Unsupervised Natural Language Parsing (Introductory Tutorial)

Kewei Tu[1], **Yong Jiang**[2], **Wenjuan Han**[3], **Yanpeng Zhao**[4]
[1]School of Information Science and Technology, ShanghaiTech University
[2]Alibaba DAMO Academy, Alibaba Group
[3]School of Computing, National University of Singapore, Singapore
[4]ILCC, University of Edinburgh
tukw@shanghaitech.edu.cn
yongjiang.jy@alibaba-inc.com
dcshanw@nus.edu.sg
yanp.zhao@ed.ac.uk

1 Introduction

Syntactic parsing is an important task in natural language processing that aims to uncover the syntactic structure (e.g., a constituent or dependency tree) of an input sentence. Such syntactic structures have been found useful in downstream tasks such as semantic parsing, relation extraction, and machine translation.

Supervised learning is the main technique used to automatically learn a syntactic parser from data. It requires the training sentences to be manually annotated with their correct parse trees. A major challenge faced by supervised parsing is that syntactically annotated sentences are not always available for a target language or domain and building a high-quality annotated corpus is very expensive and time-consuming.

A radical solution to this challenge is *unsupervised parsing*, sometimes also called *grammar induction*, which learns a parser from training sentences without parse tree annotations. Unsupervised parsing can also serve as the basis for semi-supervised and transfer learning of syntactic parsers when there exist both unannotated sentences and (in-domain or out-of-domain) annotated sentences. In addition, the research of unsupervised parsing is deemed interesting in the field of machine learning because it is a representative task of unsupervised structured prediction, and in the field of cognitive science because it inspires and verifies cognitive research of human language acquisition.

The research on unsupervised parsing has a long history, dating back to theoretical studies in 1960s (Gold, 1967) and algorithmic and empirical studies in 1970s (Baker, 1979). Although deemed an interesting topic by the NLP community, unsupervised parsing had received much less attention than supervised parsing over the past few decades.

More recently, however, there has been a resurgence of interest in unsupervised parsing, with more than ten papers on unsupervised parsing published in top NLP and AI venues over the past two years, including a best paper at ICLR 2019 (Shen et al., 2019), a best paper nominee at ACL 2019 (Shi et al., 2019), and a best paper nominee at EMNLP 2020 (Zhao and Titov, 2020). This renewed interest in unsupervised parsing can be attributed to the combination of two recent trends. First, there is a general trend in deep learning towards unsupervised training or pre-training. Second, there is an emerging trend in the NLP community towards finding or modeling linguistic structures in neural models. The research on unsupervised parsing fits these two trends perfectly.

Because of the renewed attention on unsupervised parsing and its relevance to the recent trends in the NLP community, we believe a tutorial on unsupervised parsing can be timely and beneficial to many *ACL conference attendees. The tutorial will introduce to the general audience what unsupervised parsing does and how it can be useful for and beyond syntactic parsing. It will then provide a systematic overview of major classes of approaches to unsupervised parsing, namely generative and discriminative approaches, and analyze their relative strengths and weaknesses. It will cover both decade-old statistical approaches and more recent neural approaches to give the audience a sense of the historical and recent development of the field. We also plan to discuss emerging research topics such as BERT-based approaches and visually grounded learning.

We expect that by taking this tutorial, one can not only obtain a deep understanding of the literature and methodology of unsupervised parsing and become well prepared for his own research into unsupervised parsing, but may also get inspirations from the ideas and techniques of unsupervised pars-

Proceedings of EACL: Tutorials, pages 1–5
April 19 - 20, 2021. ©2020 Association for Computational Linguistics

ing and apply or extend them to other NLP tasks that can potentially benefit from implicitly learned linguistic structures.

2 Overview

This will be a three-hour tutorial divided into five parts.

In the first part, we will introduce the unsupervised parsing task. We will start with the problem definition and discuss the motivations and applications of unsupervised parsing. For example, we will show that unsupervised parsing approaches can be extended for semi-supervised parsing (Jia et al., 2020) and cross-lingual syntactic transfer (He et al., 2019), and we will also show applications of unsupervised parsing approaches beyond syntactic parsing (e.g., in computer vision (Tu et al., 2013)). We will then discuss how to evaluate unsupervised parsing, including the evaluation metrics and typical experimental setups. We will promote standardized setups to enable meaningful empirical comparison between approaches (Li et al., 2020). Finally, we will give an overview of unsupervised parsing approaches to be discussed in the rest of the tutorial.

