An experimental study on lexicalized statistical parsing for Vietnamese

Le A.-C., Nguyen P.-T., Vuong H.-T., Pham M.-T., Ho T.-B.
College of Technology, Vietnam National University of Hanoi, E3-144 Xuan Thuy, Cau Giay, Hanoi, Viet Nam; Japan Advanced Institute of Science and Technology, Asahidai, Nomi, Ishikawa 923-1292, Japan

Abstract: Syntactic parsing is a central problem and a challenge in the field of natural language processing. It attracts many studies and consequently there exists the effective parsers for several popular languages such as English and Chinese. For Vietnamese parsing, there have been a few studies focusing on this problem, these studies lack of applying modern techniques, and no popular parser has been released. This paper presents the first study on developing a Vietnamese wide coverage parser based on lexicalized probabilistic context free grammar (LPCFG) and using a standard parsed corpus (similar to Penn Treebank). In this paper the Bikel's parser is modified to analyze Vietnamese. We also provide a comparison based on investigating different parsing models and different linguistic features. The best configuration achieves around 78% of F-score. © 2009 IEEE.
Index Keywords: Central problems; Experimental studies; F-score; Linguistic features; NAtural language processing; Parsed corpora; Probabilistic context free grammars; Syntactic parsing; Treebanks; Computational linguistics; Context free grammars; Knowledge engineering; Natural language processing systems; Query languages; Standardization; Systems engineering; Formal languages

Year: 2009
Source title: KSE 2009 - The 1st International Conference on Knowledge and Systems Engineering
Art. No.: 5361714
Page : 162-167
Link: Scorpus Link
Correspondence Address: Le, A.-C.; College of Technology, Vietnam National University of Hanoi, E3-144 Xuan Thuy, Cau Giay, Hanoi, Viet Nam
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.2009.41
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:

- Le, A.-C., College of Technology, Vietnam National University of Hanoi, E3-144 Xuan Thuy, Cau Giay, Hanoi, Viet Nam
- Nguyen, P.-T., College of Technology, Vietnam National University of Hanoi, E3-144 Xuan Thuy, Cau Giay, Hanoi, Viet Nam
- Vuong, H.-T., College of Technology, Vietnam National University of Hanoi, E3-144 Xuan Thuy, Cau Giay, Hanoi, Viet Nam
- Pham, M.-T., College of Technology, Vietnam National University of Hanoi, E3-144 Xuan Thuy, Cau Giay, Hanoi, Viet Nam
- Ho, T.-B., Japan Advanced Institute of Science and Technology, Asahidai, Nomi, Ishikawa 923-1292, Japan

References:

- Agirre, E., Baldwin, T., Martinez, D., Improving parsing and PP attachment performance with sense information (2008) Proceedings of ACL, pp. 317-32
- Bikel, D.M., David, C., Two statistical parsing models applied to the Chinese treebank (2000) Proceedings of the Second Chinese Language Processing Workshop
- Bikel, D.M., (2004) On Parameter Space of Generative Lexicalized Statistical Parsing Models, , Ph.D dissertation
- Booth, T.L., Thompson, R.A., Applying probability measures to abstract languages (1973) IEEE Transactions on Computers, C-22 (5), pp. 442-450
- Candito, M., Crabb, B., Seddah, D., On statistical parsing of french with supervised and semi-supervised strategies (2009) Proceedings of the EACL 2009, Workshop on Computational Linguistic Aspects of Grammatical Inference, pp. 49-57
- Collins, M.J., A new statistical parser based on bigram lexical dependencies (1996) Proceedings of ACL, pp. 184-191
- Collins, M., Three generative, lexicalised models for statistical parsing (1997) Proceedings of ACL-EACL '97, pp. 16-23
- Collins, M., (1999) Head-Driven Statistical Models for Natural Language Parsing, , Ph.D dissertation
- Marcus, M.P., Marcinkiewicz, M.A., Santorini, B., Building a large annotated corpus of english: The penn treebank (1993) Journal of Computational Linguistics, 22 (2), pp. 313-330
- Magermaa, D., Statistical decision-tree models for parsing (1995) Proceedings of ACL, pp. 276-283
- Nguyen, Q.T., Le, T.H., Vietnamese syntactic parsing using the lexicalized probabilistic contextfree grammar (2007) Proceedings of the FAIR conference, pp. 9-10. , Nha Trang, Vietnam
- Nguyen, P.T., Vu, X.L., Nguyen, T.M.H., Nguyen Van, H., Le, H.P., Building a large syntactically-annotated corpus of Vietnamese (2009) Proceedings of the 3rd Linguistic Annotation Workshop (LAW), , ACL-IJCNLP 2009
- Watson, R., Briscoe, T., Carroll, J., Semisupervised training of a statistical parser from unlabeled partially-bracketed data (2007) Proceedings of the Tenth International Conference on Parsing Technologies, p. 2332
- Adwait, R., A maximum entropy model for part-of-speech tagging (1996) Proceedings of the Conference on Empirical Methods in Natural Language Processing, pp. 133-142
- Cao, H., Zhao, T., Li, S., Parsing Penn Chinese treebank (CTB) with head-driven model (2007) Gaojishu Tongxin/Chinese High Technology Letters, 17 (1), pp. 15-20