Text segmentation and Chinese site search

Liyuan Zhou
NICTA & ANU
lizhou@nicta.com.au

David Hawking
Microsoft & ANU
david.hawking@acm.org

Paul Thomas
CSIRO & ANU
paul.thomas@csiro.au

ABSTRACT

Automatic segmentation and overlapping bigrams are the most common methods for overcoming the lack of explicit word boundaries in Chinese text. Past studies have compared their effectiveness, but findings have been equivocal and site search has been little studied. We compare representatives of the two approaches using a 465,000 page crawl and test queries applicable to the university context. 503 pairs of result sets were judged by 56 Chinese students.

Although there are differences on certain queries, we find no overall advantage to either method. To understand the merits of each approach, we analyze cases where they performed differently. Our analysis enumerates situations which favour segmentation, and those which favour bigrams. We observe that further improvements in segmentation accuracy will not improve retrieval effectiveness.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*search process*

Keywords

Chinese IR, segmentation, site search

1. INDEXING FOR CHINESE IR

Words are the basic units of meaning in natural language and are normally used as the fundamental indexing unit during the Information Retrieval (IR) process. In English and other European languages words are separated clearly by whitespace. However, in Chinese, text appears as a continuous character string[1]. Due to the absence of obvious word boundaries these continuous strings have to be segmented into smaller units for indexing, if familiar text retrieval technologies such as inverted files are to be used.

For example, given the English text "the inventor of the computer is John von Neumann", it is trivial to identify nine words and eight index entries. The Chinese equivalent, "计算机的发明者是约翰冯诺依曼", has no explicit boundaries but could be segmented as "计算机 / 的 / 发明者 / 是 / 约翰 / 冯 / 诺依曼" to give seven words. Note that words in Chinese, as in English, consist of a variable number of characters.

Absent any explicit word boundaries, there are three main approaches for Chinese indexing: word-based, character-based, and hybrid approaches. *Word-based* approaches attempt to segment Chinese text into semantically meaningful chunks using dictionaries, statistics or linguistic knowledge. For example: "计算机科学" (computer science) can be segmented as "计算机" (computer) and "科学" (science). On the other hand, *character-based* methods simply index a fixed length sequence of characters (n-grams) as a unit. N-grams may be disjoint or overlapping: overlapping bigrams are the most commonly used in practice. The above example has four overlapping bigrams: "计算", "算机", "机科", and "科学". Finally, *hybrids* combine character-based and word-based approaches [14].

The choice of how and whether to segment text determines which IR and language processing technologies can be brought to bear, and we may expect it to influence retrieval quality. If words are incorrectly segmented, therefore incorrectly indexed, they may not be retrieved; or downstream analysis may be more difficult. However, it is not clear whether automatic segmentation is accurate enough for retrieval purposes or when errors are important in practice.

Past work has been ambiguous on the value of segmentation, but has been limited by small corpora and artificial judgements; little past work has considered web search and none has considered search at a single website. In this work we reexamine segmentation in the context of site search; with queries from potential users; and with it situ judgements from the same potential users. We find no performance differences overall, but our case studies demonstrate where and when segmentation or bigrams are advantageous.

2. PAST WORK

A good deal of past work has considered whether segmentation or bigram indexes are more effective for Chinese IR and whether the quality of segmentation is reflected in the final quality of results. Table 1 provides a summary[2].

Past work has been equivocal, at best, on the relative effectiveness of segmentation and bigrams. On data from the TREC Chinese corpus, Tong et al. [24] reported similar effectiveness with bigrams and segmentation, conclusions which were repeated by Kwok [9] and Nie et al. [17]. On the same data, however results from Palmer and Burger [20], as well as later work by Nie et al. [18], preferred segmentation; results from Leong and Zhou [11] preferred bigrams. On other data there are claims both for segmentation [5, 8] and for bigrams [7].

[1]Other Asian languages such as Japanese, Korean, Khmer and Thai also present text without word boundaries.

[2]A full version of this table, with commentary, is available in Zhou's thesis [26].

ADCS December 08–09, 2015, Parramatta, NSW, Australia

http://dx.doi.org/10.1145/2838931.2838940.

