

# A Goodness Measure for Phrase Structure Learning via Compression with the MDL Principle

## 1 Introduction

Grammar induction from a naturally-occurring text corpus is in the general domain of inferring a theory (or model) to account for a given set of observed data. Practical techniques for grammar induction have many important applications to a wide range of natural language (NL) and speech processing tasks. Researchers have devoted tremendous effort to the research in the past decades.

Inferring a probabilistic grammar from NL data, as revealed in most foregoing practice, involves in general two essential sub-tasks, one to infer a set of phrase structure rules (or productions), another to estimate a correspondent set of probabilistic parameters (i.e., production probabilities) that optimise the fit of the grammar to the given data. The inference can be viewed as a search process for the best grammar allowable in a predefined *grammar* or *hypothesis space*. If a set of permissible rules or rule formats (e.g., *Chomsky normal form* (CNF)) are given, it is popular to apply the *Baum-Welch* (or *forward-backward*) *algorithm* [1] and its extension the *inside-outside algorithm* [25], an incarnations of the expectation-maximisation (EM) algorithm, to estimate the probabilistic parameters for regular grammars (e.g., hidden Markov models (HMMs)) and SCFGs, respectively. There are also many other sophisticated algorithms to facilitate the searching, e.g., *genetic algorithm* [19] and *stimulated annealing* algorithm [22].

However, no matter how sophisticated is the search method in use, the goodness criterion to guide the searching remains a critical issue, without which a search algorithm wouldn't know which grammar is better. In this paper we report our ongoing research on this specific task to develop a suitable criterion for unsupervised phrase structure learning from natural language corpora (e.g., the Brown [17, 18] and PTB corpus [28]). It is based on classic and algorithmic information theory [35, 12, 38, 24, 8, 26] and on the *Minimum Description/Message Length* (MDL, MML) principle [32, 33, 40, 41].

The rest of the paper is organised as the following. Section 2 reviews previous research on language learning, focusing on the goodness criterions. Section 3 discusses phrase learning via compression. Section 4 formulates description length calculation in bits, adhering to Rissanen's [32, 34] insight on the objectiveness of code length. Section 5 develops an information-theoretical measure for phrase learning, the *description length gain* or *compression effect* as an indicator to how good a word sequence is as a phrase, and a best-first learning algorithm with this measure. The last two sections report experiments and conclude the paper with a brief discussion on future work.

## 2 Learning Phrase Structure

Given no prior knowledge but a corpus as a sequence of linguistic symbols (e.g., words and characters) as data, unsupervised phrase learning needs to determine which segments of the sequence are phrases or phrase-like structures. This learning has to be carried out through guessing, guided by some reliable information-theoretical criterion on how good a fragment of the sequence is as a phrase. The guessing needs to consider all n-gram items in the corpus. Although n-grams of arbitrary lengths in a large-scale corpus are known to be huge in number, the *Virtual Corpus* (VC) [23], based on *suffix array* data structure [27, 29] and a bucket-radix sort, can be employed as a fairly efficient approach to handling them, including counting and, more importantly, storing and retrieval. It is reported that the VC system can

prepare the entire PTB WSJ part-of-speech (POS) tag corpus, of 1.3M tags, for outputting all n-grams within about 1.5 minutes on a Sun SPARC station 4. It is an ideal operational basis for inducing phrases from n-grams.

However, a more crucial issue in inducing phrases is a sound criterion to guide the guessing. Many criteria have been explored in previous research. For example, Cook *et al.* [11] explore a hill-climbing search for a grammar of a smaller weighted sum of grammar *complexity* and the *discrepancy* between grammar and corpus; Lari and Young [25], Carroll and Charniak [4, 5] induce various types of probabilistic context-free grammar (PCFG) with *inside-outside* algorithm [1] for probabilistic parameter re-estimation; Brill *et al.* [2] derive phrase structures from tagged corpus with *generalised mutual information*; and Brill and Marcus [3] attempt to induce binary branching phrases with distribution analysis using the information-theoretical measure *divergence*, derived from *relative entropy*; among many others. They have achieved significant progress, and also encountered some problems that seem unable to be resolved in their approaches, e.g., Brill *et al.* have to predefine *distitutes* as brute force to eliminate some unlikely phrases, e.g., [Noun Prep] and [Prep DT], that are favoured by their criteria. de Marcken [13] has an in-depth discussion on this kind of issues involved in pure distribution analysis and on the disadvantages of the inside-outside algorithm for grammar induction.

