How implicit negative evidence improve probabilistic grammar induction

Olgierd Unold
Department of Computer Engineering
Wrocław University of Science and Technology
27 Wybrzeże Wyspiańskiego,
50-370 Wrocław, Poland
olgierd.unold@pwr.edu.pl

Grzegorz Rorbach
Wrocław Univ Sci & Technol
rorbach.g@gmail.com

Extended abstract of the 14th International Conference on Grammatical Inference *ICGI'18*, September 5–7, 2018, Wrocław, Poland.

Abstract

A modified inside-outside algorithm to estimate probabilistic parameters over implicit positive and negative evidence was proposed. We have demonstrated that a contrastive estimation based method significantly outperforms a standard inside-outside algorithm in terms of Specificity, without any loss of Sensitivity.

Author Keywords

Probabilistic context-free grammar; inside-outside algorithm; contrastive estimation; implicit negative evidence.

Probabilistic context–free grammars (PCFGs) are applied successfully to biological sequences modeling from the early 90s [6]. Some, but not all, attempts are listed in [5].

Given a task, for example, biological sequences (RNA, DNA, proteins) to be modeled, the question then arises how to induce probabilistic grammar from unannotated data (so-called unsupervised learning). The task of learning PCFGs from data consists of two subproblems: determining a discrete structure of the target grammar and estimating probabilistic parameters in the grammar. Given the fixed topology of the grammar, the inside-outside algorithm [1, 4] is the standard method used to estimate the probabilistic parameters of a PCFG. This procedure is an expectation-maximization (EM [2]) method for obtaining maximum likelihood of the grammar's parameters. However, it requires

the grammar to be in Chomsky normal form, and it accepts only positive examples in the learning data. Note that in 1969 Horning proved [3] that for effective PCFG induction no negative evidence are obligatory. Using only positive data in learning PCFG has one significant disadvantage of inducing grammars which are not specific for a given language, i.e. not able to distinguish negative examples.

To overcome that problem we propose a modified inside-outside algorithm to estimate probabilistic parameters over implicit positive and negative evidence in learning data. We employ the concept of Contrastive Estimation (CE) [7], a method that provides a way to use implicit negative evidence. The idea of CE is to generate a given positive sentence s, a large neighborhood $N(s)$ of ungrammatical sentences as negative evidence, by perturbing s with certain operations. Note that EM can be seen as a specific case of CE, where the neighborhood $N(s)$ is the entire set of learned sentences. It should be noted that instead of the PCFG, CE uses weighted context-free grammar.

In a proposed approach, called IOCE (inside-outside CE), we introduce the CE factor of the rule:

$$C_{\varphi,CE}(A \longrightarrow \alpha) = \frac{C_{\varphi}(A \longrightarrow \alpha)}{C_{\varphi}(A \longrightarrow \alpha) + C_{\varphi,ng}(A \longrightarrow \alpha)}$$

where:
$C_{\varphi}(A \longrightarrow \alpha)$–the estimated count of the number of times that a particular rule is used in positive evidence,
$C_{\varphi,ng}(A \longrightarrow \alpha)$–the estimated count of the number of times that a a particular rule is used in a neighborhood.
The probability of the rule is calculated as follows:

$$\varphi^{'}(A \longrightarrow \alpha) = \frac{C_{\varphi}(A \longrightarrow \alpha)}{\sum_{\beta} C_{\varphi}(A \longrightarrow \beta)} \cdot C_{\varphi,CE}(A \longrightarrow \alpha)$$

We tested two different neighborhoods. In the first one (IOCEa, all negatives), the neighborhood for each positive sentence is created by choosing a determined number of sentences from the set of all available negative sentences. In the second one (IOCEs, negatives of the same length), we choose the determined number of sentences from the set of all negatives of the same length (plus/minus 2 symbols) as the positive sentence.

We compared our approach with two different neighborhoods to a standard IO over 12 test artificial languages: regular (tomita1 - tomita7) and context-free (ab, anbn, pal2, bra1, bra3). Comparative analysis of the three measures of Sensitivity, Specificity, and F1 with standard deviation) is summarized in Table 1. All experiments were done using pyGCS library [8].

Regardless of the neighborhood, the Specificity of the grammar induced by the novel method is twice as high (0.80 vs 0.40) compared with the standard IO.

We showed that PCFG trained using implicit negative evidence can drastically outperform IO-trained grammar in terms of Specificity and F1 score.

Acknowledgements

The research was supported by National Science Centre Poland (NCN), project registration no. 2016/21/B/ST6/02158.

REFERENCES

1. James K Baker. 1979. Trainable grammars for speech recognition. *The Journal of the Acoustical Society of America* 65, S1 (1979), S132–S132.

2. Arthur P Dempster, Nan M Laird, and Donald B Rubin. 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the royal statistical society.*

Algorithm	Sensitivity	Specificity	F1
IO	0.87 ± 0.01	0.40 ± 0.01	0.64 ± 0.01
IOCEa	0.91 ± 0.00	**0.80** ± 0.02	**0.84** ± 0.02
IOCEs	0.91 ± 0.01	**0.79** ± 0.03	**0.83** ± 0.01

Table 1: Performance of compared methods in terms of averaged Sensitivity, Specificity, and F1 with standard deviation.

Series B (methodological) (1977), 1–38.

3. James Jay Horning. 1969. *A study of grammatical inference.* Technical Report. Stanford Univ Calif Dept of Computer Science.
4. Karim Lari and Steve J Young. 1990. The estimation of stochastic context-free grammars using the inside-outside algorithm. *Computer speech & language* 4, 1 (1990), 35–56.
5. Hyun-Seok Park, Bulgan Galbadrakh, and Young-Mi Kim. 2011. Recent progresses in the linguistic modeling of biological sequences based on formal language theory. *Genomics & Informatics* 9, 1 (2011), 5–11.
6. Yasubumi Sakakibara, Michael Brown, Richard Hughey, I Saira Mian, Kimmen Sjölander, Rebecca C Underwood, and David Haussler. 1994. Stochastic context-free grammers for tRNA modeling. *Nucleic acids research* 22, 23 (1994), 5112–5120.
7. Noah A Smith and Jason Eisner. 2005. Contrastive estimation: Training log-linear models on unlabeled data. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*. Association for Computational Linguistics, 354–362.
8. Olgierd Unold, Grzegorz Rorbach, Mateusz Fislak, Michal Czarnecki, and Daniel Cieszko. 2018. pyGCS. `https://github.com/ounold/pyGCS`. (2018).