Modeling Acquisition of Word Structure with Lexicalized Grammar Learning

Introduction

This paper introduces a framework for learning structure in natural languages, and reports results from a simple application of it to learning word-syntax of an agglutinative language in an unsupervised manner. Arguably, the learning environment of children acquiring languages provides more information—by means of linguistic interaction and extra-linguistic information present in the learning setting— than the information provided to an unsupervised learner. However, completely unsupervised learning methods can still provide insights into how children acquire language, at least, (i) by setting a lower bound on what is learnable, (ii) by identifying type and quantity of cues in the input that is useful for successful learning, (iii) by testing different learning methods, algorithms and frameworks on the basis of how successful they are in learning from the data available to children and how well they match with the available data from developmental psycholinguistics. In this paper, we will first describe the general learning framework based on learning a lexicalized grammar, Categorial Grammar (CG, Ajdukiewicz, 1935; Bar-Hillel, 1953), then we will present our morphological learner in more detail, followed by the results obtained on testing the learner on learning morphology of Turkish from child directed speech from CHILDES database.

The learning algorithm uses techniques similar to unsupervised morphology learning systems such as Goldsmith (2001) and Creutz and Lagus (2007). However, this study tries to model human language acquisition more closely by using data from child directed speech and not assuming the complete data is available to the learner. Another major difference of this study is the emphasis on the structure learning. The model presented here learns a lexicalized word-grammar, which has similarities to other lexicalized grammar learners (e.g., Villavicencio, 2002; Zettlemoyer & Collins, 2005; Yao, Ma, Duarte, & Çöltekin, 2009).

Learning with Lexicalized Grammars

The model presented in this paper assumes that the language acquisition makes use of the process described in the following steps.

1. The learner receives a unit of input (e.g. hears an utterance).
2. Based on his/her (possibly incomplete) knowledge, the information from the environment and the interaction with the other speakers, the learner assigns an interpretation to the input utterance.
3. Based on correct interpretations learner updates his/her knowledge of language, i.e. the grammar of the target language.

For a model learning from only raw text, the step 2 above is more difficult than children learning languages. Children are aided by context and their interaction with the environment in figuring out the correct interpretations of the utterances they attend to.

The grammar formalism used in this work, CG, is a lexicalized grammar where the syntax of a language is fully specified in the lexicon. A small set of language independent rules is used for analyzing the input using the lexicalized grammar. These rules and example CG categories are presented in Figure 1, comprehensive descriptions can be found in Moortgat (2002) and Wood (1993).

By assuming such a lexicalized grammar, instead of a lexicon and separate rule set, the task of the learner is learning only a lexicon. Besides the computational convenience, the tight connection between lexicon and syntax is also in line with the experimental results from psycholinguistics (Bates & Goodman, 1997)

The use of CG for may seem an overkill for learning morphology. However, the morphologically complex languages (e.g. Turkish) may exhibit a more complicated word structure than the traditional methods for morphology assumes. This approach is also in line with the theoretical studies that postulate a *morphemic lexicon* (Bozsahin, 2002). Additionally, use of a powerful grammar formalism allows straightforward extensions of this model for learning natural language syntax.

Learning Morphology

The input to the learning algorithm is a series of unsegmented, unlabeled words. The model learns a morphemic CG lexicon, which is capable of generating and recognizing words of the input language. Each lexical item in the lexicon consists of the phonological (or orthographic) form of the morpheme associated with its CG category.

For every input received, the model first tries to find the best interpretation. The interpretation for this model consists of a segmentation of the input word and category assignments for each segment. For each input word, the model tries to parse the word. For the input words that the model cannot parse, it first tries to find segmentations of the word such that there is only one unknown segment. This results in a number of possible hypotheses about how to interpret the result. The model selects a hypotheses based on probability of segmentation, and probability of the parse given the current grammar. More formally, the model tries to find the maximum a posteriori (MAP) hypotheses. For the lexicalized grammar G, we try to find

$$\widehat{G} = \underset{G}{\operatorname{argmax}} P(G)P(input|G)$$

Using MAP estimate (Creutz & Lagus, 2007)—or equivalently minimum description length based approaches (Goldsmith, 2001)— is common in computational models of unsupervised morphology learning. There are two main differences of our model and the models cited above. First, use of a lexicalized grammar eliminates the need for estimating separate rule probabilities, and allows local changes directly related to the input at hand at every step, as well as providing potential extension of the system for learning more complex structures. Second, following a psycholinguisticaly more plausible approach, we do not provide the learner with the complete input, i.e. all the corpus, at once. We do not assume that learner has access to complete corpus, neither we assume that the learner stores all the input he/she receives. Instead, we keep some information on by updating the parameters of a number of probability distributions at every step. This is also in line with the studies that demonstrate the use of statistics by human learners (e.g., Saffran, Aslin, & Newport, 1996; Thompson & Newport, 2007).

The first component of the MAP estimate the $P(G)$ is the joint probability of the lexical items in G, where probability of each lexical item is calculated by the joint probability of the phonological form (ϕ) and the syntactic category (σ) assigned to it. With the simplifying assumption of independence of lexical items,

$$P(G) = \prod P(\phi)P(\sigma|\phi)$$

$P(\phi)$ estimated using a variation of the well known method method *letter successor variety* (LSV, Harris (1955)), for the unknown ϕ. Using this method, the phonological segments with high left and right unpredictability are assigned higher probabilities. $P(\sigma|\phi)$ is estimated from the lexicalized grammar.

The second component, $P(input|G)$, is the parse probability assigned by the probabilistic CG parser. As we do not assume the complete corpus is available to the learner, input is only the current word being processed. However, once we select an interpretation that contains a novel lexical item (ϕ, σ pair), we iterate over all lexical items containing ϕ, and re-evaluate them with the same criteria.

