Evaluation *via Negativa* of Chinese Word Segmentation for Information Retrieval*

Mike Tian-Jian Jiang[ac], Cheng-Wei Shih[bc], Chan-Hung Kuo[c], Richard Tzong-Han Tsai[d], and Wen-Lian Hsu[abc]

[a] Department of Computer Science, National Tsing Hua University,
Hsinchu, Taiwan
[b] Institute of Information Systems and Application, National Tsing Hua University,
Hsinchu, Taiwan
[c] Institute of Information Science, Academia Sinica,
Taipei, Taiwan
{tmjanng, dapi, laybow, hsu}@iis.sinica.edu.tw
[d] Department of Computer Science & Engineering, Yuan Ze University,
Taoyuan, Taiwan
thtsai@saturn.yzu.edu.te

Abstract. Numerous studies have analyzed the influences of word segmentation (WS) performance on information retrieval (IR) for Mandarin Chinese and have demonstrated a non-monotonic relationship between WS accuracy and IR effectiveness. The usefulness of the compound words that have been a focus of the IR literature is not reflected by common WS evaluation metrics of word-based precision (P) and recall (R). This investigation proposes alternative measurements of WS accuracy, which are based on negative segments that are annotated against four standards of referenced corpora, called true negative rate (TNR) and negative predictive value (NPV), and compares with P and R through search engine simulation,. Accuracy-controlled WS systems segment queries for the simulation including NTCIR collections and "Sogou" logs. Mean average precision (MAP) estimates the similarity of search results between the original and segmented queries. The statistics demonstrate that TNR and NPV are generally more closely correlated with MAP than are P and R.

Keywords: Word segmentation, information retrieval, true negative rate.

1 Introduction

Word segmentation (WS) is an essential preparatory work for Chinese text processing applications and consequently has become a focus of research during the past two decades. Numerous researchers have urged the application of various algorithms to produce "accurate" tokenized text in which the segmentations are performed to match one of the standards of Chinese corpora, such as Academia Sinica Balanced Corpus (Huang *et al.*, 1997) and Chinese Treebank of University of Pennsylvania (Xia, 1999), and evaluated using various metrics, such as word-based precision (P), recall (R) and their harmonic average, termed F_1 measure score (F-score), popularized by SIGHAN Chinese WS bakeoffs (Sproat and Emerson, 2003; Emerson, 2005; Levow, 2006; Jin and Chen, 2007; Zhao and Liu, 2010). Subsequently, WS systems that

* This research was supported in part by the National Science Council under grant NSC 100-2631-S-001-001, and the research center for Humanities and Social Sciences under grant IIS-50-23. Ted Kony is appreciated for his editorial assistance. The authors would like to thank anonymous reviewers for their constructive criticisms.

Copyright 2011 by Mike Tian-Jian Jiang, Cheng-Wei Shih, Chan-Hung Kuo, Richard Tzong-Han Tsai, and Wen-Lian Hsu

perform strongly in terms of F-score have been considered reliable and are selected as the preprocessors for the commonly accepted notion that WS accuracy is directly related to the effectiveness of follow-up applications. However, this perspective has sometimes failed to win significant support, particularly in the Chinese IR area, because of the absence of strong evidence confirming the causation. As described later in Section 2.3, some previous studies demonstrated a non-monotonic relationship between the performances of WS and IR for Mandarin Chinese, implying that better WS results do not necessarily yield better IR outcomes (Foo and Li, 2004; Kwok, 2000; Peng *et al.*, 2002), while other studies have shown the opposite (He *et al.*, 2002; Nie *et al.*, 2000; Palmer and Burger, 1997). Although these investigations provide valuable information, they remain unable to reach consensus regarding the principles or metrics involved in selecting a proper WS system for IR applications.

Developing a credible method of measuring Chinese WS system requires error analysis. However, the mainstream of WS literature overlooked how incorrect identification of segments based on particular standards affects certain type of applications, and instead focused on improving segmentation accuracy. Consequently, this study annotates negative segment (NS) against segmented corpora, which may provide clues to help select Chinese WS systems for different purposes such as IR. Following definition, the NS forms supplementary WS evaluation metrics intuitively as the concept that already existed in binary classification test: true negative rate (TNR) and negative predictive value (NPV). This investigation designs a simulation on a commercial search engine, which demonstrates how to apply TNR and NPV and why they could be useful for analyzing the influences of WS performance on IR.

The remainder of this paper is organized as follows. Section 2 briefly introduces the literature on IR and WS for Mandarin Chinese. Section 3 then describes the proposed WS evaluation metrics using qualitative analysis. Next, Section 4 summarizes the quantitative analytical results of the simulation, while Section 5 discusses these same results. Finally, Section 6 presents conclusions, along with recommendations for future research.

