Chinese Word Segmentation and Part-of-Speech Tagging in One Step*

Tom B.Y. Lai, Maosong Sun, Benjamin K. Tsou, S. Caesar Lun
cttomlai@cityu.edu.hk, rlvt6@cityu.edu.hk, rlbtsou@cpccux0.cityu.edu.hk, ctslun@cityu.edu.hk
City University of Hong Kong

Abstract

In Chinese natural language processing, word segmentation and part-of-speech tagging is generally carried out as two separate steps. Earlier, the authors introduced a tag-based Markov-model approach to word segmentation. As the tags are of a syntactic nature, this is effectively doing word segmentation and part-of-speech tagging simultaneously. We have used a best-first algorithm with empirical results showing the search for the best solution to be efficient for inputs of reasonable length. In this paper, we will see that the job can be done using an $O(n^2)$ algorithm. In our experiments, we actually had the algorithm reduced to O(n) by setting a maximum number of character for words in Chinese to a constant. We also show that performing word segmentation and part-of-speech tagging in one step will bring about improvement in accurracy.

1. Introduction

Chinese word segmentation (Chen 1992) can be done using a number of approaches (Liang 1987, He, 1991, Fan 1988, Sproat 1990 & 1996, Yeh 1991, Chang 1991, Lua 1994, Wu 1995) including the maximal-match principle, rule-based approaches and probability-based approaches. Lai 1991 suggests doing Chinese word segmentation by optimizing the product of successive tag bigram probabilities. The tags used (Sun 1992) are of a syntactic nature. Experiment results (Lai 1992) show that using the A* search algorithm is efficient for inputs of reasonable length (up to 30 characters). Bai 1995 uses tag-based bigrams to resolve ambiguities after segmenting the input. Sun 1995 uses mutual information instead of bigram probability.

---

* Supported by Research Grant #9040049, University Grant Council, Hong Kong

Markov-model approaches (Bahl 1983) have been used successfully in part-of-speech tagging (Marshall 1983, . DeRose 1988, Kupiec 1992, Chang 1993a).

In Chang 1993b, the best N outputs of a segementation module are passed to a tagging module. The two modules, operating sequentially, contribute to a score function that is used to yield the best segmentation and tagging scheme.

Lai 1991 and 1992 apply Markov Model techniques in Chinese word segmentation by using tags. As the tags are of a syntactic nature, this is effectively doing word segmentation and part-of-speech tagging at the same time. This is a genuine one-step approach. There is a well-understood linear-time dynamic programming algorithm for Markov-model-based approaches (Bahl 1983 and, e.g., DeRose 1988). However, the fact that a Chinese word can consist of a variable number of characters makes it impossible for this algorithm to be used in our approach. The A* search used in Lai 1992 is inefficient theoretically. But for inputs of less than 30 characters, space- and time-complexity are linear empirically. In this paper, we will see that genuine simultaneous word segmentation and part-of-speech tagging can nevertheless be done using an efficient dynamic programming algorithm.

2. Segmentation and part-of-speech tagging in one-step

2.1 When word segmentation and part-of-speech tagging are carried out one after another, errors in the two steps multiply. But if the two processes are integrated, then their interaction may help improve the combined accuracy. Consider:

dong1 ji4 shi4 cong2 tou2 nian2 yuan2 yue4 kai1 shi3 di2/de				(1a)
shi4	cong2_tou2	nian2	yuan2_yue4	(1b)
V	Adv	TimeN	TimeN	
copula	start afresh	year	first month of year	
shi4	cong2	tou2_nian2	yuan2_yue4	(1c)
V	P	TimeN	TimeN	
copula	from	last year	first month of year	

Input (1a) may be segmented either into (1b) or (1c). If segmentation is carried out independently before part-of-speech tagging, (1b) will probably be preferred as tou2_nian2 in (1c) is rather infrequent in Chinese text. The final tag sequence of V-Adv-TimeN-TimeN, though rather unlikely by itself, will be produced. On the other hand, if part-of-speech tagging is carried out at the same time as word segmentation, then the fact that the tag sequence V-P-

TimeN-TimeN is more likely may be enough to offset the balance to allow the correct segmentation scheme (1c) to come out as the winner.