Experiment and Results

The model described above is tested using part of the child directed speech from the Turkish section of the CHILDES database, for which we had a semi-automatically constructed gold standard. The corpus consisted of 11731 word tokens and 1794 word types, 28415 morpheme tokens and 778 morpheme types.[1] We compare the overlap of the lexicons as well as the segmentation performance. We compare the results of the model with a no-segmentation baseline and the gold standard.

In this experiment we only allowed the categories W, W/W and $W\backslash W$. The categories correspond to a word, prefix or suffix respectively. This covers only a simplified word-grammar, however, the model can be extended to use a more complex fixed grammars, or generate new categories as needed during the learning process (Zettlemoyer & Collins, 2005; Yao et al., 2009).

Table 1 presents precision recall and F1 score of the lexicon overlap against the gold standard lexicon. The values in Table 2 are the performances of the lexicalized grammars in segmentation task. Table 2 compares performance measures of no-segmentation baseline and the gold standard lexicon, as well as a computationally oriented state-of-the-art unsupervised morphology learner (Creutz & Lagus, 2007) trained and tested on the same data.

Discussion and Conclusions

This paper presented a simple unsupervised model that learns word-syntax form raw data. Application of the model to child directed speech shows that the presented model performs well over the baseline model and achieves competitive results with a computationally oriented state-of-the-art model.

Even though this paper applies it to morphology acquisition, the learning framework used in this study is directly applicable to learning other phenomena, such as word order, in human languages. The use of lexicalized grammar simplifies the grammar learning task as it reduces the need for learning a language specific rule-system.

The unsupervised learning system presented here performs better than a reasonable baseline, and shows that even without additional knowledge from the environment, the raw input to children contains cues that would help learning word structure. It should also be noted that our model does not use some obvious information, such as distribution of forms and lengths of morphemes, that can be obtained from unlabeled input. As future work, we plan to extend this model to make use of this information, as well as learning more complex grammars.

[1] In the experiments reported here, orthographic forms of the words are used instead of phonological forms. Due to the relative orthographic transparency of Turkish, using orthographic transcriptions is a common practice in studies analyzing Turkish language data.

References

Ajdukiewicz, K. (1935). Die syntaktische konnexität. *Studia Philosophica*, *1*, 1-–27. (English translation in S. McCall (ed): *Polish Logic*, 207–231, Oxford University Press, 1967)

Bar-Hillel, Y. (1953). A quasi-arithmetical notation for syntactic description. *Language*, *29*, 47-–58.

Bates, E., & Goodman, J. C. (1997). On the inseparability of grammar and the lexicon: Evidence from acquisition, aphasia and real-time processing. *Language and Cognitive Processes*, *15*(5/6), 507–584.

Bozsahin, C. (2002). The combinatory morphemic lexicon. *Computational Linguistics*, *28*(2), 145–186.

Creutz, M., & Lagus, K. (2007). Unsupervised models for morpheme segmentation and morphology learning. *ACM Trans. Speech Lang. Process.*, *4*(1), 3.

Goldsmith, J. (2001). Unsupervised learning of the morphology of a natural language. *Computational Linguistics*, *27*(2), 153–198.

Harris, Z. (1955). From phoneme to morpheme. *Language*, *31*(2), 190-–222.

Moortgat, M. (2002). Encyclopedia of cognitive science. In L. Nagel (Ed.), (Vol. 1, p. 435-447). London, Nature Publishing Group.

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month old infants. *Science*, *274*(5294), 1926–1928.

Thompson, S. P., & Newport, E. L. (2007). Statistical learning of syntax: The role of transitional probability. *Language Learning and Development*(3), 1–42.

Villavicencio, A. (2002). *The acquisition of a unification-based generalised categorial grammar*. Unpublished doctoral dissertation, University of Cambridge.

Wood, M. M. (1993). *Categorial grammars*. London: Routledge.

Yao, X., Ma, J., Duarte, S., & Çöltekin Ç. (2009). Unsupervised syntax learning with categorial grammars using inference rules. In *Proc. of the 18th annual belgian-dutch conference on machine learning.* Tilburg.

Zettlemoyer, L. S., & Collins, M. (2005). Learning to map sentences to logical form: Structured classification with probabilistic categorial grammars. In *Proceedings of the twenty first conference on uncertainty in artificial intelligence (UAI-05).*

$$\begin{array}{llllll} \text{Forward} & A/B & B & \rightarrow & A & (>) \\ \text{Backward} & B & A\backslash B & \rightarrow & A & (<) \end{array}$$

(a)

Peter	:=	NP
Mary	:=	NP
likes	:=	(S\NP)/NP
walks	:=	S\NP

(b)

$$\begin{array}{ccc} \text{Peter} & \text{likes} & \text{Mary} \\ \overline{\text{NP}} & \overline{(\text{S}\backslash\text{NP})/\text{NP}} & \overline{\text{NP}} \\ \cline{2-3} & \multicolumn{2}{c}{\mathbf{S\backslash NP} \quad >} \\ \cline{1-3} \multicolumn{3}{c}{\mathbf{S} \quad <} \end{array}$$

(c)

Figure 1: (a) CG function application rules. (b) Example CG categories for English. (c) An example CG derivation.

	Precision	Recall	F-Score
Baseline	0.25	0.58	0.39
CG Learner	0.42	0.67	0.52

Table 1: Comparison of the lexicon learned by the model and no-segmentation baseline.

	Precision	Recall	F-Score
Baseline	0.25	0.20	0.23
CG Learner	0.31	0.59	0.41
Creutz and Lagus (2007)	0.50	0.42	0.48
GS Lexicon	0.65	1.00	0.79

Table 2: Comparison of the segmentation performance.