2 Related Works

2.1 Evaluation of Chinese Information Retrieval

For various open evaluation tasks of Chinese IR, the test collections of TREC-5, TREC-6, and TREC-9 comprise 28, 26, and 25 queries in simplified Chinese, respectively, while NTCIR-2 comprises 50 queries in traditional Chinese. One problem is that the number of queries is small. Another problem with the TREC data is that the Chinese queries (topic titles) have too many keywords. According to Xu *et al.* (2010), the Chinese queries have an average length of 12.2 words; in contrast, the average length of English ad-hoc queries in TREC (English topics 251-350) is 4.7 words. Even the length of Chinese queries is measured using English translations, the average length still exceeds 7 words. Long queries introduce complex effects whose interactions are difficult to understand. These problems are part of a more general issue of sample pooling bias (Webber and Park, 2009). Common metrics for evaluating IR effectiveness include precision of top-*k* retrieval results (P@*k*), mean reciprocal recall (MRR), mean average precision (MAP), etc. These metrics of English *ad hoc* retrieval may suffer score standardization issues (Webber *et al.*, 2008). Alternative metrics were proposed and discussed, including geometric mean average precision (Robertson, 2006) or other popularity-based measurements (Mizzaro, 2008; Yilmaz and Aslam, 2006), but have not been adopted widely.

2.2 Evaluation of Chinese Word Segmentation

This study classifies known evaluation metrics of WS into three categories: boundary-based, token-based and constituent-based. Boundary recall is defined as the percentage of correct boundaries identified, while boundary precision is defined as the percentage of identified boundaries that are correct, discounting utterance boundaries (Fleck, 2008; Goldwater *et al.*, 2009; Palmer and Burger, 1997). Token-based recall (R) and precision (P) are defined analogously to their boundary-based counterparts, except the unit of measurement is tokens rather than boundaries, where tokens can be treated as morpheme instances (Ando and Lee, 2003), lexicon types (Fleck, 2008; Goldwater *et al*., 2009), or word instances. For example, given an utterance "[[meta][data]] / is / the / data / of / data" as standard in five boundaries, seven morphemes (notice the case of [[meta][data]]), six words, and five lexicon types, one possible segmentation "meta / data / is / the / data / of / data" results in one boundary error, zero morpheme error, two word errors, and one lexicon type error. Ando and Lee (2003) proposed a constituent-based metric called compatible bracket error, which counts reasonable combinations of morphemes, estimate "[meta][data]" and "[metadata]" as interchangeable. Previous studies also counted word error rates (Chiang *et al*., 1992; Teahan *et al*. 2000; Wong and Chan, 1996) and sentence accuracy (Chiang *et al*., 1992) as complementary metrics of word-based recall and precision. Several other constituent-based metrics exist, and are usually close related to IR. For example, Liu *et al*. (2008) proposed an evaluation measure "RankPrecision" based on Kendall-tau distance, which compared the similarity between the predicted rankings of "Internal Association Strength (IAS)" and the ideally sorted rankings of IAS in descending order. Methodologies resemble IAS, which may be seen as constituent-based measurements, are "Phrase Inseparability" (Shi and Nie, 2009) and the "Tightness Continuum Measure" (Xu *et al*. 2010).

2.3 WS-to-IR Performance Relationship

The influence of different Chinese WS methods on IR has attracted extensive research attention (Foo and Li, 2004; He *et al*. 2002; Kwok, 2000; Liu *et al*., 2008; Luk and Kwok, 2002; Nie *et al*., 2000; Palmer and Burger, 1997; Peng *et al*., 2002; Oard and Wang, 1999). For example, Foo and Li (2004) tested the effects of manual segmentation and various character-based segmentations and provided detailed analysis of query results. Foo and Li argued that a small test collection reduces the significance of the analytical results regarding the deduction of causal relation from WS accuracy to IR effectiveness. This argument might be considered as a Chinese-specific example of the issue of sample pooling bias mentioned in Section 2.1. Peng *et al*. (2002) compared various Chinese WS methods in IR by systematically examining retrieval effectiveness for different levels of accuracy of the word division. Experimental results indicated a non-monotonic relationship between Chinese WS accuracy (in terms of F-score) and retrieval effectiveness (in terms of average precision and R-precision), where retrieval performance increases steadily for WS accuracies between 44% and 70%, plateaus for WS accuracies between 70% and 77%, and finally decreases slightly for WS accuracies between 85% and 95%. The hypothesis is that although some Chinese WS systems tend to break compound words into smaller constituents, they cannot achieve the top F-score, but can improve the effectiveness of Chinese IR at some level. Kwok (2000) focused on various segmentation methods of query processing to obtain diverse IR results, while Luk and Kwok (2002) verified the influence of different segmentation approaches on document indexing. Kwok, sees more accurate segmentation as generally improving retrieval when content-bearing terms are involved, though exceptional cases exist that may average out the results, particularly for long queries where sufficient redundancies exist to remedy incorrect segmentations of short queries. However, the definitions for term types such as content-bearing and query lengths are unclear. Kwok also hypothesized that individual WS systems probably do not differ sufficiently in terms of accuracy to realize statistically significant differences in IR effectiveness in terms of P@*k*, MRR or MAP. Kwok thus suggested that a very high quality WS system, perhaps exceeding 95% in terms of both R and P, together with specially selected queries for query length and term type would display a different story. Subsequently, He *et al*. (2002) presented a different story, namely that a highly accurate WS system improves IR effectiveness, leaving the phenomenon of a non-monotonic performance relationship between WS and IR inconclusive. Additionally, those works do not apply the same evaluation methods, term types, or query lengths, thus and the relationship between WS and IR performance currently lacks the commensurability required to achieve a consensus among literature mentioned in this subsection.