In the second and third parts, we will introduce in detail two major classes of approaches to unsupervised parsing, generative and discriminative approaches, and discuss their pros and cons.

The second part will cover generative approaches, which model the joint probability of the sentence and the corresponding parse tree. Most of the existing generative approaches are based on generative grammars, in particular context-free grammars and dependency models with valence (Klein and Manning, 2004). There are also featurized and neural extensions of generative grammars, such as Berg-Kirkpatrick et al. (2010); Jiang et al. (2016). We will divide our discussion of learning generative grammars into two parts: structure learning and parameter learning. Structure learning concerns finding the optimal set of grammar rules. We will introduce both probabilistic methods such as Stolcke and Omohundro (1994) and heuristic methods such as Clark (2007). Parameter learning concerns learning the probabilities or weights of a pre-specified set of grammar rules. We will discuss a variety of priors and regularizations designed to improve parameter learning, such as Cohen and Smith (2010), Tu and Honavar (2012), Noji et al. (2016), and (Jin et al., 2018). We will also discuss parameter learning algorithms such as expectation-maximization (Baker, 1979; Spitkovsky et al., 2010b), MCMC (Johnson et al., 2007) and curriculum learning (Spitkovsky et al., 2010a). After introducing approaches based on generative grammars, we will discuss recent approaches that are instead based on neural language models (Shen et al., 2018, 2019).

The third part will cover discriminative approaches, which model the conditional probability or score of the parse tree given the sentence. We will first introduce autoencoder approaches such as Cai et al. (2017), which contain an encoder that maps the sentence to an intermediate representation (such as a parse tree) and a decoder that tries to reconstruct the sentence. Their training objective is typically the reconstruction probability. We will then introduce variational autoencoder approaches such as Kim et al. (2019), which has a similar model structure to autoencoder approaches but uses the evidence lower bound as the training objective. Finally, we will briefly discuss other discriminative approaches such as Grave and Elhadad (2015).

In the fourth part, we will focus on several special topics. First, while most of the previous approaches to unsupervised parsing are unlexicalized, we will discuss the impact of partial and full lexicalization (e.g., the work by Pate and Johnson (2016); Han et al. (2017)). Second, we will discuss whether and how big training data could benefit unsupervised parsing (Han et al., 2017). Third, we will introduce recent attempts to induce syntactic parses from pretrained language models such as BERT (Rosa and Mareček, 2019; Wu et al., 2020). Fourth, we will cover unsupervised multilingual parsing, the task of performing unsupervised parsing jointly on multiple languages (e.g., the work by Berg-Kirkpatrick and Klein (2010); Han et al. (2019)). Fifth, we will introduce visually grounded unsupervised parsing, which tries to improve unsupervised parsing with the help from visual data (Shi et al., 2019). Finally, we will discuss latent tree models trained with feedback from downstream tasks, which are related to unsupervised parsing (Yogatama et al., 2016; Choi et al., 2018).

In the last part, we will summarize the tutorial and discuss potential future research directions of unsupervised parsing.

3 Outline

Part 1. Introduction [20 min]

- Problem definition
- Motivations and applications
- Evaluation
- Overview of approaches

Part 2. Generative Approaches [60 min]

- Overview
- Approaches based on generative grammars
  - Structure learning
  - Parameter learning
- Approaches based on language models

Coffee Break [30 min]

Part 3. Discriminative Approaches [40 min]

- Overview
- Autoencoders
- Variational autoencoders
- Other discriminative approaches

Part 4. Special Topics [50 min]

- Lexicalization
- Big training data
- BERT-based approaches
- Unsupervised multilingual parsing
- Visually grounded unsupervised parsing
- Latent tree models with downstream tasks

Part 5. Summary and Future Directions [10 min]

4 Prerequisites for the Attendees

Linguistics Familiarity with grammars and syntactic parsing.

Machine Learning Basic knowledge about generative vs. discriminative models, unsupervised learning algorithms (such as expectation-maximization), and deep learning.

