P.O.Box 8718, Beijing 100080, China	Journal of Software, March 2007,18(3):538-546
E-mail: jos@iscas.ac.cn	ISSN 1000-9825, CODEN RUXUEW, CN 11-2560/TP
http://www.jos.org.cn	Copyright © 2007 by *Journal of Software*

基于句子对齐的汉语句法结构推导的计算模型

王厚峰, 王 波

Full-Text PDF Submission Back

王厚峰, 王 波
(北京大学 信息科学技术学院 计算语言学研究所,北京 100871)
作者简介: 王厚峰(1965－),男,湖北天门人,博士,教授,CCF高级会员,主要研究领域为自然语言处理.王波(1982－),男,硕士生,主要研究领域为自然语言处理.
联系人: 王厚峰 Phn: +86-10-62753081 ext 106, E-mail: wanghf@pku.edu.cn

Received 2006-01-26; Accepted 2006-04-12

Abstract

This paper introduces an unsupervised learning framework of Chinese syntactic structure based sentences similarity. First, all sentence pairs in the Chinese sentence corpus are aligned, and each pair is partitioned into similarity segmentations and different ones which alternately occur, Then, aligned similarity segmentations or different ones are selected as potential constituent candidates based on the strategy of similarity priority or of difference priority respectively. As the boundary friction may be introduced in the later step, its disambiguation is further carried out. Finally, by inducing sentence constituents, the syntactic structures are learned. In order to reduce word sparseness in the process, some words are replaced by classes in advance. Three forms of the sentence units, such as the sequence of words, the sequence of POS (part of speech)-tags and the sequence of words with POS-tag, are examined and the learned syntactic structures are evaluated respectively. The results show that different priority strategy achieves a better performance than the similarity one, and the Fs are above 46% for all three forms, with the best one being 49.52%, which is better than those having been reported.

Wang HF, Wang B. A computational model for Chinese syntactic structure induction based on sentence alignment. *Journal of Software*, 2007,18(3):538-546.
DOI: 10.1360/jos180538
http://www.jos.org.cn/1000-9825/18/538.htm

摘要

基于句子的相似性,提出了无指导的汉语句法结构推导方法.基本思想是:首先,在汉语句子库的基础上,通过句对之间的对齐,得到交替的相同片断和相异片断.然后,根据相同片断优先或相异片断优先策略,选取相应的对齐片断作为句子成分候选,并对可能因片断交叉而导致边界摩擦的候选进行歧义消解.最后,通过逐步归约句子成分,推导出汉语句法结构树.为了避免对齐过程中词的稀疏问题,还对部分具有明显规律的词事先作了归类处理.分别以词、词性以及词联合词性作为句子基本构成单元,评测了推导的句法结果.测试结果表明:对于3种构成单元,相异片断优先归约得到的结果的F值都超过了46%,均优于相同片断优先归约所得到的结果,最好的达到了49.52%,好于已报道的结果.

基金项目：Supported by the National Natural Science Foundation of China under Grant Nos.60473138, 60675035 (国家自然科学基金)

References:

[1] Brill E. Automatic grammar induction and parsing free text: A transformation-based approach. In: Proc. of the 31st Annual Meeting of the Association for Computational Linguistics. 1993. 259-265. http://acl.ldc.upenn.edu/P/P93/

[2] Pereira F, Schabes Y. Inside-Outside reestimation from partially bracketed corpora. In: Pros. of the 30th Annual Meeting of the Association for Computational Linguistics. 1992. 128-135. http://acl.ldc.upenn.edu/P/P92/

[3] Nakamura K, Matsumoto M. Incremental learning of context free grammar. In: Adriaans P, et al., eds. Proc. of the Grammatical Inference: Algorithms and applications (ICGI-2002). LNAI 2484, Springer-Verlag, 2002. 174-184.

[4] Grunwall P. A minimum description length approach to grammar inference. In: Wermter S, Riloff E, Scheler G, eds. Proc. of the Symbolic, Connectionist and Statistical Approaches to Learning for Natural Language Processing. LNCS 1040, Springer-Verlag, 1996. 203-216.

[5] Wolff GJ. Unsupervised grammar induction in a framework of information compression by multiple alignment, unification and search. In: de la Higuera C, Adriaans P, van Zaanen M, Oncina J, eds. Proc. of the Workshorp at ECML/PKDD2003: Learning Context-Free Grammars. 2003. 114-124. http://ilk.uvt.nl/~mvzaanen/ECMLPKDD/talks.html

[6] Klein D, Manning CD. A generative constituent-context model for improved grammar induction. In: Proc. of the 40th Annual Meeting of the Association for Computational Linguistics. 2002. 128-135. http://acl.ldc.upenn.edu/P/P02/

[7] Klein D. The unsupervised learning of natural language structure [Ph.D. Thesis]. Stanford University, 2005.

[8] Clark A. Unsupervised induction of stochastic context-free grammars using distributional clustering. In: Daelemans W, Zajac R, eds. Proc. of the CoNLL 2001. Morgan Kaufmann. 2001. 105-112.

[9] Adriaans P, Trautwein M, Vervoort M. Towards high speed grammar induction on large text corpora. In: Hlavac V, Feffrey G, Wiedermann J, eds. Proc. of the SOFSEM-2000, Theory and Practice of Informatics. LNCS 1963, Springer-Verlag, 2000. 173-186.

[10] van Zaanen M. Bootstrapping syntax and recursion using alignment-based learning. In: Langley P, ed. Proc. of the 17th Int'l Conf. on Machine Learning. Morgan Kaufmann. 2000. 1063-1070.

[11] van Zaanen M, Adriaans P. Alignment-Based learning versus EMILE: A comparison. In: Krose B, de Rijke M, Schreiber G, van Someren M, eds. Proc. of the Belgian-Dutch Conf. on Artificial Intelligence (BNAIC). 2001. 315-322. http://www.ics.mq.edu.au/~menno/research/publications

[12] Cicekli I, Guvenir HA. Learning translation templates from bilingual translation examples. In: Applied Intelligence, vol.15. 2001. 57-76.

[13] Smith NA, Eisner J. Annealing techniques for unsupervised statistical language learning. In: Proc. of the 42nd Annual Meeting the Association for Computational Linguistics. 2004. 487-94. http://acl.ldc.upenn.edu/P/P04/

[14] Yu SW, et al. The Grammatical Knowledge-Base of Contemporary Chinese—A Complete Specification. 2nd ed., Beijing: Tsinghua University Press, 2003 (in Chinese).

附中文参考文献:
[14] 俞士汶,等.现代汉语语法信息词典详解.第2版,北京:清华大学出版社,2003.