3 Qualitative Analysis

3.1 Error Case Study

Chinese IR systems frequently employ WS systems as preprocessing tools. WS errors thus may propagate to follow-up Chinese IR systems. Kwok (2002) has investigated this issue and classified relationship between IR effectiveness and segmented terms for query or indexing into three categories. The first is insensitive: segmented terms whose errors do not influence IR effectiveness fall into this category. Most terms in this category are highly frequent and non-content-bearing terms, which are generally treated as stop-words and eventually removed in IR process. The second is monotonic: segmented terms in this category are high frequency content-bearing terms. WS accuracy is monotonically related to IR effectiveness. The third is non-monotonic. As the category named, IR effectiveness is not monotonically related to the accuracy of segmented terms in this category, which can be further divided into three sub-categories: 1) segmented terms are infrequent non-content-bearing terms that are segmented correctly and retained after stop-word removal, and that then may decrease IR effectiveness; 2) segmented terms that are semantically correct but do not help retrievals; 3) segmented terms are foreign names. Take the query phrase “1999年西土耳其地震” (the 1999 Western Turkey Earthquake) for example, where “西土耳其” (Western Turkey) is the primary subject, which belongs to both the second and the third sub-classes. Correctly segmenting “西土耳其” (Western Turkey) as one term were observed that does not guarantee good IR effectiveness, because “西土耳其” is far less frequent than “土耳其” (Turkey). A particular observation is that “西土耳其” may not appear in some documents directly. Instead, “土耳其西部” (western part of Turkey) or “土耳其西部省分” (western provinces of Turkey) are mentioned in these documents that cannot be retrieved using the correctly segmented term “西土耳其”.

Analysis of the above involves semantic and syntactical structures behind surface patterns of words, yet introduces potential debate regarding definitions such as “content-bearing term” and requires more complicated procedures such as sentence parsing that lie beyond the ability of simple WS systems. For example, “农作物” (agricultural plants) is segmented as two words “农” (agriculture) and “作物” (plants) and could be useful for recalling more relevant retrieval results, but is usually seen as an error from the perspective of WS accuracy. A more surprising case is “旱灾” (drought disaster). While one WS system segments it incorrectly into “旱” (drought) and “灾” (disaster) by following a certain standard, some IR systems actually welcome such queries because they have a better chance of recalling loosely relevant results such as “春旱” (Spring drought disaster) or “旱区” (area of drought disaster). In this case, surprises might result from semantic relations such as synonyms or hyponyms, which are frequently considered query expansions of IR rather than the duty of WS. Furthermore, the trade-off between IR recall and precision is also an issue. For instance, a single WS system may not recognize “皮纳图博火山” (Mount Minatubo) as a word and the resultant segmentation could be either “皮/纳/图/博/火山” as five words with “火山” (volcano) as one recognized word or simply “皮/纳/图/博/火/山” as six single-character words. These inaccurate segmentations in terms of WS performance evaluation metrics generally lead to the effective reduction of IR in terms of MAP. However, other studies may interpret similar segmentations as unexpected positive influences on IR in terms of MRR.

This study thus examines errors involving surface patterns rather than deep structures. Examples of the above imply that IR systems prefer segmentations as simple as containing some single-character words, but not as simple as consisting only of single-character words. This particular behavior could be visible in preferences regarding evaluation metrics, which leads this study to investigate properties of word-based WS evaluation metrics in Section 3.2.