2.2 Word segmentation and part-of-speech tagging can be carried out simultaneously using a tag-based Markov model (Lai 1992). Consider:

yu3 zhong1 guo2 you3 guan1 lian2 (2)
yu3 / zhong1 guo2 / you3 guan1 / lian2 (2a)
yu3 / zhong1 guo2 / you3 / guan1 lian2 (2b)
yu3 / zhong1 guo2 / you3 / guan1 / lian2 (2c)
yu3 / zhong1 / guo2 you3 / guan1 lian2 (2d)
yu3 / zhong1 / guo2 you3 / guan1 / lian2 (2e)
yu3 / zhong1 / guo2 / you3 guan1 / lian2 (2f)
yu3 / zhong1 / guo2 / you3 / guan1 / lian2 (2g)

Input sentence (2) can be segmented into 2(a) to (2g). Words in (2a) to (2g) above may have the following tags: tag(yu3) = {jom, pom}, tag(zhong1 guo2) = {spd}, tag(zhong1) = {fom, spm, vnm}, tag(you3 guan1) = {qd, vnd}, tag(you3) = {vy}, tag(lian2) = {vnm, cnr, bom, pom}, tag(guo2 you3) = {aod}, tag(guo2) = {nam}, tag(guan1 lian2) = {vnd}, tag(guan1) = {vnm, ncm}.

As a word can have more than one tag, each segmentation scheme in (2a) to (2g) will correspond to a number tag sequences. The correct segmentation (2b), corresponding to tag sequence

pom /spd / vy / vnd (2b*),

is found by maximizing the product of successive tag bigrams. (Lai 1992 for details.)

A closer look at the tags reveals that they are essentially of a syntactic nature. For example, *pom* and *spd* in (2b*) are monosyllabic preposition and poly-syllabic proper noun respectively. With the correctly segmentated character string (2b), we also obtain the part-of-speech tagging information in (2b*). We are thus effectively performing segmentation and part-of-speech tagging at the same time.

One problem with this approach is that syntactic class information has to be encoded in the lexicon. This is expensive in terms of resources. However, it should be noted that such information is required for part-of-speech tagging anyway.

Another problem is computational efficiency. The well-understood linear-time algorithm for Markov-model based part-of-speech tagging (e.g. DeRose 1988) cannot be used. The best-

first search algorithm in Lai 1992 is exponential, though experimental results show that it is efficient in practical situations (with the input containing up to 30 characters). Addressing this issue, we have designed an $O(n^2)$ dynamic programming algorithm (described elsewhere) for finding the best segmentation-tagging alternative. By setting the maximum number of characters in a word to a constant, this algorithm can be further reduced to O(n).

3. Experimental results

Using a 486 PC (66 Mhertz, 8 M), we have performed experiments (using bigrams) on 729 sentences with a total of 10734 character tokens. The $O(n^2)$ algorithm was reduced to O(n) by setting the maximum number of characters per word to 7. For the sake of clarity, define:

FP: all segmentation possibilities taken into account, each of which further expanded to a corresponding tag lattice;

MM: maximal-match used, so only one segmentation candidate and one tag lattice;

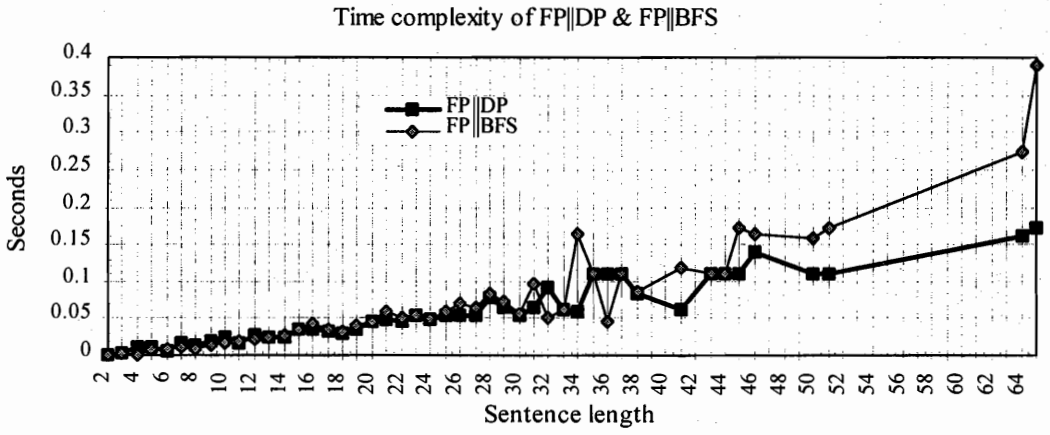
DP: dynamic programming used to find the most likely path through the tag lattice;

