The 56th Annual Meeting of the Association for Computational Linguistics

ACL 2018

Author Response

Title: Unsupervised Depth-bounded Grammar Induction Model for PCFG with Inside-sampling
Authors: Lifeng Jin, William Schuler, Finale Doshi-Velez, Timothy Miller and Lane Schwartz

Instructions

The author response period has begun. The reviews for your submission are displayed on this page. If you want to respond to the points raised in the reviews, you may do so in the boxes provided below.

We encourage you to explicitly respond to every point raised in **Weaknesses** and **Questions to Authors** sections in the reviews. Use the template provided in the response box as far as possible, but this is not obligatory.

Please note: *you are not obligated to respond to the reviews*.

For reference, you may see the review form that reviewers used to evaluate your submission. If you do not see some of the filled-in fields in the reviews below, it means that they were intended to be seen only by the committee. See the review form HERE.

Review #1

Appropriateness	Appropriate
Adhere to ACL 2018 Guidelines	Yes
Adhere to ACL Author Guidelines	Yes
Handling of Data / Resources	N/A
Handling of Human Participants	N/A

Summary and Contributions

Summary:

The paper presents a Bayesian model for inferring the constituency annotation that should be assigned to a corpus of natural language sentences in an unsupervised setting, i.e., given only the sentences in the corpus without POS annotation. The model is based on a Dirichlet-Multinomial model for the rules in a grammar, which is combine with a constraint on the maximum depth of a parse tree, under a psycholinguistically motivated notion of depth. In order to do inference for this model, the paper proposes a matrix based definition for the operations of a sampler, which allows for a GPU based implementation that can better exploit modern hardware. Included in the paper is a detailed discussion of the influence of different parameter settings on the performance of the overall system for inference on an English corpus. The best parameter settings are then used to evaluate the system on Chinese and German data. The results are compared to two other systems for unsupervised annotation.

Contribution 1:

The main contribution is the proposed model, which is similar, but not identical to the model used by [1]. The model is applied to unsupervised constituency parsing

with reasonable results.

Contribution 2:

The paper has a nice discussion of the optimal parameter settings for applying the model to English and generally does a much better than usual job of presenting a statistical exploration of the behavior of the unsupervised parsing system, which would be helpful for other authors working in this domain.

Contribution 3:

The paper explains how to implement a known sampler with matrix operations, which allows the use of efficient GPU operations, which might speed up the algorithm.

[1] https://arxiv.org/abs/1802.08545

Strengths

Strength argument 1:

I really liked the in-depth discussion of the parameter settings and the fact that there was a discussion of the variance of the results depending on parameters and sampler initializations. I think anyone working on unsupervised grammar induction problems knows that this problem exists for quite a few systems. The paper also discusses how much this problem can be overcome by inspecting metrics such as the log-likelihood of the data over different initializations, which again should be helpful for others.

Strength argument 2:

The paper attempts to use a well known and, presumably, widely applicable constraint on human language processing to guide unsupervised grammar induction. This seems like the right approach to improve grammar induction systems and it should be easy to extend it with further constraints.

Strength argument 3:

I think that most of the paper, aside from section 3, is very well written and understandable.

Weaknesses

Weakness argument 1:

I think the model is sufficiently straightforward to not be very new for any reader interested in Bayesian methods and grammar induction. The implementation is also primarily based on existing techniques. I think that I have learned less new information from this paper than I would expect from a paper that is a certain accept for ACL.

Weakness argument 2:

It is not clear how well the parameter settings found on English generalize to other languages. While it is normal that systems perform worse on languages other than English (which seems structurally simpler than most resource rich languages), it is problematic that F-Score for the system is only better than in the comparison systems on English data, while the comparison systems perform better for Chinese and German (for German substantially so). This suggests that the parameters found for English might have been overfitted to English. The fact that the system presented in the paper always beats the other systems in recall is most likely due to the fact that (as far as the paper suggests), the system always proposes full binary branching trees, while the competing systems build non-binary trees and proposing more constituents will always improve recall. I would have liked an exploration of best parameter settings on other languages, e.g., by applying the model to the copora in the universal dependencies (UD) treebanks and seeing

which settings lead to the greatest agreement between the generated trees and the dependency structures annotated in the UD.