3.2 Properties of Word-based WS Evaluation Metrics

Main WS events like the SIGHAN Chinese WS bakeoff traditionally apply word-based evaluation metrics such as recall (R), precision (P), and F_1-score (F) to both in-vocabulary and out-of-vocabulary data to measure participating system performance. Although these measurements are intuitive and uncomplicated, they still suffer the weakness of not considering incorrect segments that variously influence applications such as web search or text retrieval. This section presents qualitative analysis between word-based WS evaluation metrics and the proposed metrics. Notably, this study follows the naming convention proposed by Gao *et al.* (2005) and uses "segmentation unit" (segment in short, hereafter) rather than "word" as the conceptual element of qualitative analysis.

An input sequence "XYX" and its four possible segmentation gold standards G_1, G_2, G_3 and G_4 are "X/Y/X," "X/YX," "XY/X" and "XYX", respectively. Furthermore, four different systems S_1, S_2, S_3, and S_4 output "X/Y/X," "X/YX," "XY/X," and "XYX," respectively. Table 1 lists their recall (R) and precision (P).

Table 1: Performance Evaluation Samples of SIGHAN's Metrics

	S_1: X/Y/X		S_2: X/YX		S_3: XY/X		S_4: XYX	
	R	P	R	P	R	P	R	P
G_1: X/Y/X	3/3	3/3	1/3	1/2	1/3	1/2	0/3	0/1
G_2: X/YX	1/2	1/3	2/2	2/2	0/2	0/2	0/2	0/1
G_3: XY/X	1/2	1/3	0/2	0/2	2/2	2/2	0/2	0/1
G_4: XYX	0/1	0/3	0/1	0/2	0/1	0/2	1/1	1/1

One notable phenomenon listed in Table 1 is that the weighting of partial credits increases with the number of single-character segments in a system, which may mislead applications developers. The output of incorrect segments by different systems may influence the following applications such as IR with different degrees that P and R cannot reflect. For example, "施 / 政 / 偉" (*shih* / *zheng* / *wei*) and "施政 / 偉" (*shih-zheng*; *practice policy* / *wei*; *great*) are both treated as incorrect segments, but the former decreases IR effectiveness less than the latter, because the former only interprets "施政偉" vaguely, while the latter improperly alters the meaning of "施政偉" through its substring "施政." To reveal complementary facts besides R and P of a Chinese WS system, assuming a dictionary D contains segments {X, Y, XY, YX, XYX}, for an input sequence "XYX" in question and its four possible gold standards {X/Y/X, X/YX, XY/X, XYX}, namely G_1, G_2, G_3 and G_4, respectively, their corresponding positive segments (PS) and negative segments (NS) would be

- $PS_{G1} = \{X_1, Y_2, X_3\}$ and $NS_{G1} = \{XY_1, YX_2, XYX_1\}$;
- $PS_{G2} = \{X_1, YX_2\}$ and $NS_{G2} = \{X_2, X_3, XY_1, XYX_1\}$;
- $PS_{G3} = \{XY_1, X_3\}$ and $NS_{G3} = \{X_1, Y_2, YX_2, XYX_1\}$;
- $PS_{G4} = \{XYX_1\}$ and $NS_{G4} = \{X_1, Y_2, X_3, XY_1, YX_2\}$,

where digital subscriptions for segments of the above (such as $_1$ for X_1) denote the beginning position of that segment in the original input sequence. The general definitions of PS and NS are as follows.

- PS: For all segments a_i of a predefined dictionary D, a_i appears at position i of the input sequence and is identical to the segment at position i of the corresponding sequence from a particular gold standard.
- NS: For all segments a_j of a predefined dictionary D, a_j appears at position j of the input sequence and is **not** the same as the segment at position j of the corresponding sequence from a particular gold standard.

When a given gold standard of PS_G and NS_G compared with PS_S and NS_S obtained from a certain Chinese WS system, four true or false predictions form, as shown in Table 2, in the context of binary classification tasks. PS_G and PS_S symmetrically generate recall (also known as TPR) based on $|TP| / |PS_G|$ and precision (also known as PPV) based on $|TP| / |PS_S|$, yet since NS is defined, true negative rate (TNR) and negative predictive value (NPV) exist as $|TN| / |NS_G|$ and $|TN| / |NS_S|$, respectively. For consistency and readability, hereafter this study uses TPR and PPV rather than word-based recall and precision of WS evaluation, respectively. Using the same

Table 2: The Four Situations of Chinese WS as Binary Classification

	PS_S	NS_S	
PS_G	True positive (TP)	False negative (FN)	True positive rate (TPR) = $\|TP\| / \|PS_G\|$
NS_G	False positive (FP)	True negative (TN)	True negative rate (TNR) = $\|TN\| / \|NS_G\|$
	Positive predictive value (PPV) = $\|TP\| / \|PS_S\|$	Negative predictive value (NPV) = $\|TN\| / \|NS_S\|$	