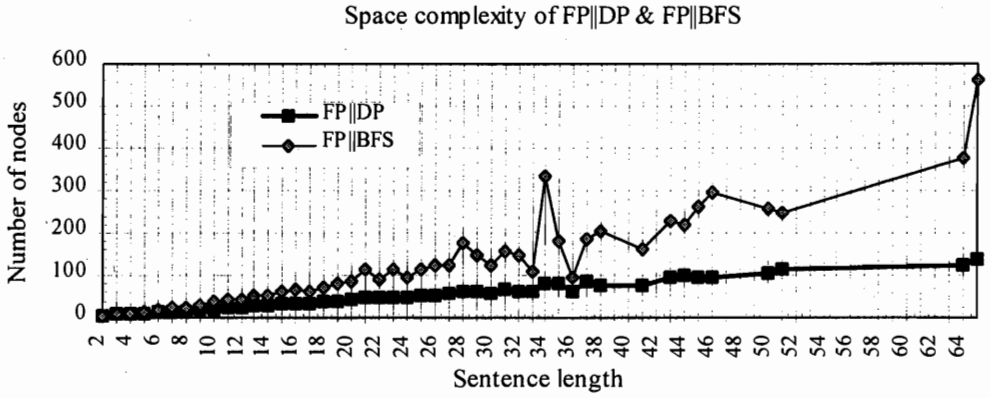
BFS: best-first search used to find the most likely path through the tag lattice;

x||y: procedures x and y carried out simultaneously;

x+y: procedures x and y carried out successively

3.1 Comparing FP||DP and FP||BFS





Sentences: 729 Character tokens: 10734

	Total time (seconds.)	Average time per character (seconds)	Total no. of nodes	Average no. of nodes per character	Total no. of arcs	Average no. of arcs per character
FP||DP	22.74	0.00	22520	2.10	21791	2.03
FP||BFS	23.24	0.00	44766	4.17	44037	4.10

(1) the time and space graphs for FP||DP are approximately linear;

(2) the number of nodes/arcs created by FP||DP is about half of that created by FP||BFS;

(3) time-efficiency improvement is signifcant for input more than 30 characters long.

3.2 Comparing FP||DP and MM+DP.

While the efficiency (and linearity) of our algorithm is established above, it is to be expected that finding the best segmentation-tagging acheme in a one step involves a larger search space than finding the best segmentation alternative and the best tagging scheme thereof in two separate steps. MM+DP, for example, should be less expensive than FP||DP. However, our results show that FP||DP does not compare too unfavourably with MM+DP.

Sentences: 729 Character tokens: 10734

	Total time (seconds.)	Average time per character (seconds)	Total no. of nodes	Average no. of nodes per character	Total no. of arcs	Average no. of arcs per character
FP||DP	22.74	0.00	22520	2.10	21791	2.03
MM+DP	9.70	0.00	10617	0.99	9888	0.92

FP||DP is just a little more than twice more expensive than MM+DP in terms of both space

and time. If MM were replaced with a procedure that returned more than one segmentation scheme for the DP tagging component to work on, the combined procedure would also be more expensive than MM+DP. FP||DP is thus efficient compared to sequentially combined segmentation and tagging.

3.3 Effectiveness of FP||DP: accuracy (precision) improvements

	Correctly segmented sentence+DP	MM+DP	FP||DP
Accuracy of word segmentation (A)	100.00%	98.35%	99.66%
Accuracy of POS tagging (B)	95.06%	93.13%	94.66%
Estimation of the total performance(A*B)	95.06%	91.65%	94.34%

The left-most column gives the upper bounds (no segmentation errors). Compared with MM+DP, FP||DP has a 1.31% improvement in segmentation, a 1.47% improvement in POS tagging, and a 2.69% improvement in the combined process. This shows that doing segmentation and part-of-speech tagging simultaneously is indeed more effective than performing the two tasks in two separate steps.

4. Conclusion

We have shown that segmenting a sentence and marking parts of speech of the words identified simultaneously will improve both segmentation accuracy and part-of-speech tagging accuracy. Using our tag-based Markov-model approach, this can be done effectively, with an $O(n^2)$ algorithm.

References

Bahl, L.R., F. Jelinek and R.L. Mercer, "A Maximum Likelihood Approach to Continuous Speech Recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol *PAMI*-5, No. 2, March 1993, pp. 179-190.

Bai, S.H., "An Integrated Model of Chinese Word Segmentation and Part-of-Speech Tagging (in Chinese)," in *Advances and Applications on Computational Linguistics (Selected Papers*

from the 3rd National Conference on Computational Linguistics, Shanghai, Nov. 5-7. 1995), Tsinghua University Press, 1995, pp. 56-61.

Chen, K.J. and S.H. Liu, "Word Identification for Mandarin Chinese Sentences," *COLING9-92*, Nantes, 23-28 Aug., 1992, pp. 101-107.

