

A hybrid approach to word segmentation of Vietnamese texts

Phuong L.H., Huyen N.T.M., Roussanaly A., Vinh H.T.

LORIA, Nancy, France; Vietnam National University, Hanoi, Viet Nam; IFI, Hanoi, Viet Nam

Abstract: We present in this article a hybrid approach to automatically tokenize Vietnamese text. The approach combines both finite-state automata technique, regular expression parsing and the maximal-matching strategy which is augmented by statistical methods to resolve ambiguities of segmentation. The Vietnamese lexicon in use is compactly represented by a minimal finite-state automaton. A text to be tokenized is first parsed into lexical phrases and other patterns using pre-defined regular expressions. The automaton is then deployed to build linear graphs corresponding to the phrases to be segmented. The application of a maximal- matching strategy on a graph results in all candidate segmentations of a phrase. It is the responsibility of an ambiguity resolver, which uses a smoothed bigram language model, to choose the most probable segmentation of the phrase. The hybrid approach is implemented to create vnTokenizer, a highly accurate tokenizer for Vietnamese texts. ?? 2008 Springer-Verlag Berlin Heidelberg.

Index Keywords: Applications; Computational linguistics; Finite automata; Graph theory; Linguistics; Natural language processing systems; Robots; Semantics; Statistical methods; Translation (languages); Bigram language models; Hybrid approaches; Linear graphs; Regular expressions; Tokenizer; Word segmentations; Automata theory

Year: 2008

Source title: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)

Volume: 5196 LNCS

Page : 240-249

Cited by: 1

Link: [Scopus Link](#)

Correspondence Address: Phuong, L. H.; LORIA, Nancy, France

Sponsors: Fundacio Caixa Tarragona

Conference name: 2nd International Conference on Language and Automata Theory and Applications, LATA 2008

Conference date: 13 March 2008 through 19 March 2008

Conference location: Tarragona

Conference code: 74250

ISSN: 3029743

ISBN: 3540882812; 9783540882817

DOI: 10.1007/978-3-540-88282-4_23

Language of Original Document: English

Abbreviated Source Title: Lecture Notes in Computer Science (including subseries Lecture Notes in

1. Phuong, L.H., LORIA, Nancy, France
2. Huy?n, N.T.M., Vietnam National University, Hanoi, Viet Nam
3. Roussanaly, A., LORIA, Nancy, France
4. Vinh, H.T., IFI, Hanoi, Viet Nam

References:

1. Maurel, D., (2003) Electronic Dictionaries and Acyclic Finite-State Automata: A , State of The Art. Grammars and Automata for String Processing
2. Daciuk, J., Mihov, S., Watson, B.W., Watson, R.E., Incremental Construction of Minimal Acyclic Finite-State Automata (2000) Computational Linguistics, 26 (1)
3. Language Resource Management-Word Segmentation of Written Texts for Mono-lingual and Multi-lingual Information Processing - Part I: General Principles and Methods. Technical Report, ISO (2006), ISO/TC 37/SC 4 AWI N309Jelinke, F., Mercer, R.L., Interpolated estimation of Markov source parameters from sparse data (1980) Proceedings of the Workshop on Pattern Recognition in Practice, , The Netherlands
4. Schmid, H.: Tokenizing. In: L?deling, A., Kyt??, M. (eds.) Corpus Linguistics. An International Handbook. Mouton de Gruyter, Berlin (2007)Gao, J., (2006) Chinese Word Segmentation and Named Entity Recognition: A Pragmatic Approach, , Computational Linguistics
5. Chen, S.F., Goodman, J., An Empirical Study of Smoothing Techniques for Language Modeling (1996) Proceedings of the 34th Annual Meeting of the ACL
6. Wong, P., Chan, C., Chinese Word Segmentation based on Maximum Matching and Word Binding Force (1996) Proceedings of the 16th Conference on Computational Linguistics, Copenhagen, DK