Table 3: Performance Evaluation Samples of TNR and NPV

	S_1: X/Y/X		S_2: X/YX		S_3: XY/X		S_4: XYX	
	TNR	NPV	TNR	NPV	TNR	NPV	TNR	NPV
G_1: X/Y/X	3/3	3/3	2/3	2/3	2/3	2/3	2/3	2/5
G_2: X/YX	2/4	2/3	3/3	3/3	1/3	1/3	2/3	2/5
G_3: XY/X	2/4	2/3	1/3	1/3	3/3	3/3	2/3	2/5
G_4: XYX	2/5	2/3	2/5	2/3	2/5	2/3	5/5	5/5

assumptions as Table 1 and with the help of definitions of PS and NS, Table 3 lists simulated statistics of TNR and NPV.

For example, system S_1 segments "XYX" as "X/Y/X", suggesting its NS_{S1} are $\{XY_1, YX_2, XYX_1\}$. Comparing system S_1 with the gold standard G_4 "XYX" and NS_{G4} as $\{X_1, Y_2, X_3, XY_1, YX_2\}$, TN_{S1} is $\{XY_1, YX_2\}$, and thus TNR is $|TN_{S1}| / |NS_{G4}| = 2 / 5$ and NPV is $|TN_{S1}| / |NS_{S1}| = 2 / 3$. Simultaneously considering Table 1 and Table 3, TNR and NPV clearly balance some trends of TPR and PPV. For instance, according to the gold standard G_4, Chinese WS system S_1 gains no credits in TPR and PPV, yet obtains non-zero scores in TNR and NPV.

4 Quantitative Analysis

Reasonably good IR evaluation requires controlling an excessive number of variables. On the one hand, model indexing/retrieving involves numerous techniques such as stemming, stop-word elimination, word-based indexing, character-based n-gram indexing, hybrid indexing, vector space modeling, probabilistic modeling, etc. On the other hand, not only do test collections for ad-hoc retrieval have some unbalanced characteristics, such as mostly being involved in long queries, but choices must also be made between different IR evaluation metrics and their potential bias. This investigation respects these research difficulties and proposes a temporary solution based on popularity. Considering a real world search engine widely used in daily IR, indexing/retrieving approaches are considered a black box that is likely to be a reasonably good (and probably commercial confidential) set of treatments that includes all useful techniques. Each query and its ranked results based on this black box are paired as test collections. Original queries from different sources are collected to obtain a balanced simulation regarding term length and type. Different query representations are prepared by segmenting original queries using accuracy-controlled WS systems. Similarities in query results between original and correspondingly segmented queries should provide a series of numbers indicating the degree of how closely the segmented queries match the preferences of the black box. Calculating Pearson product-moment correlation coefficients between query result similarities and WS performance in several evaluation metrics can obtain a clearer picture of the influence of WS on IR effectiveness.

4.1 Query Collections

Owing to the sample pooling bias issues mentioned in Section 2.1, this study collected queries from two sources in the hope of obtaining a balanced samples: test sets of NTCIR CLIR 3, 4, 5 and 6 (NTCIR query as in short, hereafter) comprise 197 traditional Chinese queries with average length of 9.38 characters that may contains multiple query strings in a single query, representing academic samples of queries; meanwhile, query logs arranged by a commercial search engine company “Sogou” (Sogou query as in short, hereafter), written in simplified Chinese, comprising 432 consecutive query strings with an average length of 4.36 characters, provide a practical view of end users.

4.2 Accuracy-controlled WS Systems

This study implements several WS systems based on the-state-of-the-art approach that uses conditional random fields with a 6-tag labeling scheme in bi-directional unigram, bi-gram and pair contexts (Zhao *et al.*, 2006). Four gold standards involving Chinese WS corpus from SIGHAN 2005 WS bakeoffs, including Academia Sinica (AS), City University of Hong Kong (CityU), Microsoft Research (MSR), and Peking University (PKU), are adopted as training data (cf. footnote). By randomly dividing the training data into exponentially smaller parts ranging from 1/2 to 1/16384, WS systems trained using corresponding parts perform with linearly decreasing accuracy in TPR and PPV, just as demonstrated by Zhao *et al*. (2010).

4.3 Simulation Design

For traditional Chinese, NTCIR queries are segmented by the accuracy-controlled WS systems AS and CityU. For simplified Chinese, Sogou queries are segmented by the accuracy-controlled WS systems MSR and PKU. The average segment numbers of segmented NTCIR and Sogou queries are 5.8 and 3.1, respectively. All the segmented tokens of a query are quoted, adhered by white space, and formed a segmented query. The original query and all the segmented queries from different WS systems are forwarded to Google to retrieve the top-100 search results. Search results of original queries are considered as the gold standards of evaluating the consistency of querying popularity between original and segmented queries. The consistency evaluation result is measured by using mean average precision (MAP).