Weakness argument 3:

Minor: I think section 3 was harder to read for me than necessary because of the use of the non-standard (in NLP) approach of writing grammars in matrix form. My understanding could have been greatly aided by a few short sentences explaining the meaning of each matrix used. Also, at some points I suspect that there would have been a more straightforward way to write definitions, e.g., I think that (1) could have been written as:

G_{abc} = P(c -> a,b)

However, I think that the main points of the paper still come through clearly enough.

Weakness argument 4:

Minor: The authors could have cited additional relevant literature on using matrix operations to enable work with CFG based formalisms on GPUs [1,2] and on unsupervised constituency parsing without gold POS tags [3].

[1] D. Hall, T. Berg-Kirkpatrick, J. Canny, and D. Klein. 2014. Sparser, Better, Faster GPU Parsing; Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics

[2] J. Canny, D. Hall, and D. Klein. 2013. A multi-Teraflop Constituency Parser using GPUs; Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing

[3] Christian Hänig. 2010. Improvements in unsupervised co-occurrence based parsing; Proceedings of the Fourteenth Conference on Computational Natural Language Learning

Questions to Authors (Optional)

I do not have any specific questions which I need the authors to answer for my evaluation of this paper.

NLP Tasks / Applications	N/A
Methods / Algorithms	Marginal contribution
Theoretical / Algorithmic Results	N/A
Empirical Results	Marginal contribution
Data / Resources	N/A
Software / Systems	N/A
Evaluation Methods / Metrics	N/A
Other Contributions	Moderate contribution
Originality (1-5)	3
Soundness/Correctness (1-5)	2
Substance (1-5)	4
Replicability (1-5)	4
Meaningful Comparison (1-5)	4
Readability (1-5)	3
Overall Score (1-6)	3

Additional Comments (Optional)

I like the paper and think it does a lot of things right. If either of the main contributions where more substantial, then I think the paper would be a clear accept recommendation. This could mean expanding the discussion of the

behavior of the algorithms or experimenting with models that include additional constraints.

How important is the implementation on graphics cards for the experiment to be feasible?

I would have liked a more in-depth discussion of the specific depth constraint used and why the authors think it is generally helpful. I am willing to believe them, I just would have liked an argument on why it should generalize from English, which is right branching, to Chinese, which is left-branching, or German, which has a freer word-order. If the paper presented experiments with other definitions of depth for the other languages and found them to perform worse, then I would have more faith that this is indeed the right constraint.

Would it be possible to use the posterior probability of constituents to discard constituents? This would make it possibel to obtain a parse that is not completely binary. Also, would it be possible to combine the presented system with existing systems for a better or more stable result? E.g., by initializing the presented system with the results from the other ones?

I would have used a shuffling based test [1] for establishing significance in system comparisons, because I never feel certain that the required pre-conditions are met for a paired t-test to apply, but that may just be personal preference.

[1] A. Yeh. 2000. More accurate tests for the statistical significance of result differences. Proceedings of COLING

Review #2

Appropriateness	Appropriate
Adhere to ACL 2018 Guidelines	Yes
Adhere to ACL Author Guidelines	Yes
Handling of Data / Resources	N/A
Handling of Human Participants	N/A

Summary and Contributions

Summary: This paper extends a previous depth-bounded grammar induction method by running Gibbs sampling with PCFGs. In the generative model, the PCFG is converted into a depth-bounded PCFG before being used to generate parse trees and sentences. Gibbs sampling is used to alternately sample the (unbounded) grammar and the parse trees of the training sentences. Experiments show that the proposed approach achieves competitive results in unsupervised constituency parsing.