Chang, C.H. and C.D. Chen, "HMM-based Part-of-Speech Tagging for Chinese Corpora," *Proc. Workshop on Very Large Corpora: Academic and Industrial Perspectives*, Columbus, Ohio, June 1993, pp. 40-47. (1993a)

Chang, C.H. and C.D. Chen, "A Study on Integrating Chinese Word Segmentation and Part-of-Speech Tagging," *Communictions of COLIPS*, Vol. 3, No. 1, 1993, pp. 69-77. (1993b)

Chang, J.S., J,I. Chang and S.D. Chen, "A Method of Contstraint Satisfaction and Statistical Optimization for Chinese Word Segmentation," *Proc. 1991 ROCLING*, Kenting, Taiwan, August 1991.

DeRose, S.J., "Grammatical Category Disambiguation by Statistical Optimization," *Computational Linguistics*, Vol. 14, No. 1, Winter 1988, pp. 31-39.

Fan, C.K. and W.H. Tsai, "Automatic Word Identification in Chinese Sentences by the Relaxation Technigue," *Computer Processing of Chinese and Oriental Languages*, Vol.4, No.1, November 1988, pp. 33-56.

He, K.K., Xu, H. and B. Sun, "Design Principles of a Expert System for Word Segmentation for Written Chinese Text (in Chinese)," *Journal of Chinese Information Processing*, 5-2, 1991.

Lai, T.B.Y., S.C. Lun, C. F. Sun and M.S. Sun, "A Maximal Match Chinese Text Segmentation Algorithm Using Mainly Tags for Resolution of Ambiguities (in Chinese)." *Proc. of ROC Computational Linguistics Conference*, Kenting, Taiwan, August 1991, pp. 135-146.

Lai, T.B.Y., S.C. Lun, C.F. Sun and M.S. Sun, "A Tagging-based First-Order Markov Model Approach to Automatic Word Identification for Chinese Sentences," *Proc. 1992 International Conference on Computer Processing of Chinese and Oriental Languages*, Tampas, Fl, 15-18 Dec., 1992, pp. 17-23.

Liang, N.Y, "Automatic Segmentation of Chinese Words and the Related Theory." Proc. 1987 *International Conference on Chinese Information Processing*, Beijing, 1987, pp. 454-9.

Lua, K.T. and G.W. Gan, "An Application of Information Theory in Chinese Word

Segmentation," *Computer Processing of Chinese & Oriental Langauges*, Vol. 8, No.1, June 1994, pp. 115-124.

Marshall, I., "Choice of Grammatical Word-Class Without Global Syntactic Analysis: Tagging Words in the LOB Corpus." *Computers in the Humanities*, Vol. 17, 1983, pp. 139-150.

Sproat, R. and C. Shih, "A Statistical Method for Finding Word Boundaries in Chinese Text," *Computer Processing of Chinese and Oriental Languages*, Vol. 4, No. 4, March 1990, pp. 336-351.

Sproat, R., C. Shih, W. Gale and N. Chang, "A Stochastic Finite-State Word-Segmentation Algorithm for Chinese," *Computational Linguistics*, 22(3), 1996, pp. 377-404.

Sun, M.S., T.B.Y. Lai, S.C. Lun and C.F. Sun. "A Tagset for Automatic Chinese Text Segmentation," First International Conference on Chinese Linguistics, Singapore, June 1992. Printed in *Working Papers in Languages and Linguistics*, No. 5, City University of Hong Kong, April 1993, pp. 127-134.

Sun, M.S., T.B.Y. Lai, T.B.Y., S.C. Lun and C.F. Sun, "Some Issues in the Statistical Approach to Chinese Word Indentification," *Proc. 1992 International Conference on Chinese Information Processing*, Oct. 26-28, 1992, Beijing, pp. 246-253.

Sun, M.S. and B.K. Tsou, "Ambiguity Resolution in Chinese Word Segmentation," *Proceedings of the 10th Pacific Asia Conference*, Dec. 27-28, 1995, Hong Kong, pp. 121-126.

Wu, D.K, "Stochastic Inversion Transduction Grammars, with Application to Segmentation, Bracketing, and Alignment of Parallel Corpora," *Proceedings of IJCA-95 (Fourteenth International Joint Conference on Artificial Intelligence)*, Montreal, 1995, pp. 1328-1334.

Yeh, C.L. and H.J. Lee, "Rule-based Word Identification for Mandarin Chinese Sentences - A Unification Approach," *Computer Processing of Chinese and Oriental Langauges*, Vol 5, No. 2, March 1991, pp. 97-118.