



Search bar with magnifying glass icon and settings gear icon

Home Contact Us

Look Inside

Get Access

Find out how to access preview-only content

### Chapter

Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data

Volume 8202 of the series Lecture Notes in Computer Science pp 238-246

# Exploiting Lexicalized Statistical Patterns in Chinese Linguistic Analysis

Yu Zhao, Maosong Sun



Yu Zhao and Maosong Sun  
Department of Computer Science and Technology,  
State Key Lab on Intelligent Technology and Systems,  
National Lab for Information Science and Technology,  
Tsinghua University, Beijing 100084, China

Other

- » Ex
- » Ab
- » Re
- » Pe
- » Ad

**Abstract.** The web corpus has been used for linguistic analysis with the help of search engines. In this paper, we describe the concept of lexicalized patterns, which we exploit to obtain statistical information using the simple string matching strategy via search engines. We discuss the usage of lexicalized statistical patterns at three linguistic levels of Chinese analysis: lexical, syntactic and semantic. We develop a specialized search engine to get frequency counts for these patterns on SogouT corpus. Experimental results show that lexicalized statistical patterns are effective on analyzing the cohesion of phrases, determining the phrasal category and discovering patient objects.

**Keywords:** Lexicalized statistical patterns, Chinese linguistic analysis, Web corpus, Natural language processing.

#### 1 Introduction

Most of current statistical natural language processing (NLP) systems are based on manually annotated corpora. For example, the Penn Treebank is highly time-consuming and labor-intensive to build. The main processing task is the lack of reliable information. Let us consider the task of determining the most significant issues in a document. For instance, 会议(Conference) 正在(are) 进行(being) 讨论(discussing) 问题(problem) 的(the) 报告(report) 是(is) 由(by) 谁(who) 提出(put forward) 的(the) 呢(ne)?

Therefore, a growing number of researchers have been realizing the potential of automatic corpus for NLP tasks. The key advantage of web corpora lies in that they are easy to obtain and update.

The 2008 version is available online at <http://www.ccl.umd.edu/2008/>.

M. Sun et al. (Eds.), CCL and NLP-NAACL 2013, LNCS 8202, pp. 238–246, 2013.  
© Springer-Verlag Berlin Heidelberg 2013

Look Inside

Buy chapter Buy this eBook

Get Access \* Final gross prices may vary according to local VAT.

## Abstract

The web corpus has been used for linguistic analysis with the help of search engines. In this paper, we describe the concept of lexicalized patterns, which we exploit to obtain statistical information using the simple string matching strategy via search engines. We discuss the usage of lexicalized statistical patterns at three linguistic levels of Chinese analysis: lexical, syntactic and semantic. We develop a specialized search engine to get frequency counts for these patterns on SogouT corpus. Experimental results show that lexicalized statistical patterns are effective on analyzing the cohesion of phrases, determining the phrasal category and discovering patient objects.

## Keywords

Lexicalized statistical pattern Chinese linguistic analysis Web corpus  
Natural language processing

### Supplementary Material (0)

### References (9)

#### References

1. Bansal, M., Klein, D.: *Web-scale features for full-scale parsing*. In: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, vol. 1 (2011)
  2. Curran, J.R., Moens, M.: *Scaling context space*. In: *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, Philadelphia, PA, USA* (2002)
  3. Keller, F., Lapata, M., Ourioupina, O.: *Using the web to overcome data sparseness*. In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing, Philadelphia* (2002)
  4. Lapata, M., Keller, F.: *The Web as a baseline: Evaluating the performance of unsupervised Web-based models for a range of NLP tasks*. In: *Proceedings of HLT-NAACL* (2004)
  5. Volk, M.: *Exploiting the WWW as a corpus to resolve PP attachment ambiguities*. In: *Proceedings of the Corpus Linguistics 2001 Conference, Lancaster, UK*, pp. 601–606 (2001)
  6. Yates, A., Schoenmackers, S., Etzioni, O.: *Detecting parser errors using web-based semantic filters*. In: *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics* (2006)
  7. Yuan, Y.: *A Cognitive Investigation and Fuzzy Classification of Word-class in Mandarin Chinese*. Shanghai Educational Publishing House (2009)
  8. Yuan, Y.: *Beijing Language and Culture University Press* (2010)
  9. Zhang, Y., Clark, S.: *Syntactic processing using the generalized perceptron and beam search*. *Computational Linguistics* 37(1), 105–151 (2011)
- [CrossRef](#)

### About this Chapter

#### Title

Exploiting Lexicalized Statistical Patterns in Chinese Linguistic Analysis

#### Book Title

Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data

#### Book Subtitle

12th China National Conference,

#### Topics

Language Translation and Linguistics  
Artificial Intelligence (incl. Robotics)

Document Preparation and Text Processing  
Information Systems and Communication Service

#### Keywords

Lexicalized statistical pattern  
Chinese linguistic analysis

#### Editors

Maosong Sun  (19)  
Min Zhang  (20)  
Dekang Lin  (21)  
Haifeng Wang  (22)

#### Editor Affiliations

19. Department of Computer Science and Technology, Tsinghua University  
20. Horizon Doctoral Training Centre, School of Computer Science, University of Nottingham

CCL 2013 and First International Symposium, NLP-NABD 2013, Suzhou, China, October 10-12, 2013. Proceedings

---

**Pages**

pp 238-246

---

**Copyright**

2013

---

**DOI**

10.1007/978-3-642-41491-6\_22

---

**Print ISBN**

978-3-642-41490-9

---

**Online ISBN**

978-3-642-41491-6

---

**Series Title**

[Lecture Notes in Computer Science](#)

---

**Series Volume**

8202

---

**Series ISSN**

0302-9743

---

**Publisher**

Springer Berlin Heidelberg

---

**Copyright Holder**

Springer-Verlag Berlin Heidelberg

---

**Additional Links**

[About this Book](#)

Web corpus

Natural language processing

---

**Industry Sectors**

[Electronics](#)

[Telecommunications](#)

[IT & Software](#)

---

**eBook Packages**

[eBook Package english](#)

[Computer Science](#)

[eBook Package english full Collection](#)

21. Google Inc.

22. Baidu Inc.

---

**Authors**

[Yu Zhao](#) <sup>(23)</sup>

[Maosong Sun](#) <sup>(23)</sup>

---

**Author Affiliations**

23. Department of Computer Science and Technology, State Key Lab on Intelligent Technology and Systems, National Lab for Information Science and Technology, Tsinghua University, Beijing, 100084, China

Over 9 million scientific documents at your fingertips

Browse by Discipline



---

**Our Content**

Journals

Books

Book Series

---

**Other Sites**

Springer.com

SpringerProtocols

SpringerMaterials

---

**Help & Contacts**

Contact Us

Feedback Community

Impressum

Protocols

Reference Works

© Springer International Publishing AG, Part of Springer Science+Business Media | [Privacy Policy](#), [Disclaimer](#), [General Terms & Conditions](#)

Media

沪ICP备13017623号

Not logged in Unaffiliated 122.70.132.162