Work	Data collection	Evaluation method(s)	Conclusion(s)
Tong et al. [24]	TREC	Recall, AP, R-prec., P-R curves	Equal
Nie et al. [16]	1270KB	P-R curves	Direct relationship
Kwok [10]	TREC	AP, R-prec., P@N	No direct relationship
Kwok [9]	TREC	Rel. retrieved@1000, AP, R-prec., P@N	Equal, no direct relationship
Palmer and Burger [20]	TREC	AP, R-prec.	Segmentation better, no direct relationship
Nie et al. [17]	TREC	AP, R-prec	Equal
Nie et al. [18]	TREC	AP	Equal
He et al. [5]	TREC, NTCIR-2	AP	Direct relationship
Peng et al. [21]	TREC	AP, R-prec.	No direct relationship
Cao et al. [2]	CIRB010 [15]	Precision, recall, AP	No direct relationship
Kang et al. [7]	NTCIR-4	AP	Bigrams better, direct relationship
Foo and Li [3]	266 files	AP, R-prec., recall	Not direct relationship
Kim and Ming [8]	1200 articles	P-R curves	Segmentation better, direct relationship
Jin et al. [6]	85352 news pages	AP	No direct relationship

Table 1: Previous work has been inconclusive on the performance of bigram and segmented indexes, and on the relationship between segementation and retrieval quality. A fuller version of this table, with commentary, is available [26].

The relationship between segmentation accuracy and IR performance has also been unclear. In much previous work, segmentation accuracy is explicitly or implicitly assumed to be an important factor in determining IR effectiveness [4, 12, 13]; however, experimentation provides little evidence for this. Some studies find that segmentation accuracy directly determines IR result [5, 7, 8, 19], while others find no relationship [2, 3, 6, 9, 20, 21].

We also note some limitations. All of the cited studies make use of small corpora (less than 100,000 documents) and only one corpus includes web pages—although these were news articles in a formal style. None of the studies make use of web-style queries and all focus on queries with informational intent [1]. They use judgments and measures, such as mean average precision (MAP), which may not be appropriate for navigational or transactional intents.

3. EXPERIMENT

As part of a study of segmentation for Chinese site search [26], we built a corpus of 466,897 pages crawled from Northeast Normal Univeristy (NENU, `nenu.edu.cn`). Chinese content in the crawl was indexed two ways: once ("Seg") using standard word-based indexing, using a commercial enterprise search engine and the IK segmenter[3]; and once ("Grams") as overlapping bigrams. Queries to each index were similarly segmented, or interpreted as bigrams, but the engines were otherwise identical.

We recruited 56 partipants from the ANU and from NENU; they were familiar with university-related information needs and we expect that their queries and judgments are representative of those of a significant population of the site's users. Participants were asked to submit queries which they expected to be satisfied by NENU web pages. A randomized list of query suggestions was provided to each participant, although they were free to enter any query they liked.

Judges used a side-by-side evaluation interface [23], with a single query entry box and two side-by-side panels of results. Grams results appeared in one panel, chosen at random, and those from Seg in the other. Judges chose between buttons labelled "Prefer left", "equally good", "equally bad", and "prefer right".

We received several queries which were obscene, unrelated to NENU, or consisted of a single English character; we removed these, as well as 13 submissions of the same query by a single participant. In the remaining 503 votes, from 56 participants, we saw no clear preference for either index method. In 42% of cases, participants saw no difference at all (133 votes were for "equally good", 77 for "equally bad"). Across the other cases the preference for one engine over the other was not significant: 137 votes were cast for Grams and 156 for Seg (sign test, $p = 0.15$). We also saw no preference when we divided the queries according to Broder's taxonomy [1], although there was an apparent advantage to Grams for short queries (2–3 characters) and to Seg for long queries (6 characters or more).

[3] `https://code.google.com/p/ik-analyzer/`

4. CASE STUDIES

As with past work, our high-level results are inconclusive and possibly dependent on the details of our particular audience, segmenter, and IR system.

The data here comes from only one segmenter and search engine, and we cannot claim segmentation in general is more or less effective than bigrams. Particular cases, however, are instructive and give some insight as to why—and when—one method or the other may be preferred, regardless of the particular technology in use. Here we present some illustrative cases from our partcipants' queries and judgements.

4.1 Segmentation preferred, when correct

A bigram index can introduce inappropriate matches on unrelated words, or parts of longer words. In these cases, segmentation leads to higher precision:

"文学院" (College of Liberal Arts) This query was segmented as a single word, and Seg results were restricted to web pages specifically mentioning the College of Liberal Arts. In contrast Grams indexed "文学" (liberal arts) and "学院" (college), and consequently also matched another NENU college with a similar name: "人文学院" (College of Humanities).

"化学学院" (Faculty of Chemistry) This query was segmented by the system as "化学" (chemistry) and "学院" (faculty), and the Seg system was able to retrieve web pages which mention these two words. Grams, on the other hand, indexed this query as three items: "化学" (chemistry), "学学", and "学院" (faculty) which shares two common index items with "历史文化学院" (School of History and Culture). Again this led to the bigrams-based search engine returning lots of web pages from the School of History and Culture, many of which also matched on the false, or nonsense, word "学学".

