Dependency Parsing: Past, Present, and Future

Wenliang Chen and Zhenghua Li and Min Zhang
School of Computer Science and Technology,
Soochow University, China
{wlchen, Zhli13, mzhang}@suda.edu.cn

Dependency parsing has gained more and more interest in natural language processing in recent years due to its simplicity and general applicability for diverse languages. The international conference of computational natural language learning (CoNLL) has organized shared tasks on multilingual dependency parsing successively from 2006 to 2009, which leads to extensive progress on dependency parsing in both theoretical and practical perspectives. Meanwhile, dependency parsing has been successfully applied to machine translation, question answering, text mining, etc.

To date, research on dependency parsing mainly focuses on data-driven supervised approaches and results show that the supervised models can achieve reasonable performance on in-domain texts for a variety of languages when manually labeled data is provided. However, relatively less effort is devoted to parsing out-domain texts and resource-poor languages, and few successful techniques are bought up for such scenario. This tutorial will cover all these research topics of dependency parsing and is composed of four major parts. Especially, we will survey the present progress of semi-supervised dependency parsing, web data parsing, and multilingual text parsing, and show some directions for future work.

In the first part, we will introduce the fundamentals and supervised approaches for dependency parsing. The fundamentals include examples of dependency trees, annotated treebanks, evaluation metrics, and comparisons with other syntactic formulations like constituent parsing. Then we will introduce a few mainstream supervised approaches, i.e., transition-based, graph-based, easy-first, constituent-based dependency parsing. These approaches study dependency parsing from different perspectives, and achieve comparable and state-of-the-art performance for a wide range of languages. Then we will move to the hybrid models that combine the advantages of the above approaches. We will also introduce recent work on efficient parsing techniques, joint lexical analysis and dependency parsing, multiple treebank exploitation, etc.

In the second part, we will survey the work on semi-supervised dependency parsing techniques. Such work aims to explore unlabeled data so that the parser can achieve higher performance. This tutorial will present several successful techniques that utilize information from different levels: whole tree level, partial tree level, and lexical level. We will discuss the advantages and limitations of these existing techniques.

In the third part, we will survey the work on dependency parsing techniques for domain adaptation and web data. To advance research on out-domain parsing, researchers have organized two shared tasks, i.e., the CoNLL 2007 shared task and the shared task of syntactic analysis of non-canonical languages (SANCL 2012). Both two shared tasks attracted many participants. These participants tried different techniques to adapt the parser trained on WSJ texts to out-domain texts with the help of large-scale unlabeled data. Especially, we will present a brief survey on text normalization, which is proven to be very useful for parsing web data.

In the fourth part, we will introduce the recent work on exploiting multilingual texts for dependency parsing, which falls into two lines of research. The first line is to improve supervised dependency parser with multilingual texts. The intuition behind is that ambiguities in the target language may be unambiguous in the source language. The other line is multilingual transfer learning which aims to project the syntactic knowledge from the source language to the target language.

Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Tutorial Abstracts, pages 14–16, Dublin, Ireland, August 23-29 2014.

In the fifth part, we will conclude our talk by discussing some new directions for future work.

Outline

- Part A: Dependency parsing and supervised approaches
  - A.1 Introduction to dependency parsing
  - A.2 Supervised methods
  - A.3 Non-projective dependency parsing
  - A.4 Probabilistic and generative models for dependency parsing
  - A.5 Other work
- Part B: Semi-supervised dependency parsing
  - B.1 Lexical level
  - B.2 Partial tree level
  - B.3 Whole tree level
  - B.4 Other work
- Part C: Parsing the web and domain adaptation
  - C.1 CoNLL 2007 shared task (domain adaptation subtask)
  - C.2 Works on domain adaptation
  - C.3 SANCL 2012 (parsing the web)
  - C.4 Text normalization
  - C.5 Attempts and challenges for parsing the web
- Part D: Multilingual dependency parsing
  - D.1 Dependency parsing on bilingual text
  - D.2 Multilingual transfer learning for resource-poor languages
  - D.3 Other work
- Part E: Conclusion and open problems

Instructors

Wenliang Chen received his Bachelor degree in Mechanical Engineering and Ph.D. degree in computer science from Northeastern University (Shenyang, China) in 1999 and 2005, respectively. He joined Soochow University since 2013 and is currently a professor in the university. Prior to joining Soochow University, he was a research scientist in the Institute for Infocomm Research of Singapore from 2011 to 2013. From 2005 to 2010, he worked as an expert researcher in NICT, Japan. His current research interests include parsing, machine translation, and machine learning. He is currently working on a syntactic parsing project where he applies semi-supervised learning techniques to explore the information from large-scale data to improve Dependency parsing. Based on the semi-supervised techniques, he developed a dependency parser, named DuDuPlus (http://code.google.com/p/duduplus/), for the research and industry communities.

Zhenghua Li is currently an assistant professor in Soochow University. He received his PhD in Computer Science and Technology from Harbin Institute of Technology (HIT) in April 2013. Zhenghuas research interests include natural language processing and machine learning. More specifically, his PhD research focuses on the dependency parsing of the Chinese language using discriminative machine-learning approaches. He has been working on joint POS tagging and dependency parsing (EMNLP 2011, COLING 2012), and multiple treebank exploitation for dependency parsing (ACL 2012).

Min Zhang, a distinguished professor and Director of the Research Center of Human Language Technology at Soochow University (China), received his Bachelor degree and Ph.D. degree in computer science from Harbin Institute of Technology in 1991 and 1997, respectively. From 1997 to 1999, he worked as a postdoctoral research fellow in Korean Advanced Institute of Science and Technology in South Korea. He began his academic and industrial career as a researcher at Lernout & Hauspie Asia Pacific (Singapore) in Sep. 1999. He joined Infotalk Technology (Singapore) as a researcher in 2001 and became a senior research manager in 2002. He joined the Institute for Infocomm Research (Singapore) as a research scientist in Dec. 2003. His current research interests include machine translation, natural language processing, information extraction, social network computing and Internet intelligence. He has co-authored more than 150 papers in leading journals and conferences, and co-edited 10 books/proceedings published by Springer and IEEE. He was the recipient of several awards in China and oversea. He is the vice president of COLIPS (2011-2013), the elected vice chair of SIGHAN/ACL (2014-2015), a steering committee member of PACLIC (2011-now), an executive member of AFNLP (2013-2014) and a member of ACL (since 2006). He supervises Ph.D students at National University of Singapore, Harbin Institute of Technology and Soochow University.