Contribution 1: A novel method that integrates depth-bounding into the sampling-based approach to PCFG induction.

Strengths

Strength argument 1: The integration of depth-bounding and sampling-based PCFG learning is novel and interesting.

Strength argument 2: Good empirical results on the very difficult task of PCFG induction.

Strength argument 3: The empirical analysis is comprehensive and informative.

Weaknesses

Weakness argument 1: I find some part of section 3 hard to follow. For example, in Eq.2, G_{1,d} and G_{2,d} are undefined and may be mistaken for elements in G.

Weakness argument 2: In section 4.2, it is said that beta=0.5 is over-regularizing the model. This may be incorrect because the Dirichlet prior becomes weaker when beta is closer to 1 (it becomes uniform when beta is 1). Besides, larger values of beta (>1) should be tested to see the effect of smoothing.

Questions to Authors (Optional)

Question 1: During Gibbs sampling, you sample an unbounded grammar from the parse trees that are sampled from a bounded grammar. The current procedure of this sampling step regards the parse trees as if they are sampled from an unbounded grammar. Can you prove or explain why this is correct, considering that the probability of a parse tree is no longer the product of rule probabilities from the unbounded grammar.

Question 2: What is your training data for experiments in section 4.1 and 4.2? The paper only states that the first half of WSJ20 is used as the development set.

NLP Tasks / Applications	N/A
Methods / Algorithms	Moderate contribution
Theoretical / Algorithmic Results	N/A
Empirical Results	N/A
Data / Resources	N/A
Software / Systems	N/A
Evaluation Methods / Metrics	N/A
Other Contributions	N/A
Originality (1-5)	3
Soundness/Correctness (1-5)	4
Substance (1-5)	4
Replicability (1-5)	4
Meaningful Comparison (1-5)	4
Readability (1-5)	3
Overall Score (1-6)	4

Additional Comments (Optional)

Figure 5: could you also show the depth distribution of the gold parses?

Line 467: following -> follow

Line 699: compute -> computing

Review #3

Appropriateness	Appropriate
Adhere to ACL 2018 Guidelines	Yes
Adhere to ACL Author Guidelines	Yes
Handling of Data / Resources	N/A
Handling of Human Participants	N/A

Summary and Contributions

Summary: This paper presents a method of depth-bounded grammar induction from raw text based on Bayesian modeling and Gibbs sampling. The authors successfully achieves competitive results on difficult unsupervised grammar induction task.

Contribution 1: An extensive and comprehensive study for the model behavior and hyperparameter settings of depth-bounded grammar induction.

Strengths

Strength argument 1: I think the main contribution of this paper is an extensive and comprehensive study for the model behavior and hyperparameter settings of depth-bounded grammar induction. Especially, showing the variance of the performance, the sensitivity of hyperparameters, etc. reveal the difficulty of the unsupervised grammar induction.

Strength argument 2: The proposed method achieves competitive or superior performance compared with conventional ones on unsupervised grammar induction.

Strength argument 3:

Strength argument 4:

Strength argument 5:

Weaknesses

Weakness argument 1: The originarity of the proposed method is weak. I understand the sampling procedure is somewhat novel, but basically it is based on the exsiting techniques from Bayesian unsupervised grammar induction.

Weakness argument 2: The model definition (section 3) may be hard to follow for many readers. The notations and explanation could be improved.