4.2 Segmentation incorrect but preferred

Let us define *incorrect segmentation* to mean segmentation which introduces obviously unrelated words, or fails to recognize new words, person names or proper nouns. We manually identified all such cases of incorrect segmentation among the submitted queries.

There were relatively few cases of incorrect segmentation (12 queries, 19 of 503 query instances), but amongst these there were cases where Seg was still preferred. Examples include:

"东师幼儿园" (NENU kindergarten) The segmentation method failed to recognize "东师", the abbreviation of "东北师范大学" (Northeast Normal University), as a single word, and instead indexed two words: "东" (east) and "师" (teachers). It indexed the word "幼儿园" (kindergarten) correctly. However, there is only one kindergarten in the range of NENU, so there is little chance of the IR system missing the relevant results.

We saw a similar fault and effect with "东师会馆" (NENU hotel), which was indexed as "东" (east), "师" (teachers), "会馆" (hotel).

"净月校区办公室" (Jinyue District Office) A proper noun "净月" was segmented as two words: "净" (clean) and "月" (moon), while words "校区" (district) and "办公室" (office) were indexed correctly. Because there are only two districts in NENU, as long as the words "district" and "office" were present, the IR system was able to find relevant pages.

There is little evidence that correcting the segmentation errors (in both queries and documents) would increase the preference for Seg.

4.3 Bigrams preferred

There were a variety of queries where Grams was preferred over Seg, including:

"东北师大" (NENU) This is a short form of "东北师范大 学" (Northeast Normal University) and is indexed as one word. But unfortunately, not one web page in NENU contains this word, so Seg returns zero results. For Grams this query was indexed as three items: "东北" (northeast), "北师" and "师大" (normal university). Two of these bigrams are found within "东北师范大 学" (Northeast Normal University).

"博士条件" (PhD requirements) This was segmented as "博士" (PhD), "条件" (requirements); a related query was segmented as "博 士" (PhD), "申请" (application). In both cases, many relevant web pages express the same meaning with slightly different representations: "博士生条件" (PhD student requirements) and "博 士生申请" (PhD student application) which were segmented as "博士 生" (PhD student), "条件" (requirements) and "博士生" (PhD student), "申请" (application). Thus, for Seg, "博士" (PhD) and "博士生" (PhD student) failed to match, while "条 件" (requirements) and "申请" (application) are common in a university web site and resulted in many irrelevant hits.

"学校网络" (campus networks) This query was segmented as one word by Seg. However, relevant documents contain the phrase "学校网络中 心" (campus networks center) which was segmented as "学校" (campus) and "网络中心" (networks center), and could not be matched by Seg. Grams indexed this phrase as "学校" (campus), "校网", "网络"networks, "络中" and "中心" (center), which matched all the query bigrams: "学校" (campus), "校网" and "网 络" (networks), leading to successful retrieval.

"科学家" (scientist) This query was indexed as one word in Seg. However, many derivative words appear in web pages: "生物学 家" (biologist), "物理学家" (physicist), "数学 家" (mathematician), "计算机学家" (computer scientist), etc. All of them were indexed as a whole word by Seg, so none would match the query, but they share a common suffix: "学家" (-ist). For bigrams, "科学家" (scientist) was indexed as "科学" (science) and "学家" (-ist). So "数学家" (mathematician) for example, which is indexed as "数学" (mathematics) and "学家" (-ist), matches on one index item.

"学校地址" (campus address) The segmentation algorithm treated this query as one word. However, relevant documents contain another word "学校地理" (campus address) instead, and therefore, Seg retrieved zero results. Grams performance was dramatically better than Seg because two of three bigrams overlap.

"孔俊" (Jun Kong) This is a person name which was segmented as two separate commonly used Chinese characters "孔" (a common Chinese family name) and "俊" (beautiful/handsome/smart) by Seg, leading to many irrelevant results. In contrast, Grams only indexed "孔俊", and was thus able to return more precise results.

We saw a similar effect with "导员" (tutor), which was segmented as "导" (lead, guide, conduct, teach) which might appear in words such as "导师" (supervisor), "导论" (introduction), or "领 导" (leader) in the university context; and "员" (-or, -er), a suffix in Chinese to represent people.

Only in the last two of these cases can the preference for Grams be attributed to incorrect segmentation. In general, we can see that bigrams indexing is better able to handle person names, abbreviations, synonyms and sub-word matching.