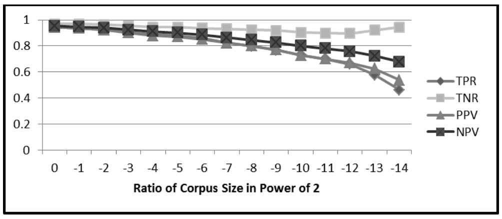

Figure 1: Performances of WS systems trained by different sizes of AS corpus.

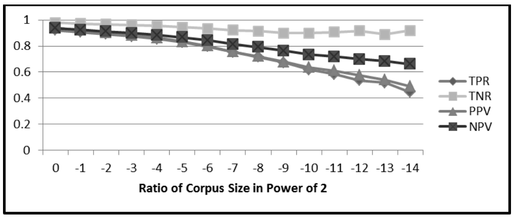

Figure 2: Performances of WS systems trained by different sizes of CityU corpus.

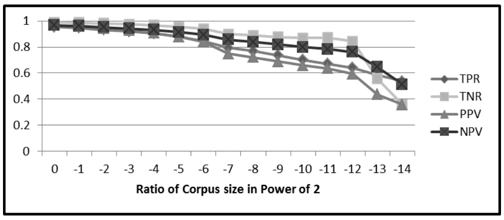

Figure 3: Performances of WS systems trained by different sizes of MSR corpus.

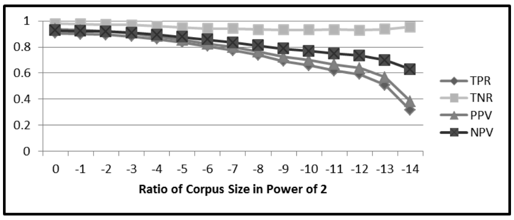

Figure 4: Performances of WS systems trained by different sizes of PKU corpus.

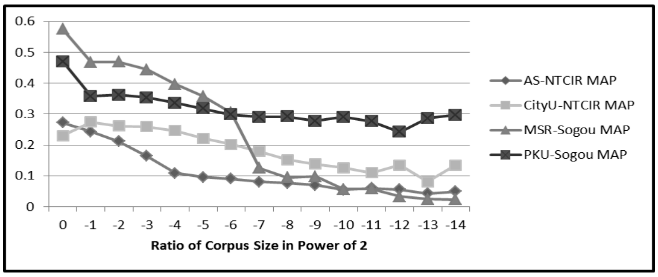

Figure 5: MAP of retrieval results of NTCIR/Sogou queries segmented by different WS systems.

4.4 Result of the Simulation

Figures 1 to 4 show the performances of accuracy-controlled WS systems in terms of TPR, TNR, PPV, and NPV. The MAP values of the search results retrieved by the original and segmented queries are illustrated in Figure 5. Finally, Table 4 compares correlation coefficients.

Table 4: Correlation between search similarities and WS metrics

	AS	CityU	MSR	PKU
TPR	0.742	0.916	0.933	0.615
TNR	0.836	0.944	0.649	0.870
PPV	0.788	0.931	0.901	0.634
NPV	0.800	0.940	0.819	0.692

5 Discussion

Table 4 shows that TNR and NPV are more closely correlated than TPR and PPV, except in the case of MSR. Notably, since Corpus quality assurance process (Sun *et al*., 2005) and analysis for out-of-vocabulary issue (Li *et al*., 2005) of MSR use similar definitions of NS, they provide a good demonstration that considering both the PS and NS oriented metric can identify differences between WS systems trained using different standards of corpora. For example, a Sogou query "上海滩" (*shang-hai-tan; The bund of Shanghai*) were segmented into "上海滩," "上海 / 滩" (*Shanghai* / *bund*), or "上 / 海 / 滩" (*up* / *sea* / *bund*) using the accuracy-controlled system MSR, while the same query were segmented into "上海滩," "上海 / 滩," or "上 / 海滩" (*go to* / *beach*) using the accuracy-controlled system PKU. The TPR and PPV values generally do not significantly differentiate "上 / 海 / 滩" and "上 / 海滩", but the TNR and NPV values do.

