Team member:
Jeremy Choo(1602380), Siyuan Liu(1589879), Youwei Chen (1591895)

Title:
A Replication Study to Evaluate the Performance of Unsupervised Chunking as Syntactic Structure Induction with a Knowledge-Transfer Approach

1 Introduction / Background

Nowadays the high-performance model that is designed for understanding the linguistic structure of language, requires lots of massive data to be labelled. Where Unsupervised learning methods provide a way that does not need human labelling. We chose the paper *Unsupervised Chunking as Syntactic Structure Induction with a Knowledge-Transfer Approach*, which was published in 2021. "The paper proposes a knowledge-transfer approach that heuristically induces chunk labels from state-of-the-art unsupervised parsing models"[1]. Unsupervised learning methods can potentially discover linguistic structures for low-resource languages. As the paper proposed, "Further constructing new treebanks for low-resource languages is expensive"[1]. And the Unsupervised learning "can also be a first pass in annotating large treebanks for them".

The paper proposes a knowledge-transfer approach to unsupervised chunking by hierarchical recurrent neural networks (HRNN). Firstly, the paper's team uses a simple heuristic to achieve a noisy and reasonable chunking performance[1]. Then, apply HRNN learning to smooth out such noise and yield more meaningful chunks[1]. For the result, the paper evaluates on teacher model Compound PCFG, with the selection from student models: HRNN Only, BERT+1-layer RNN, BERT+2-layer RNN, BERT+HRNN (hard) and BERT+HRNN on CoNLL-2000 (English), CoNLL-2003 (German), English Web Treebank datasets. And as a comparison study, The paper also analyzed the chunking heuristics: 1-word & 2-word chunks, Maximal right-branching, Maximal left-branching.

Our replication will focus on evaluating the performance of the proposed method and conclude whether the results obtained are valid or not. In our project, we might not use the same computer configurations as given in the original paper. We are planning to use Compute Canada servers to run the experiment.

2 Related Work

The idea in the original paper is originally from Li et al. (2019), who proposed a method to transfer knowledge between different unsupervised parsers to achieve better performance[1]. As unsupervised learning parsers are gaining interest, increasingly more unsupervised parsers are proposed by researchers. The parser used to induce chunk labels is the Compound PCFG, which is an excellent parser proposed by Kim et al. (2019). Then, a pre-trained language model BERT (Kenton et al., 2019) is first applied to let the model understand a general view of the global context. After that, a hierarchical RNN has trained to denoise these chunk labels[1]. In the end, the result is derived and used to compare the performance with other pre-trained language models.

3 Purpose / Focus of this study

The purpose of this study is to evaluate the effectiveness of the proposed method by replicating the experiment following the algorithm stated in the paper. To conclude, we will try to apply the same dataset (CONLL - 2000) and check if our result matches the one provided. Furthermore, we can also carry out the evaluation using the extra datasets given in Kaggle, for example, COLL 2002, 2007, etc. to enhance the statement in our conclusion.

We decided to replicate this paper for various reasons: (a) the proposed approach is related to state-of-the-art unsupervised parsing models, (b) the result of the proposed model provided in the paper is fascinating and it dominates the other unsupervised approaches such as LM Chunker and Compound PCFG Chunker, and (c) the steps to implement the algorithm as well as the datasets used for evaluation is mentioned clearly so that we can follow the instructions to produce the corresponding result.

References

[1] Anup Anand Deshmukh, Qianqiu Zhang, Ming Li, Jimmy Lin, and Lili Mou, David R. Unsupervised Chunking as Syntactic Structure Induction with a Knowledge-Transfer Approach. In: EMNLP Conference on Short Papers; 2021 -[2]Yoon Kim, Chris Dyer, and Alexander M Rush. 2019b. Compound probabilistic context-free grammars for grammar induction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2369–2385.

[3]Bowen Li, Lili Mou, and Frank Keller. 2019. An imitation learning approach to unsupervised parsing. In *Proceedings of the 57th Annual Meeting of the Asso- ciation for Computational Linguistics*, pages 3485– 3492.

[4]Jacob Kenton, Devlin Ming-Wei Chang, and Lee Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186.