Weakness argument 3:

Weakness argument 4:

Weakness argument 5:

NLP Tasks / Applications	N/A
Methods / Algorithms	Marginal contribution
Theoretical / Algorithmic Results	N/A
Empirical Results	N/A
Data / Resources	N/A
Software / Systems	N/A
Evaluation Methods / Metrics	N/A
Other Contributions	N/A
Originality (1-5)	2
Soundness/Correctness (1-5)	3
Substance (1-5)	2
Replicability (1-5)	3
Meaningful Comparison (1-5)	2
Readability (1-5)	3
Overall Score (1-6)	2

Author Response

ATTENTION: this time, we plan to do some analytics on anonymized reviews and rebuttal statements, upon the agreement of the reviewers and authors, with the purpose of improving the quality of reviews. The data will be compiled into a unique corpus, which we potentially envisage as a great resource for NLP, e.g. for sentiment analysis and argumentation mining, and made available to the community properly anonymized at earliest in 2 years. We hope to provide data on "how to review" to younger researchers, and improve transparency of the reviewing process in ACL in general.

By default, you agree that your anonymised rebuttal statement can be freely used for research purposes and published under an appropriate open-source license within at earliest 2 years from the acceptance deadline.

Select "No" if you would like to opt out of the data collection: Yes

Review Quality Survey

The quality of the reviews significantly influences the quality of the conference. To evaluate the quality of each review and reviewer, we invite the authors to rate the reviews they have received. Note that the survey results will only be presented to Programme Chairs and the corresponding Area Chairs, and will *NOT* be disclosed to the reviewers. Use the guidelines here to answer the questions for each review:

Quality of the Review

Do the reviews address the main strengths/weaknesses of the paper? Does the reviewer have a good understanding of the paper? Do the reviews include some insightful comments?

1. Nonsense: the reviews are mostly confusing and hard to understand
2. Below average: the reviews are largely about minor points, and there are significant misunderstandings of the paper
3. Mediocre: the reviews give some valuable comments, but also ignore/degrade some important strengths/potentials of this paper
4. Above average: the reviews point out some main strengths and weaknesses of the paper and provide good justifications
5. Insightful: the reviews not only grasp the main strengths/weaknesses of the paper, but also give convincing analyses and insightful comments

Helpfulness of the Review

How helpful are the reviews? Can you use the reviews to improve the paper? Do the reviews give some insightful comments that can be helpful to your future research?

1. Poor: the reviews make not much sense and are largely useless
2. Somewhat helpful: the reviews can help me to do some minor changes, but no major improvements can be made
3. Helpful: the reviews give an objective and fairly comprehensive evaluation of the paper, and can help me marginally improve the quality of the paper
4. Very helpful: with the help of these reviews, some major weaknesses of this paper can be strengthened and some strengths can be reinforced, thus the quality of the paper can be significantly improved
5. Out of my expectation: the reviews not only give important comments on how to improve this paper, but also give some insightful analyses and advice that can be helpful to my ongoing research and future works

---

Submit Response to Reviewers

Use the following boxes to enter your response to the reviews. Please limit the total amount of words in your comments to 1000 words (longer responses will not be accepted by the system).

Response to Review #1:

Reply to weakness argument 1:

Reply to weakness argument 2:

Reply to weakness argument 3:

Reply to weakness argument 4:

Reply to weakness argument 5:

Reply to question 1:

Reply to question 2:

Reply to question 3:

Quality of Review #1: 1

Helpfulness of Review #1: 1

Response to Review #2:

Reply to weakness argument 1:

Reply to weakness argument 2:

Reply to weakness argument 3:

Reply to weakness argument 4:

Reply to weakness argument 5:

Reply to question 1:

Reply to question 2:

Reply to question 3:

Quality of Review #2: 1

Helpfulness of Review #2: 1

Response to Review #3:

Reply to weakness argument 1:

Reply to weakness argument 2:

Reply to weakness argument 3:

Reply to weakness argument 4:

Reply to weakness argument 5:

Reply to question 1:

Reply to question 2:

Reply to question 3:

Quality of Review #3: 1

Helpfulness of Review #3: 1

Response to Chairs

Use this textbox to contact the area chairs directly only when there are serious issues regarding the reviews. Such issues can include reviewers who grossly misunderstood the submission, or have made unfair comparisons or requests in their reviews. Most submissions should not need to use this facility.

Submit
