vary very substantially, clearly indicating that precision heavily favours under-splitting systems.

When λ equals 0, we obtain near-zero scores with trivial under-splitting. In contrast, the over-splitting segmenter with λ is equal to 1 yields a notable amount of true positives, due to the fact that there is a considerable amount of single-character words, especially in Chinese. This implies that actually trivial over-splitting is relatively better than under-splitting in practise, even though it is not favoured by precision.

For Chinese, the optimal λ_P^* for precision is 0.6, whereas λ_R^* for recall is 0.65. They would be different for English as well if a smaller interval of λ were adopted. λ_R^* corresponds to the system with most correctly segmented words, whereas λ_P^* is slightly biased towards under-splitting systems. The difference between λ_P^* and λ_R^* is marginal only when the segmenter performs very well as in the case of English.

Next, we investigate how the metrics behave in a learning curve experiment with ordinary argmax decoding. Instead of using the complete training set, for each test experiment, a controlled number of sentences are used for training the segmenter. The results are shown in Figure 2, in which the training set is extended gradually by 200 sentences. As expected, the segmenter is better trained and more accurate with a larger training set, which is in accordance with recall as it always increases when the training set is expanded. However, despite being closely correlated with recall in general, precision notably drops for Chinese when enlarging the train set from 800 to 1,000 as well as from 1,800 to 2,000, implying the segmenter becomes relatively over-splitting and obtains lower precision despite having more correctly segmented words. Similarly for English, the precision decreases when the training set is enlarged from 1,200 to 1,400.

The experimental results of *TNR* is also consistent with our analysis in the previous section. In WS, the values of both *RN* in the reference as well as *PN* by the system are drastically greater than the corresponding values of the positives. Thus, *TN* is high regardless of how the segmenter performs, which makes *TNR* very insensitive and inappropriate as an evaluation metric for WS.

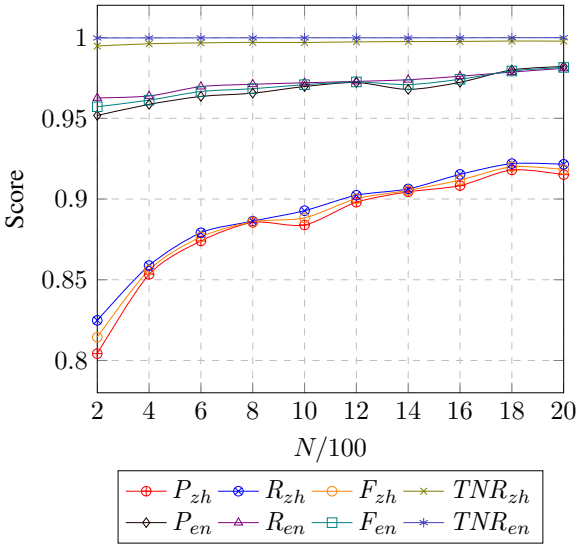

Figure 2: Evaluation scores on Chinese (zh) and English (en) in precision (P), recall (P), F1-score (F) and true negative rate (TNR) with different numbers of training instances *N*.

4 Conclusion

We discuss and analyse precision, recall in addition to true negative rate (*TNR*) as the evaluation metrics for WS both theoretically and experimentally in this paper. Unlike standard evaluation for IR, all the metrics are positively correlated in general. It is non-trivial to optimise the segmenter towards either precision or recall. The difference between precision and recall is notable only if the segmenter is strongly over- or under-splitting. In either case, precision as the evaluation is misleading as it heavily favours under-splitting systems. Additionally, *TNR* is very insensitive and not suitable to evaluate WS either.

Under the circumstances, we propose that precision should be excluded from the conventional evaluation metrics. As opposed to precision, recall is hard-constrained by the reference and therefore not biased towards neither under-splitting nor over-splitting systems. It explicitly measures the correctly segmented words that are meaningful to higher level tasks. Employing recall solely is therefore sufficient and more adequate as the evaluation metric for WS.

Acknowledgments

We acknowledge the computational resources provided by CSC in Helsinki and Sigma2 in Oslo through NeIC-NLPL (www.nlpl.eu). This work is supported by the Chinese Scholarship Council (CSC) (No. 201407930015).

References

Kyunghyun Cho, Bart Van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the properties of neural machine translation: Encoder-decoder approaches. *arXiv preprint arXiv:1409.1259* .

Sharon Goldwater, Thomas L Griffiths, Mark Johnson, et al. 2007. Distributional cues to word boundaries: Context is important. In *Proceedings of the 31st Annual Boston University Conference on Language Development*. pages 239–250.

Mike Tian-Jian Jiang, Cheng-Wei Shih, Richard Tzong-Han Tsai, and Wen-Lian Hsu. 2011. Evaluation via negativa of Chinese word segmentation for information retrieval. In *Proceedings of the 25th Pacific Asia Conference on Language, Information and Computation*. Institute of Digital Enhancement of Cognitive Processing, Waseda University, Singapore, pages 100–109.

Allen Kent, Madeline M Berry, Fred U Luehrs, and James W Perry. 1955. Machine literature searching VIII. operational criteria for designing information retrieval systems. *Journal of the Association for Information Science and Technology* 6(2):93–101.

David Palmer and John Burger. 1997. Chinese word segmentation and information retrieval. In *AAAI Spring Symposium on Cross-Language Text and Speech Retrieval*. pages 175–178.

David Martin Powers. 2011. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation .

Xipeng Qiu, Peng Qian, Liusong Yin, Shiyu Wu, and Xuanjing Huang. 2015. Overview of the NLPCC 2015 shared task: Chinese word segmentation and POS tagging for micro-blog texts. In *National CCF Conference on Natural Language Processing and Chinese Computing*. Springer, pages 541–549.

Yan Shao, Christian Hardmeier, Jörg Tiedemann, and Joakim Nivre. 2017. Character-based joint segmentation and POS tagging for Chinese using bidirectional RNN-CRF. *arXiv preprint arXiv:1704.01314* .

Richard Sproat and Thomas Emerson. 2003. The first international Chinese word segmentation bakeoff. In *Proceedings of the second SIGHAN workshop on Chinese language processing-Volume 17*. Association for Computational Linguistics, pages 133–143.