Recent work in language modelling as Stolcke [39] and Chen [9, 10] are in the theoretical framework of Bayesian modelling. Basically, Stolcke's work follows Cook *et al.*'s [11] paradigm of searching by hill-climbing, but guided by maximum likelihood instead. Chen follows Solomonoff's [36, 37, 38] thoughts of inductive learning, and uses the universal prior probability  $p(G) = 2^{-l(G)}$  for grammar induction. Their learning systems are reported to work well on small to medium size artificial corpora, in terms of the theoretical measures like entropy, perplexity or likelihood. However, more concrete evaluations are expected to be based on the performance of learning from a large-scale naturally-occurring corpus like the Brown or PTB corpus.

### 3 Learning via Compression with the MDL principle

The idea of learning via compression has been practised by researchers for a long time. A notable early discussion on the dual relationship between learning (i.e., detecting regularities in data) and compression is in Solomonoff's pioneering work [37, 38]. Some earlier related thoughts may be traced back to, for instance, Zipf's

Compression, or a procedure equivalent to compression has been widely exploited for unsupervised language learning at different linguistic levels, e.g., Oliver [30], Wolff [42, 43], Ellison [16], Cartwright and Brent [6, 7], Stolcke [39], Chen [9, 10] and de Marcken [14, 15], among many others. Recently, the MDL principle [32, 33] is extensively applied as a guidance to language learning, e.g., as in [39] [14, 15] mentioned. The central idea in MDL is that a model  $M$  (e.g., a grammar  $G$ ) for a set of data  $X$  (e.g., a corpus  $C$ ) with the shortest description length of both the model and the data (given the model), i.e.,  $|M| + |X_{\text{given } M}|$  in bits, is the best model.

The difference among various learning-via-compression approaches lies in how the compression is carried out. It is unnecessary to really have a compression procedure but a calculation of the probability of a model,  $p(M)$ , in terms of its structure (e.g., grammar rules) and parameters. For example, Stolcke's [39] and Chen's [9, 10] learning procedures are both to search for a model  $M$  to maximise the *posterior probability*  $p(M|X)$ , using the Bayes' rule

$$M_{max} = \arg \max_M p(M|X) = \arg \max_M \frac{p(X|M)p(M)}{p(X)} = \arg \max_M p(X|M)p(M) \quad (1)$$

where  $p(X)$  is a constant. Stolcke and Chen both use the prior  $p(M) = c^{-l(M)}$  to calculate  $p(M)$ , where  $l(M)$  is the description length of  $M$  encoded by a coding scheme chosen, and  $c = 2$  if  $l(M)$  is in bits. Different *ad hoc* coding schemes are used by different researchers to estimate  $l(M)$ .

Theoretically, searching for the most likely model for a given data set in a Bayesian modelling framework is equivalent to searching for a model with minimum description length, since using the negative

logarithm, (1) above becomes

$$M_{max} = \arg \min_M -\log p(X|M) - \log p(M) = \arg \min_M |X_{\text{given } M}| + |M| \quad (2)$$

It straightforwardly follows from information theory [35, 12]. A Bayesian interpretation of MDL principle can also be found in [31, 26].

There is another option that strictly follows the algorithmic information theory and the MDL principle to calculate bits, instead of probabilities: the model  $M$  and the data  $X$  (generated by  $M$ ) can be encoded as one, instead of as two separate parts, using a universal coding. The universal coding is assumed, in theory, to be a coding scheme that is as good as any other scheme in coding a random sequence. For the purpose of language learning, a coding scheme such as Shannon-Fano (or Huffman) code [35, 20, 12] or arithmetic coding [32, 33] would suffice for the purpose of description length calculation. Consequently, a general criterion for selecting phrase candidates among n-grams within a given corpus  $C$  is established as this: an n-gram is a better candidate if putting it as a phrase into  $G$  can lead to a shorter  $|G + C_{\text{given } G}|$ . That is, only n-grams with a good compression effect are candidates. Accordingly, phrase learning is to search for a subset  $G$  of this candidate set such that  $|G + C_{\text{given } G}|$  is minimal.

## 4 Calculating Description Length

The application of MDL is independent of encoding scheme [33, 31]. To resolve the critical issue of calculating the description length  $|G| + |C_{\text{given } G}|$ , what we need is an ideal encoding scheme for calculation, rather than a real compression program implemented.