5. ADVANTAGES OF BIGRAMS

A generally accepted explanation of the relatively good performance of bigrams indexes in IR is that most Chinese words are two characters [7, 20] and that consequently many correct words are indexed. But after examining the cases in our experiment, additional characteristics of bigrams indexing can be seen to be beneficial.

Matching sub-words. Bigrams are not only capable of representing long words, they are also able to index meaningful sub-words of a relatively long word. Segmentation can only achieve this at the expense of an exhaustive strategy [19].

Synonyms. Synonyms in Chinese always share one or more characters. For example, "发明" (develop) and "发现" (discover) share the same character "发" (emit). If "发明" (develop) appears in the query, and there are no web pages containing "发 明" (develop) but some have "发", then a bigrams-based system can at least match one character and possibly retrieve web pages which mention "发现" (discover). A segmentation-based system, on the other hand, finds no instances. We see this in the "campus address" and "scientist" cases above.

Abbreviations. Abbreviations in Chinese always retain at least one key character from the original word. For example, as discussed above, the abbreviation "东北师大" (NENU) includes four characters from the full version of "东北师范大学" (Northeast Normal University). Seg indexes the latter as a single word, meaning that it cannot even partially match the abbreviation.

Unknown personal names. It is very common to see names which include frequent words. For example, "李发展" includes "发 展" (develop). Grams indexes it as "李发" and "发展" (develop), whereas Seg indexes it as "李" (Lee) and "发展" (develop) assuming that the full name is not in the dictionary. Because "李" (Lee) is a very common family name in China and "发 展" (develop) is a frequent

word, segmenting search engines are highly likely to return results which mention another person with the same family name Lee and which talk about development. We saw this in our data with "孔俊" (Jun Kong).

Indexing a query as overlapping bigrams will potentially introduce unrelated words or parts of other longer unrelated words. For example, "严守一把手机关了" (Shouyi Yan shutdown his phone) would be indexed as the following bigrams: "严守" (adherence), "守一", "一把" (once), "把手" (handle), "手机" (phone), "机关" (office, mechanism, stratagem) and finally "关了" (shutdown). In this example, four unrelated words are introduced, and the probability of retrieving irrelevant documents will increase accordingly.

Previous studies suggest that false words in bigram indexes have a negative affect on IR efficiency and effectiveness [25]. However, we saw only an 11% increase in the size of the inverted file, no apparent difference in speed, and no difference in overall preferences. On the other hand, we find that some false words are potentially useful for the IR system. In particular, they allow *matches on multi-character words:* although AB and BC are false words, documents containing the three-character word ABC can still be matched. They also allow *matching words in sequence:* for example, allowing us to distinguish "狗咬了人" (dog bites man) from "人咬了狗" (man bites dog).

6. DISCUSSION AND CONCLUSIONS

Accurate segmentation can indeed help to improve IR performance when overlapping ambiguity occurs in queries such as in the Shouyi Yan example. As seen above, however, accurate segmentation cannot always guarantee good IR; but nor do segmentation errors always hurt.

We observed no clearcut advantage of segmentation over bigrams in document matching and ranking. We note that navigational searches in our website context worked well when link and anchortext were exploited, regardless of whether bigrams or words were indexed. Our results suggest that the bigrams method may work better for queries of three characters or less while segmentation was clearly superior for queries longer than six characters. We observed few cases of failed segmentation cases and noted that they did not measurably harm performance. Failure to recognize novel words had only very limited influence on the IR results when other context information was preserved. Finally, we identified a number of patterns in which false words may confer advantage to bigrams.

Sophisticated IR systems do however include features which rely on semantics captured better in words than in bigrams: query expansion through synonyms, negated antonyms and other transformations based on thesaurus lookups; named entity extraction; question answering; spelling correction; semantic inference based on natural language processing (NLP) techniques. It might be possible to base relevance feedback on bigrams but this would have to be verified; it seems more natural to use words [16, 22].

Further improvements may be possible with exhaustive segmentation [19], recognising synonyms and abbreviations, and hybrids of Grams and Seg indexes. Investigating these approaches in Chinese website search is left for future work.

References

[1] A. Broder. A taxonomy of web search. *SIGIR Forum*, 36(2):3–10, 2002.

[2] G. Cao, P. He, G. Wu, and S. Nie. 中文分词对中文信息检索系统性能的影响 (Impact of Chinese Segmentation to Chinese Information Retrieval). *Computer Engineering and Applications*, 19:78–79, 2003.

[3] S. Foo and H. Li. Chinese word segmentation and its effect on information retrieval. *Inf. Proc. & Management*, 40(1):161–190, 2004.

