Transformation rule learning without rule templates: A case study in part of speech tagging

Bach N.X., Cuong L.A., Ha N.V., Binh N.N.

College of Technology, Vietnam National University, Hanoi; Information Technology Institute, Vietnam National University, Hanoi

Abstract: Part of speech (POS) tagging is an important problem and is one of the first steps included in many tasks in natural language processing. It affects directly on the accuracy of many other problems such as Syntax Parsing, Word Sense Disambiguation, and Machine Translation. Stochastic models solve this problem relatively well, but they still make mistakes. Transformation-based learning (TBL) is a solution which can be used to improve stochastic taggers by learning a set of transformation rules. However, its rule learning algorithm has the disadvantages that rule templates must be prepared by hand and only rules are instances of rule templates can be generated. In this paper, we propose a model to learn transformation rules without rule templates. This model considers the rule learning problem as a feature selection problem. Experiments on Penn TreeBank showed that the proposal model reduces errors of stochastic taggers with some tags. ?? 2008 IEEE.

Index Keywords: Artificial intelligence; Computational linguistics; Computer aided language translation; Education; Feature extraction; Information technology; Information theory; Laws and legislation; Learning algorithms; Learning systems; Linguistics; Mathematical models; Natural language processing systems; Speech; Speech processing; Speech transmission; Stochastic programming; Technology; Case studies; Feature selection; International conferences; Language processing; Machine translation; NAtural language processing; Part-of-Speech tagging; Rule learning; Transformation rules; Transformation-based learning; Treebank; Web information; Word-sense disambiguation; Stochastic models

Year: 2008
Source title: Proceedings - ALPIT 2008, 7th International Conference on Advanced Language Processing and Web Information Technology
Art. No.: 4584333
Page : 9-14
Link: Scorpus Link
Correspondence Address: Bach, N. X.; College of Technology, Vietnam National University, Hanoi
Conference name: ALPIT 2008, 7th International Conference on Advanced Language Processing and Web Information Technology
Conference date: 23 July 2008 through 25 July 2008
Conference location: Liaoning
Conference code: 73559
ISBN: 9.78E+12
DOI: 10.1109/ALPIT.2008.73
Language of Original Document: English

Abbreviated Source Title: Proceedings - ALPIT 2008, 7th International Conference on Advanced Language Processing and Web Information Technology

Document Type: Conference Paper

Source: Scopus

Authors with affiliations:

1. Bach, N.X., College of Technology, Vietnam National University, Hanoi
2. Cuong, L.A., College of Technology, Vietnam National University, Hanoi
3. Ha, N.V., Information Technology Institute, Vietnam National University, Hanoi
4. Binh, N.N., College of Technology, Vietnam National University, Hanoi

References:

1. Berger, A.L., Della Pietra, S.A., Della Pietra, V.J., A Maximum Entropy Approach to Natural Language Processing (1996) Association for Computational Linguistics, pp. 39-71
2. Brants, T., TnT, A., Statistical Part-of-Speech Tagger (2000) Proc. of the 6th Applied NLP Conf, pp. 224-231
3. Brill, E., Proc. of the 12th National Conference on AI (1994) A Report of Recent Progress in Transformation-Based Error-Driven Learning, pp. 722-727
4. Brill, E., Transformation-Based Error-Driven Learning and Natural Language Processing: A Case Study in Part of Speech Tagging (1995) Association for Computational Linguistics, pp. 543-565
5. Carrasco, R., Gelbukh, A., Evaluation of TnT Tagger for Spanish (2003) Proc. of the 4th Mexican Inter Conf on Computer Science, p. 18
6. Florian, R., Ngai, G., Transformation-Based Learning in the fast lane (2001) Proc. of North America ACL, pp. 1-8
7. Gimenez, J., Marquez, L., SVMTool: A general POS tagger generator based on Support Vector Machines (2004) Proc. of the 4th Inter Conf on Language Resources and Evaluation, , Portugal
8. Gimenez, J., Marquez, L., (2006) SVMTool Technical Manual v1.3, , TALP Research Center, Universitat Politcnica de Catalunya, Barcelona
9. Kohavi, R., A study of cross-validation and bootstrap for accuracy estimation and model selection (1995) Proc. of the 14th Inter Joint Conf on AI, pp. 1137-1145
10. The Penn Treebank Project, , http://www.cis.upenn.edu/tree-bank, University of Pennsylvania, Available at
11. Ratnaparkhi, A., A Maximum Entropy Model for Part-Of-Speech Tagging (1996) Association for Computational Linguistics, pp. 133-142
12. Ratnaparkhi, A., (1998) Maximum Entropy Models For Natural Language Ambiguity Resolution, , Ph.D. Dissertation, the University of Pennsylvania
13. Statistical natural language processing and corpus-based computational linguistics: An annotated list of resources, , http://wwwnlp.stanford.edu/links/statnlp.html, Available at
14. Tan, L., Taniar, D., Adaptive estimated maximumentropy distribution model (2007) Information Sciences: An International Journal, pp. 3110-3128
15. Zitouni, I., Sorensen, J.S., Sarikaya, R., Maximum Entropy Based Restoration of Arabic Diacritics (2006) Proc. of the 21st Inter Conf on Computational Linguistics, pp. 577-584