A hybrid approach to vietnamese word segmentation using part of speech tags

Pham D.D., Tran G.B., Pham S.B.

Human Machine Interaction Laboratory, College of Technology, Vietnam National University, Hanoi, Viet Nam

Abstract: Word segmentation is one of the most important tasks in NLP. This task, within Vietnamese language and its own features, faces some challenges, especially in words boundary determination. To tackle the task of Vietnamese word segmentation, in this paper, we propose the WS4VN system that uses a new approach based on Maximum matching algorithm combining with stochastic models using part-of-speech information. The approach can resolve word ambiguity and choose the best segmentation for each input sentence. Our system gives a promising result with an F-measure of 97%, higher than the results of existing publicly available Vietnamese word segmentation systems. ?? 2009 IEEE.

Index Keywords: Boundary determination; F-measure; Hybrid approach; Maximum matchings; New approaches; Part Of Speech; Part-of-speech tags; Word segmentation; Word segmentation systems; Stochastic models; Systems engineering; Knowledge engineering

Year: 2009
Source title: KSE 2009 - The 1st International Conference on Knowledge and Systems Engineering
Art. No.: 5361713
Page : 154-161
Link: Scorpus Link
Correspondence Address: Pham, D. D.; Human Machine Interaction Laboratory, College of Technology, Vietnam National University, Hanoi, Viet Nam; email: dangpd@vnu.edu.vn
Sponsors: College of Technology;Vietnam National University
Conference name: 1st International Conference on Knowledge and Systems Engineering, KSE 2009
Conference date: 13 October 2009 through 17 October 2009
Conference location: Hanoi
Conference code: 79895
ISBN: 9.78E+12
DOI: 10.1109/KSE.2009.44
Language of Original Document: English
Abbreviated Source Title: KSE 2009 - The 1st International Conference on Knowledge and Systems Engineering
Document Type: Conference Paper
Source: Scopus
Authors with affiliations:

1. Pham, D.D., Human Machine Interaction Laboratory, College of Technology, Vietnam National University, Hanoi, Viet Nam
2. Tran, G.B., Human Machine Interaction Laboratory, College of Technology, Vietnam National University, Hanoi, Viet Nam

3. Pham, S.B., Human Machine Interaction Laboratory, College of Technology, Vietnam National University, Hanoi, Viet Nam

References:

1. Jacobs, A.J., Wong, W.Y., Maximum entropy word segmentation of Chinese text (2006) Proceedings of the 5th SIGHAN Workshop on Chinese Language Processing, pp. 108-117
2. Capenter, B., Character language models for Chinese word segmentation and named entity recognition (2006) ACL, pp. 169-172
3. Chang, C.H., Chen, C.D., A study on integrating Chinese word segmentation and part-of-speech tagging (1993) Communications of COLIPS, 3 (2), pp. 69-77
4. Papageorgiou, C.P., Japanese word segmentation by hidden markov model (1994) Proceedings of the HLT Workshop, pp. 283-288
5. Nguyen, C.T., Nguyen, T.K., Phan, X.H., Nguyen, L.M., Ha, Q.T., Vietnamese word segmentation with CRFs and SVMs: An investigation (2006) Proceedings of the 20th PACLIC, pp. 215-222. , Wuhan, China
6. Dinh, D., Vu, T., A maximum entropy approach for vietnamese word segmentation (2006) Proceedings of 4th RIVF VietNam, pp. 12-16
7. Dinh, D., Hoang, K., Nguyen, V.T., Vietnamese word segmentation (2001) The 6th NLPRS, pp. 749-756. , Tokyo, Japan
8. Peng, F., Feng, F., McCallum, A., Chinese segmentation and new word detection using conditional randome field (2004) Proceedings of COLING, pp. 562-568
9. Peng, F., Huang, X., Schuurmans, D., Wang, S., Text classification in Asian language without word segmentation (2003) Proceedings of 6th IRAL, pp. 41-48
10. Le, H.P., Nguyen, T.M.H., Roussanaly, A., Ho, T.V., A hybrid approach to word segmentation of vietnamese text (2008) Proceedings of 2nd LATA
11. Zhou, J.S., Dai, X.Y., Ni, R.Y., Chen, J.J., A hybrid approach to Chinese word segmentation around CRFs Proceedings of 4nd SIGHAN Workshop on Chinese Language Processing
12. Chen, K.J., Liu, S.H., (1992) Proceedings of the 15nd COLING
13. Ha, L.A., (2003) Proceedings of the Corpus Linguistics 2003, , Lancaster, UK
14. Xue, N., Converse, S.P., Combining classifiers for Chinese word segmentation (2002) First SIGHAN Workshop attached with the 19th COLING, pp. 57-63
15. Nguyen, P.T., Nguyen, V.V., Le, A.C., Vietnamese word segmentation using hidden markov model (2003) International Workshop for Computer, Information, and Communication Technologies on State of the Art and Future Trends of Information Technologies in Korea and Vietnam
16. Foo, S., Li, H., (2004) Information Processing & Management: An International Journal, 40 (1), pp. 161-190
17. Nguyen, T.H., Word segmentation for vietnamese text categorization: An online corpus approach (2006) Proceedings of 4th RIVF, , Ho Chi Minh, VietNam
18. Lu, X., Towards a hybrid model for Chinese word segmentation (2005) Proceedings of 4th SIGHAN Workshop on Chinese Language Processing
19. Vu, D., Nguyen, N.L., Dinh, D., (2006) ICT.rda'06, D? Lat.
20. Nguyen, T.M.H., Vu, X.L., Le, H.P., (2003) ICT.rda'03, Ha Noi.s