Once an ideal scheme that can reach the best compression on a random sequence is assumed, the description length  $DL(X)$  on a given finite data set  $X = x_1 x_2 \cdots x_n$ , with a vocabulary  $V$ , can be calculated as below in terms of the *empirical entropy*  $\hat{H}(X)$ , which is the minimum expected average codeword length per symbol in  $X$ .

$$DL(X) = n\hat{H}(X) = -n \sum_{x \in V} \hat{p}(x) \log \hat{p}(x) = - \sum_{x \in V} c(x) \log [c(x)/n] \quad (3)$$

where  $c(x)$  is  $x$ 's count in  $X$ , and  $\hat{p}(x) = c(x)/n$ . The relation between this description length and the perplexity  $PP(X) = \hat{p}(x_1 x_2 \cdots x_n)^{-1/n}$  under a uni-gram model is straightforward as below.

$$\begin{aligned} \log PP(X) &= -\frac{1}{n} \log \hat{p}(x_1 x_2 \cdots x_n) = -\frac{1}{n} \log [\hat{p}(x_1) \hat{p}(x_2) \cdots \hat{p}(x_n)] \\ &= -\frac{1}{n} \sum_{i=1}^n \log \hat{p}(x_i) = -\frac{1}{n} \sum_{x \in V} c(x) \log \hat{p}(x) = \frac{1}{n} DL(X) \end{aligned} \quad (4)$$

Perplexity is an indication to the quality of a language model: a lower perplexity indicates a better model [21]. The description length in (3) and its average (i.e., the empirical entropy) in (4) play the same role.

## 5 Learning Algorithm

Following the DL calculation above, the description length gain of selecting each instance of the substring  $x_i x_{i+1} \cdots x_j$  (denoted as  $x_{i..j}$  for simplicity) ( $i < j$ ) in a given corpus  $X$  as a phrase is defined as

$$DLG(x_{i..j} \in X) = [DL(X) - DL(X[r \rightarrow x_{i..j}] \oplus x_{i..j})] / c(x_{i..j}) \quad (5)$$

where  $X[r \rightarrow x_{i..j}]$  represents a resultant corpus from the operation of replacing all instances of  $x_{i..j}$  with  $r$  throughout  $X$ , and  $\oplus$  denotes a string concatenation operation with a delimiter<sup>1</sup> inserted in between its two operands, i.e., the current corpus  $X$  and the newly learned phrase  $x_{i..j}$ .

<sup>1</sup>For the sake of simplicity, the sign  $\oplus$  is also used to denote a delimiter, in the context of no confusion caused.

Accordingly, a best-first learning algorithm using this goodness measure is put forward as below. It reaches a lower DL when more phrases are acquired.

1. Input:  $X_0$  (with a vocabulary  $V_0$ ); set  $k = 0$ .
2. Examine all substrings  $x_i x_{i+1} \dots x_j$  (denoted as  $x_{i..j}$  for simplicity) ( $i < j$ ) in  $X_k$ ,
  - (a) If no more  $x_{i..j}$  that  $DL(X_k[r \rightarrow x_{i..j}] \oplus x_{i..j}) < DL(X_k)$ , output phrases and exit;
  - (b) Else  $r^k = \arg \max_{x_{i..j} \in X_k} DLG(X_k[r \rightarrow x_{i..j}])$ .
3.  $X_{k+1} = X_k[r^k \rightarrow x_{i..j}] \oplus x_{i..j}$ ,  $V_{k+1} = V_k \cup \{r^k\}$ ,  $k = k + 1$ , goto 2.

It is worth noting that only a learned phrase, but not its index  $r^k$ , is concatenated to the updated corpus  $X_k[r^k \rightarrow x_{i..j}]$ . This eliminates redundant  $r^k$ 's in grammar. The phrase that a  $r^k$  represents immediately follows the  $k$ .th  $\oplus$  in the corpus.

One may see that this learning algorithm may not reach the shortest description length, since it is a best-first strategy that stops at a local minima. To alleviate this problem, an optimisation process is necessary. We are developing a learning algorithm to derive an optimal bracketing for a sentence with regard to DLG criterion<sup>2</sup>. Beyond this, It is highly possible that the criterion can be integrated into a GA or other sophisticated searching algorithms. Our goal here is to test the effectiveness of the DLG criterion for phrase learning. Also, the above learning need to observe some natural constraints (but not brute-force ones like *distitutes* [2, 3]) on phrases e.g., a phrase does not cross a sentence boundary.

## 6 Experiments

A number of preliminary experiments on unsupervised phrase and lexical learning have been conducted on parts of the PTB corpus with the above. The experimental outcomes show that the learning approach with the DLG measure gives promising results. It illustrates a very good capacity to capture the regularities in the input data.