[4] Q. Fu. 基于搜索统计技术中文分词算法的应用研究 (Application of statistical techniques to Chinese word segmentation algorithm). China Sciencepaper Online, 2007. http://www.paper.edu.cn/releasepaper/content/200704-749.

[5] H. He, P. He, J. Gao, and C. Huang. Finding the better indexing units for Chinese information retrieval. In *Proc. SIGHAN Workshop on Chinese Language Processing*, pages 1–7, 2002.

[6] P. Jin, Y. Liu, and S. Wang. 汉语分词对中文搜索引擎检索性能的影响 (Influence of Chinese word segmentation on web information retrieval). *Journal of the China Society for Scientific and Technical Information*, 25(1):21–24, 2006.

[7] I.-S. Kang, S.-H. Na, and J.-H. Lee. Combination approaches in information retrieval: words vs. n-grams, and query translation vs. document translation. In *Proc. NTCIR*, 2004.

[8] D. Kim and S. Ming. Effectiveness of segmentation granularity and indexing units for worst case evaluation in Chinese information retrieval. *Proc. Int. Conf. Internet Information Retrieval*, pages 177–180, 2005.

[9] K. L. Kwok. Comparing representations in Chinese information retrieval. *SIGIR Forum*, 31:34–41, 1997.

[10] K. L. Kwok. Lexicon effects on Chinese information retrieval. In *Proc. Empirical Methods in NLP*, pages 141–8, 1997.

[11] M.-K. Leong and H. Zhou. Preliminary qualitative analysis of segmented vs bigram indexing in Chinese. In *Proc. TREC-6*, pages 551–557, 1997.

[12] X. Liu, Y. Hu, and X. Ai. 开源中文分词器在 web 搜索引擎中的应用 (The application of open source Chinese tokenizer in web search engine). *Computer Engineering & Software*, 34(3):80–83, 2013.

[13] S. Long, Z. Zhao, and H. Tang. Overview on Chinese Segmentation Algorithm. *Computer Knowledge and Technology*, 5(10):2605–2607, 2009.

[14] R. W. Luk, K.-F. Wong, and K.-L. Kwok. Hybrid term indexing: an evaluation. In *Proc. NTCIR*, pages 130–136, 2001.

[15] National Taiwan University. Chinese Information Retrieval Benchmark version 1.0 (CIRB010). Web site, Jun 2000. http://lips.lis.ntu.edu.tw/cirb/releases/CIRB010.htm.

[16] J. Y. Nie, M. Brisebois, and X. Ren. On Chinese text retrieval. In *Proc. SIGIR*, pages 225–233, 1996.

[17] J. Y. Nie, J. P. Chevallet, and M. F. Bruandet. Between terms and words for European language IR and between words and bigrams for Chinese IR. In *Proc. TREC-6*, pages 697–710, 1998.

[18] J. Y. Nie, J. Gao, J. Zhang, and M. Zhou. On the use of words and n-grams for Chinese information retrieval. In *Proc. Int. Work. Information Retrieval with Asian Languages*, pages 141–148, 2000.

[19] D. W. Oard and J. Wang. Effects of term segmentation on Chinese/English cross-language information retrieval. In *Proc. SPIRE*, pages 149–157, 1999.

[20] D. Palmer and J. Burger. Chinese word segmentation and information retrieval. In *Proc. AAAI Spring Symposium*, pages 175–178, 1997.

[21] F. Peng, X. Huang, D. Schuurmans, and N. Cercone. Investigating the relationship between word segmentation performance and retrieval performance in Chinese IR. In *Proc. COLING*, pages 1–7, 2002.

[22] M. Sun, J. Zou, et al. 汉语自动分词研究评述(Chinese Automatic Segmentation Research Review). *Contemporary Linguistics*, 3(1): 22–32, 2001.

[23] P. Thomas and D. Hawking. Evaluation by comparing result sets in context. In *Proc. of CIKM 2006*, pages 94–101, 2006.

[24] X. Tong, C. Zhai, N. Millic Frayling, and D. A. Evans. Experiments on Chinese text indexing: CLARIT TREC-5 Chinese track report. In *Proc. TREC-5*, pages 335–339, 1997.

[25] S. Wang. 面向大规模信息检索的中文分词技术研究 *(Chinese Words Segmentation Technology in Large-scale Information Retrieval)*. PhD thesis, Beijing: Institute of Computing Technology Chinese Academy Of Sciences, 2006.

[26] L. Zhou. Investigating indexing units for Chinese web information retrieval: Chinese word segmentation versus N-grams. Master's thesis, Australian National University, 2013.