5 Reading List

Klein and Manning (2004) – An influential generative approach to unsupervised dependency parsing that is the basis for many subsequent papers.

Jiang et al. (2016) – A neural extension of Klein and Manning (2004). One of the first modern neural approaches to unsupervised parsing.

Stolcke and Omohundro (1994) – One of the first structure learning approaches of context-free grammars for unsupervised constituency parsing.

Tu and Honavar (2012) – A parameter learning approach to unsupervised dependency parsing based on unambiguity regularization.

Cai et al. (2017) – An autoencoder approach to unsupervised dependency parsing.

Kim et al. (2019) – A variational autoencoder approach to unsupervised constituency parsing.

6 Presenters

Kewei Tu (ShanghaiTech University)
http://faculty.sist.shanghaitech.edu.cn/faculty/tukw/

Kewei Tu is an associate professor with the School of Information Science and Technology at ShanghaiTech University. His research lies in the areas of natural language processing, machine learning, and artificial intelligence in general, with a focus on the representation, learning and application of linguistic structures. He has over 50 publications in NLP and AI conferences and journals including ACL, EMNLP, AAAI, IJCAI, NeurIPS and ICCV. He served as an area chair for the syntax track at EMNLP 2020.

Yong Jiang (Alibaba DAMO Academy)
http://jiangyong.site

Yong Jiang is a researcher at Alibaba DAMO Academy, Alibaba Group. He received his Ph.D. degree from the joint program of ShanghaiTech University and University of Chinese Academy of Sciences. He was a visiting student at University of California, Berkeley in 2016. His research interest mainly focuses on machine learning and natural language processing, especially multilingual and unsupervised natural language processing. His research has been published in top-tier conferences and journals including ACL, EMNLP and AAAI.

Wenjuan Han (National University of Singapore)
http://hanwenjuan.com
Wenjuan Han is a research fellow at National University of Singapore. She received her Ph.D. degree from the joint program of ShanghaiTech University and University of Chinese Academy of Sciences. She was a visiting student at University of California, Los Angeles in 2019. Her research focuses on natural language understanding, especially unsupervised syntactic parsing. Her research has been published in top-tier *ACL conferences.

Yanpeng Zhao (University of Edinburgh)
https://zhaoyanpeng.github.io/
Yanpeng Zhao is a Ph.D. student in the Institute for Language, Cognition and Computation (ILCC) at the University of Edinburgh. His research interests lie in structured prediction and latent variable models with a focus on syntactic parsing and relation induction. His work was nominated for the best paper award at ACL 2018 and received an honorable mention for best paper award at EMNLP 2020.

References

J. K. Baker. 1979. Trainable grammars for speech recognition. In *Speech Communication Papers for the 97th Meeting of the Acoustical Society of America*.

Taylor Berg-Kirkpatrick, Alexandre Bouchard-Côté, John DeNero, and Dan Klein. 2010. Painless unsupervised learning with features. In *NAACL*.

Taylor Berg-Kirkpatrick and Dan Klein. 2010. Phylogenetic grammar induction. In *ACL*.

Jiong Cai, Yong Jiang, and Kewei Tu. 2017. CRF autoencoder for unsupervised dependency parsing. In *EMNLP*.

Jihun Choi, Kang Min Yoo, and Sang-goo Lee. 2018. Unsupervised learning of task-specific tree structures with tree-lstms. *AAAI*.

A. Clark. 2007. Learning deterministic context free grammars: The omphalos competition. *Machine Learning*, 66.

Shay B Cohen and Noah A Smith. 2010. Covariance in unsupervised learning of probabilistic grammars. *The Journal of Machine Learning Research*.

E. Mark Gold. 1967. Language identification in the limit. *Information and Control*, 10(5):447–474.

Edouard Grave and Noémie Elhadad. 2015. A convex and feature-rich discriminative approach to dependency grammar induction. In *ACL-IJCNLP*.

Wenjuan Han, Yong Jiang, and Kewei Tu. 2017. Dependency grammar induction with neural lexicalization and big training data. In *EMNLP*.

Wenjuan Han, Ge Wang, Yong Jiang, and Kewei Tu. 2019. Multilingual grammar induction with continuous language identification. In *EMNLP*.