## 7 Concluding remarks

We have developed a information-theoretical criterion, the *description length gain*, for unsupervised learning of phrase structures from natural language corpus within the MDL learning paradigm. The preliminary learning experiments show promising results.

This learning approach has two distinct features. First, it counts the description length in bits, instead of calculating probabilities then estimating compression effect in terms of the probability of the data given the model. Counting bits and calculating probability are theoretically equivalent, but operationally different in practice. This leads to another distinct feature of our approach: the corpus as input and the grammar as learning result are encoded together as one by the same coding scheme, instead of as two separate parts encoded by two different schemes. A universal compression is assumed for calculating description length. In practice, the Shannon-Fano coding can suffice for phrase learning and text compression. In this sense, our learning approach is theoretically and operationally elegant, in addition to its effectiveness.

Although experiments show encouraging results, in order to have a clearer idea on how good the GLD criterion is, we still need to have more thorough evaluations on it with more sophisticated learning algorithms, e.g., the optimal hierarchical chunking under development. The evaluations and advanced learning algorithms are the two main tasks in our future work.

---

<sup>2</sup>We wish to report our ongoing work in this direction with an *optimal hierarchical chunking* algorithm in a coming paper soon.

## References

- [1] J. Baker. Trainable grammars for speech recognition. In J. J. Wolf and D. H. Klatt, editors, *Speech Communication Papers Presented at the 97th Meeting of the Acoustical Society of America*, Cambridge, MA, 1979. MIT Press.
- [2] Eric Brill, David Magerman, Mitchell Marcus, and Beatrice Santorini. Deducing linguistic structure from the statistics of large corpora. In *Proceedings of the DARPA Speech and Natural Language Workshop*, pages 275–282, 1990.
- [3] Eric Brill and Mitchell Marcus. Automatically acquiring phrase structure using distributional analysis. In *Proceedings of 1992 DARPA Speech and Language Workshop*, Harriman, N.Y., 1992.
- [4] G. Carroll and E. Charniak. Learning probabilistic dependency grammars from labelled text. In *AAAI-92*, 1992.
- [5] G. Carroll and E. Charniak. Two experiments on learning probabilistic dependency grammars from corpora. Technical Report CS-92-16, Department of Computer Science, Brown University, 1992.
- [6] T. A. Cartwright and M. R. Brent. Segmenting speech without a lexicon: Evidence for a bootstrapping model of lexical acquisition. In *Proceedings of the 16th Annual Conference of the Cognitive Science Society*, pages 148–152, Erlbaum, Hillsdale, NJ, 1994.
- [7] T. A. Cartwright and M. R. Brent. Segmenting speech without a lexicon: The roles of phonotactics and speech source. In *First Meeting of the ACL Special Interest Group in Computational Phonology*, pages 83–90. Association for Computational Linguistics, 1994.
- [8] Gregory Chaitin. On the length of programs for computing finite binary sequences. *JACM*, 13:547–569, 1966.
- [9] Stanley F. Chen. Bayesian grammar induction for language modelling. In *ACL-95*, pages 228–235, Cambridge, Massachusetts, June 1995.
- [10] Stanley F. Chen. *Building Probabilistic Models for Natural Language*. PhD thesis, Harvard University, Cambridge, Massachusetts, 1996.
- [11] C. M. Cook, A. Rosenfeld, and A. R. Aronson. Grammatical inference by hill climbing. *Information Science*, 10:59–80, 1976.
- [12] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. John Wiley and Sons, Inc., New York, 1991.
- [13] Carl de Marken. Lexical heads, phrase structure and the induction of grammar. In D. Yarowsky and K. Church, editors, *Third Workshop on Very Large Corpora*, pages 14–26, Cambridge, Massachusetts, June 1995.
- [14] Carl de Marken. The unsupervised acquisition of a lexicon from continuous speech. Technical Report A.I. Memo No. 1558, AI Lab., MIT, Cambridge, Massachusetts, November 1995.
- [15] Carl de Marken. Linguistic structure as composition and perturbation. In *ACL-96*, pages 335–341, Santa Cruz, California, 1996.
- [16] T. Mark Ellison. *The Machine Learning of Phonological Structure*. PhD thesis, University of Western Australia, 1992.
- [17] W. Francis. Problems in assembling, describing and computerizing large corpora. In H. Bergenholtz and B. Schaefer, editors, *Empirische Textwissenschaft: Aufbau und Auswertung von Text-Corpora*, pages 110–123. Scriptor Verlag, Königstein, 1979.