6 Conclusions and Future Work

Chinese WS achieve competitive performances based on TPR and PPV, meanwhile numerous academic studies have analyzed the influence of WS performance on IR and yet failed to reach a verdict since some of them demonstrate that WS accuracy is non-monotonically related to IR effectiveness. The usefulness of compounds and constituents of interest to the literature on Chinese IR might not be reflected by common WS evaluation metrics. This study thus develops alternative WS performance measurements, TNR and NPV, by defining negative segments, and then observes their correlation with results obtained using a major search engine. The experiment design aims to provide quantitative comparisons of Chinese search engine results that retrieved using different segmented query representations. The original queries are collected from NTCIR CLIR test sets and "Sogou" search engine logs. Segmented queries are generated using accuracy-controlled WS systems. Mean average precision (MAP) scores of Chinese IR results yield similarities between original and segmented queries. The statistics demonstrate that when obtaining comparisons using MAP, Chinese WS scores in TNR and NPV provide a complementary trend of correlation coefficient to traditional evaluation metrics TPR and PPV.

While Sproat and Emerson (2003) mentioned that WS performance on short Chinese strings has been only studied indirectly via IR evaluation of TREC, this study could also contribute to the research aspect on short Chinese strings since Sogou queries fall into this category. Although this study appears to omit the out-of-vocabulary (OOV) issue, the intention is to provide an alternative perspective in the form of token-based metrics based on NS, which can easily be expanded to cover specific token type of OOV. Sproat and Emerson believe that WS evaluation with alternative segmentation is straightforward, and this study thus roughly describes an implementation to serve this purpose. Assuming the definition of PS is relaxed, such that for a named entity "施政偉" as a gold standard, segments "施 / 政偉" and "施 / 政 / 偉" are also acceptable, the variations of NS are reduced to "施政 / 偉". Under this circumstance, a silver standard of WS can be developed using only highly unacceptable NS annotated in the corpus. Evaluation metrics tied up with this silver standard are naturally TNR and NPV, where traditional TPR and PPV can remain intact with the original gold standard that does not really include alternative segments of “施政偉”. Evaluations thus could be more feasible for the WS system with customizable granularity of named entity (Wu, 2003; Gao *et al*., 2005) or the concept of character combination (Dong *et al*., 2010). Hence, not only the OOV issue could be transformed into character combination validation, but the word-hood debate regarding Chinese was perceived from different angles, thus probably reducing less disagreement, *via negativa*.

References

Ando, R. K. and L. Lee. 2003. Mostly-Unsupervised Statistical Segmentation of Japanese Kanji Sequences. *Natural Language Engineering*, 9(2), 127-149.

Chiang, T.-H., J.-S. Chang, M.-Y. Lin and K.-Y. Su. 1992. Statistical Models for Word Segmentation and Unknown Word Resolution. *Proceedings of ROCLING V*, 121-146.

Dong, Z.-D., Q. Dong and C.-L. Hao. 2010. Word Segmentation Needs Change – From a Linguist’s View. *Proceedings of the First CIPS-SIGHAN Joint Conference on Chinese Language Processing*, 1-7.

Emerson, T. 2005. The Second International Chinese Word Segmentation Bakeoff. *Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing*, 123-133.

Fleck, M. M. 2008. Lexicalized Phonotactic Word Segmentation. *Proceedings of the 46th Annual Meeting of the ACL*. 130-138.

Foo, S. and H. Li. 2004. Chinese Word Segmentation and its Effect on Information Retrieval. *Information Processing & Management*, 40(1), 161-190.

Gao, J.-F, M. Li, A. Wu and C.-N. Huang. 2005. Chinese Word Segmentation and Named Entity Recognition: A Pragmatic Approach. *Computational Linguistics*, 31(4), 531-574.

Goldwater, S., T. L. Griffiths and M. Johnson. 2009. A Bayesian Framework for Word Segmentation: Exploring the Effects of Context. *Cognition*, 112(1), 21-54.

He, H.-Z., P.-L. He, J.-F. Gao and C.-N. Huang. 2002. Finding the Better Indexing Units for Chinese Information Retrieval. *Proceedings of the First SIGHAN Workshop on Chinese Language Processing*, 1-7.

Huang, C.-R., K.-J. Chen, F.-Y. Chen and L.-L. Chang. 1997. Segmentation Standard for Chinese Natural Language Processing. *International Journal of Computational Linguistics and Chinese Language Processing*, 2(2), 47-62.

Jin, G.-J. and X. Chen. 2007. The Fourth International Chinese Language Processing Bakeoff: Chinese Word Segmentation, Named Entity Recognition and Chinese POS Tagging. *Proceedings of the Sixth SIGHAN Workshop on Chinese Language Processing*, 69-81.

K. L. Kwok. 2000. Improving English and Chinese Ad-Hoc Retrieval: A Tipster Text Phase 3 Project Report. *Information Retrieval*, 3(4), 313-338.

Levow, G. A. The Third International Chinese Language Processing Bakeoff: Word Segmentation and Named Entity Recognition. *Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing*, 108-117.