Junxian He, Zhisong Zhang, Taylor Berg-Kirkpatrick, and Graham Neubig. 2019. Cross-lingual syntactic transfer through unsupervised adaptation of invertible projections. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3211–3223, Florence, Italy. Association for Computational Linguistics.

Zixia Jia, Youmi Ma, Jiong Cai, and Kewei Tu. 2020. Semi-supervised semantic dependency parsing using CRF autoencoders. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6795–6805, Online. Association for Computational Linguistics.

Yong Jiang, Wenjuan Han, and Kewei Tu. 2016. Unsupervised neural dependency parsing. In *EMNLP*.

Lifeng Jin, Finale Doshi-Velez, Timothy Miller, William Schuler, and Lane Schwartz. 2018. Depth-bounding is effective: Improvements and evaluation of unsupervised PCFG induction. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2721–2731, Brussels, Belgium. Association for Computational Linguistics.

Mark Johnson, Thomas L. Griffiths, and Sharon Goldwater. 2007. Bayesian inference for pcfgs via markov chain monte carlo. In *HLT-NAACL*, pages 139–146.

Yoon Kim, Alexander M Rush, Lei Yu, Adhiguna Kuncoro, Chris Dyer, and Gábor Melis. 2019. Unsupervised recurrent neural network grammars. In *NAACL*.

Dan Klein and Christopher D. Manning. 2004. Corpus-based induction of syntactic structure: Models of dependency and constituency. In *ACL*.

Jun Li, Yifan Cao, Jiong Cai, Yong Jiang, and Kewei Tu. 2020. An empirical comparison of unsupervised constituency parsing methods. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3278–3283, Online. Association for Computational Linguistics.

Hiroshi Noji, Yusuke Miyao, and Mark Johnson. 2016. Using left-corner parsing to encode universal structural constraints in grammar induction. In *EMNLP*.

John K Pate and Mark Johnson. 2016. Grammar induction from (lots of) words alone. In *COLING*.

Rudolf Rosa and David Mareček. 2019. Inducing syntactic trees from bert representations. In *BlackboxNLP*.

Yikang Shen, Zhouhan Lin, Chin wei Huang, and Aaron Courville. 2018. Neural language modeling by jointly learning syntax and lexicon. In *International Conference on Learning Representations*.

Yikang Shen, Shawn Tan, Alessandro Sordoni, and Aaron Courville. 2019. Ordered neurons: Integrating tree structures into recurrent neural networks. In *International Conference on Learning Representations*.

Haoyue Shi, Jiayuan Mao, Kevin Gimpel, and Karen Livescu. 2019. Visually grounded neural syntax acquisition. *arXiv preprint arXiv:1906.02890*.

Valentin I. Spitkovsky, Hiyan Alshawi, and Daniel Jurafsky. 2010a. From Baby Steps to Leapfrog: How "Less is More" in unsupervised dependency parsing. In *NAACL*.

Valentin I Spitkovsky, Hiyan Alshawi, Daniel Jurafsky, and Christopher D Manning. 2010b. Viterbi training improves unsupervised dependency parsing. In *CoNLL*.

Andreas Stolcke and Stephen Omohundro. 1994. Inducing probabilistic grammars by bayesian model merging. In *International Colloquium on Grammatical Inference*.

Kewei Tu and Vasant Honavar. 2012. Unambiguity regularization for unsupervised learning of probabilistic grammars. In *EMNLP-CoNLL*.

Kewei Tu, Maria Pavlovskaia, and Song-Chun Zhu. 2013. Unsupervised structure learning of stochastic and-or grammars. In *Advances in Neural Information Processing Systems 26*, pages 1322–1330.

Zhiyong Wu, Yun Chen, Ben Kao, and Qun Liu. 2020. Perturbed masking: Parameter-free probing for analyzing and interpreting BERT. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4166–4176, Online. Association for Computational Linguistics.

Dani Yogatama, Phil Blunsom, Chris Dyer, Edward Grefenstette, and Wang Ling. 2016. Learning to compose words into sentences with reinforcement learning. In *ICLR*.

Yanpeng Zhao and Ivan Titov. 2020. Visually grounded compound PCFGs. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4369–4379, Online. Association for Computational Linguistics.