- [18] W. N. Francis and H. Kucera. *Frequency Analysis of English Usage: Lexical and Grammar*. Houghton-Mifflin, Boston, 1982.
- [19] J. H. Holland. *Adaptation in Natural and Artificial Systems*. University of Michigan Press, Ann Arbor, 1975.
- [20] D. A. Huffman. A method for construction of minimum-redundancy codes. *Proceedings IRE*, 40:1098–1101, 1952.
- [21] F. Jelinek. Self-organized language modeling for speech recognition. Technical report, IBM T.J. Watson Research Center, Continuous Speech Recognition Group, Yorktown Heights, NY, 1985. Also in A. Waibel and K. F. Lee (eds.), *Readings in Speech Recognition*, pages 450-506, Morgan Kaufmann, San Mateo, California, 1990.
- [22] S. Kirkpatrick, C. D. Gelatt Jr., and M. P. Vecchi. Optimisation by stimulated annealing. *Science*, 200:671–680, 1983.
- [23] Chunyu Kit. Speeding up the Virtual Corpus approach to derivng and retrieving n-grams for any n from large-scale corpora. Manuscript, Department of Computer Science, University of Sheffield, August 1995.
- [24] A. N. Kolmogorov. Three approaches for defining the concept of ‘information quantity’. *Problem of Information Transmission*, 1:4–7, 1965.
- [25] K. Lari and S. J. Young. The estimation of stochastic context-free grammars using the inside-outside algorithm. *Computer Speech and Language*, 4:35–56, 1990.
- [26] Ming Li and P. M. B. Vitányi. *Introduction to Kolmogorov Complexity and its Applications*. Springer-Verlag, New York, 1993.
- [27] U. Manber and E. Myers. Suffix array: a new method for on-line string searches. In *First ASM-SIAM Symposium on Discrete Algorithms*, pages 319–327, Providence, 1990. American Mathematical Society.
- [28] M. Marcus, B. Santorini, and M. Marcinkiewicz. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330, 1993.
- [29] Makoto Nagao and Shinsuke Mori. A new method of n-gram statistics for large number of n and automatic extraction of words and phrases from large text data of japanese. In *COLING-94*, pages 611–615, 1994.
- [30] Donald Cort Olivier. *Stochastic Grammars and Language Acquisition Mechanisms*. PhD thesis, Harvard University, Cambridge, MA, 1968.
- [31] J. Ross Quinlan and Ronald L. Riverst. Inferring decision tree using the Minimum Description Length principle. *Information and Control*, 80:227–248, 1989.
- [32] Jorma Rissanen. Modelling by shortest data description. *Automatica*, 14:465–471, 1978.
- [33] Jorma Rissanen. *Stochastic Complexity in Statistical Inquiry*. World Scientific, N.J., 1989.
- [34] Jorma Rissanen and Eric Sven Ristad. Language acquisition in the MDL framework. In E. Ristad, editor, *Language Computations*. American Mathematical Society, Philadelphia, PA, 1994.
- [35] Claude Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27:379–423, 623–656, 1948.
- [36] R. J. Solomonoff. A new method for discovering the grammars of phrase structure languages. *Information Processing*, pages 258–290, 1959.

- [37] R. J. Solomonoff. The mechanization of linguistic learning. In *Proceedings of the 2nd International Conference on Cybernetics*, pages 180–193, 1960.
- [38] R. J. Solomonoff. A formal theory of inductive inference, part 1 and 2. *Information Control*, 7:1–22, 224–256, 1964.
- [39] Andreas Stolcke. *Bayesian Learning of Probabilistic Language Models*. PhD thesis, University of California at Berkeley, Berkeley, CA, 1994.
- [40] C. S. Wallace and D.M Boulton. An information measure for classification. *Computer Journal*, 11(2):185–194, 1968.
- [41] C. S. Wallace and P. R. Freeman. Estimation and inference by compact coding. *Journal of the Royal Statistical Society*, 49:240–251, 1987. discussion pages 251-265.
- [42] Gerard J. Wolff. Language acquisition and the discovery of phrase structure. *Language and Speech*, 23:255–269, 1980.
- [43] Gerard J. Wolff. Language acquisition, data compression and generalisation. *Language and Communication*, 2(1):57–89, 1982.

## Appendix A

– removed by the student session chair because the paper was too long –

## Appendix B

– removed by the student session chair because the paper was too long –