Liu, Y.-X., B. Wang, F. Ding and S. Xu. 2008. Information Retrieval Oriented Word Segmentation based on Character Associative Strength Ranking. *Proceedings of the Conference on Empirical Methods in Natural Language Processing 2008*, 1061-1069.

Li, H.-Q., C.-N. Huang, J.-F. Gao and X.-Z. Fan. 2005. The Use of SVM for Chinese New Word Identification. *Proceedings of IJCNLP 2004*, 723-732.

Luk, R. W. P. and K. L. Kwok. 2002. A Comparison of Chinese Document Indexing Strategies and Retrieval Models. *ACM Transactions on Asian Language Information Processing*, 1(3), 225-268.

Mizzaro, S. 2008. The Good, the Bad, the Difficult, and the Easy: Something Wrong with Information Retrieval Evaluation? *Proceedings of the IR Research, 30th European Conference on Advances in Information Retrieval*, 642-646.

Nie, J.-Y., J.-F. Gao, J. Zhang and M. Zhou. 2000. On the Use of Words and N-grams for Chinese Information Retrieval. *Proceedings of the Fifth International Workshop on Information Retrieval with Asian Languages*, 141-148.

Oard, D. W. and J.-Q. Wang. 1999. Effects of Term Segmentation on Chinese/English Cross-Language Information Retrieval. *Proceedings of the String Processing and Information Retrieval Symposium & International Workshop on Groupware*, 149-157

Palmer, D. D. and J. D. Burger. 1997. Chinese Word Segmentation and Information Retrieval, *AAAI Spring Symposium on Cross-Language Text and Speech Retrieval*.

Peng, F. C., X.-J. Huang, D. Schuurmans and N. Cercone. 2002. Investigating the Relationship between Word Segmentation Performance and Retrieval Performance in Chinese IR. *Proceedings of the 19th International Conference on Computational Linguistics*, 1-7.

Robertson, S. 2006. On GMAP – and Other Transformations. Proceedings of the 15th ACM International Conference on Information and Knowledge Management, 78-83.

Shi, L. and J.-Y. Nie. 2009. Integrating Phrase Inseparability in Phrase-Based Model. Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval, 708-709.

Sproat, R. and T. Emerson. 2003. The First International Chinese Word Segmentation Bakeoff. *Proceedings of the Second SIGHAN Workshop on Chinese Language Processing*, 113- 143.

Sun, C.-J., C.-N. Huang, X.-L. Wang and M. Li. 2005. Detecting Segmentation Errors in Chinese Annotated Corpus. *Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing*.

Teahan, W. J., Y.-Y. Wen, R. McNab and I. H. Witten. 2000. A Compression-based Algorithm for Chinese Word Segmentation. *Computational Linguistics*, 26(3), 375-393.

Webber, W., A. Moffat and J. Zobel. 2008. Score Standardization for Inter-Collection Comparison of Retrieval Systems. *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 51-58.

Webber, W. and L. A. F. Park. 2009. Score Adjustment for Correction of Pooling Bias. Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval, 444-451.

Wong, P. and C. Chan. 1996. Chinese Word Segmentation Based on Maximum Matching and Word Binding Force. *Proceedings of the 16th Conference on Computational Linguistics*, 200-203.

Wu, A. 2003. Customizable Segmentation of Morphologically Derived Words in Chinese. *International Journal of Computational Linguistics and Chinese Language Processing*, 8(1), 1-27.

Xia, F. 1999. *Segmentation Guideline, Chinese Treebank Project*. Technical report, University of Pennsylvania.

Xu, Y., R. Goebel, C. Ringlstetter and G Kondrak. 2010. Application of the Tightness Continuum Measure to Chinese Information Retrieval. *Proceedings of the 23rd International Conference on Computational Linguistics*, 55-63.

Yilmaz, E. and J. A. Aslam. 2006. Estimating Average Precision with Incomplete and Imperfect Judgments. *Proceedings of the 15th ACM International Conference on Information and Knowledge Management*, 102-111.

Zhao, H, Y. Song and C.-Y. Kit. 2010. How Large a Corpus Do We Need: Statistical Method Versus Rule-based Method. *Proceedings of the Seventh conference on International Language Resources and Evaluation*.

Zhao, H, C.-N. Huang and M. Li. 2006. An Improved Chinese Word Segmentation System with Conditional Random Field. *Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing*, 162-165.

Zhao, H.-M. and Q. Liu. 2010. The CIPS-SIGHAN CLP2010 Chinese Word Segmentation Backoff. *Proceedings of the First CIPS-SIGHAN Joint Conference on Chinese Language Processing*, 199-209.