Learning grammatical categories using paradigmatic representation: Substitute words for language acquisition

Mehmet Ali Yatbaz[a,*], Volkan Cirik[a], Deniz Yuret[a]

[a]*Koç University, Istanbul, Turkey*

Abstract

Keywords: Language acquisition, Grammatical categorization, Distributional information, Corpus analysis, Computational modeling, Paradigmatic approach

1. Introduction

1.1. Psycholinguistic evidence relevant to substitutes

1.2. Comparison with previous distributional approaches

2. Substitute Words

In this study, we predict the syntactic category of a word in a given context based on its most likely substitute words. Note that the substitute word distribution is a function of the context only and is indifferent to the target word.

St Clair et al. (2010) demonstrated that learning left and right bigrams together was much more effective than learning them individually. Thus it is best to use both the left and the right context when estimating the probabilities for potential lexical substitutes. For example, in *"He lived in San Francisco suburbs."*, the token *San* would be difficult to guess from the left context but it is almost certain looking at the right context. We define

*Corresponding author. Address: Department of Computer Engineering, Koç University, 34450, Istanbul, Turkey

Email addresses: `myatbaz@ku.edu.tr` (Mehmet Ali Yatbaz), `vcirik@ku.edu.tr` (Volkan Cirik), `dyuret@ku.edu.tr` (Deniz Yuret)

c_w as the $2n-1$ word window centered around the target word position: $w_{-n+1} \dots w_0 \dots w_{n-1}$. The probability of a substitute word w in a given context c_w can be estimated as:

$$\begin{aligned} P(w_0 = w|c_w) &\propto P(w_{-n+1} \dots w_0 \dots w_{n-1}) & (1) \\ &= P(w_{-n+1})P(w_{-n+2}|w_{-n+1}) \dots P(w_{n-1}|w_{-n+1}^{n-2}) & (2) \\ &\approx P(w_0|w_{-n+1}^{-1})P(w_1|w_{-n+2}^{0}) \dots P(w_{n-1}|w_0^{n-2}) & (3) \end{aligned}$$

where w_i^j represents the sequence of words $w_i w_{i+1} \dots w_j$. In Equation 1, $P(w|c_w)$ is proportional to $P(w_{-n+1} \dots w_0 \dots w_{n+1})$ because the words of the context are fixed. Terms without w_0 are identical for each substitute in Equation 2 therefore they have been dropped in Equation 3. Finally, because of the Markov property of n-gram language model, only the closest $n-1$ words are used in the experiments.

Near the sentence boundaries the appropriate terms were truncated in Equation 3. Specifically, at the beginning of the sentence shorter n-gram contexts were used and at the end of the sentence terms beyond the end-of-sentence utterance were dropped.

To compute substitute probabilities we trained a language model using approximately 6.8 million tokens of child-directed speech data from the CHILDES corpus (MacWhinney, 2000) (excluding sections of [test-set]) We used SRILM (Stolcke, 2002) to build a 4-gram language model with Kneser-Ney discounting. Words that were observed less than 2 times in the LM training data were replaced by UNK tags, which gave us a vocabulary size of 21734. [What is the test data? Where should we put this?] [perplexity]

1. We need to clarify the tag set that is used during the experiments. May be it is better to give the whole mapping as an Appendix section.
2. Data statistics (it is common to all experiments)

3. Experiment 1: corpus analysis

In this section we replicate the corpus analyses of St Clair et al. (2010) and Mintz (2003).

3.1. Input Corpora

In order to be consistent with St Clair et al. (2010) and Mintz (2003), we use the same six corpora of child-directed speech from the CHILDES corpus (MacWhinney, 2000): Anne and Aran (Theakston, Lieven, Pine & Rowland, 2001), Eve (Crystal, 1974), Naomi (Sachs, 1983), Nina (Suppes, 1974), Peter (Bloom, Hood & Lightbown, 1974; Bloom, Lightbown, Hood, Bowerman, Maratsos & Maratsos, 1975). Following Mintz (2003) we only analyze the adulth utterances in sessions where the target child is 2.6 years old or younger.

3.1.1. Preprocessing

The grammatical category of words in CHILDES are extracted by first applying the MOR parser (MacWhinney, 2000) and then using the POST disambiguator (Sagae, MacWhinney & Lavie, 2004). The accuracy of CHILDES grammatical categories is approximately 95% (Parisse et al., 2000) and it is encoded in the MOR line of the CHILDES corpus.

We apply the following pre-processing steps (St Clair, Monaghan & Christiansen, 2010) initial to our analyses:

- All punctation, pause, trailing off and interruption marks are treated as utterance boundary marks.
- Repetitions of a word are kept in the text and their grammatical categories are automatically set to the grammatical category of the original word.
- Words that are grammatically necessary but not spoken are deleted (grammatical omissions).
- **Short usages ??**

3.2. Method

3.3. Results

4. Experiment 2: computational modelling of substitutes

St Clair et al. (2010) compare the effect of distributional cues from various type of frames on the learning grammatical category problem by using a feedforward connectionist model. We adopt their framework to compare the paradigmatic representation (substitute words) with the syntagmatic representation (flexible frames).

A prototypical connectionist model consists of input, hidden and output layers. Input and output layers are connected to each other through the hidden layer. The behavior of the output units are determined by the activity of the hidden layers which is triggered by the input layer.

4.1. Method

We train two connectionist models to compare flexible frames ($aX + Xb$) to the substitute words ($a * b$).

4.2. Architecture

For each model we represent the input in the following way,

- $aX+Xb$ **model:** The first and second half of the input units correspond to the preceding bigram (a) and the succeeding bigram (b), respectively. Thus two input units are activated for each target word.
- $a*b$ **model:** Each input unit represents a distinct substitute and input units that correpond to the substitutes are set to the number of their occurences in the target word substitutes.

Each output unit represents a distinct grammatical category therefore the models are expected to produce only one active ouput unit for each target word. Both models have 10 output units due to the standard labelling (Mintz, 2003).

Number of hidden units is set to 200 and initialized randomly for each model. **backprobagation, learning rate, sigmoid...**

4.3. Training and Testing

Table 1: 10 fold cross-validation accuracy of flexible frame ($aX + Xb$) and substitute based models on each child corpus are summarized. The training phase on each corpus is stopped after 50K word patterns are presented and the standard labelling is used. Substitute based model samples 16 substitutes for each target word. Standard errors are reported in parentheses.

Child	$aX + Xb$	*Substitutes*
Anne	.7545 (.0147)	**.8273 (.0087)**
Aran	.7164 (.0151)	**.8136 (.0096)**
Eve	.7605 (.0104)	**.8378 (.0199)**
Naomi	.7438 (.0156)	**.8165 (.0147)**
Nina	.7745 (.0199)	**.8494 (.0073)**
Peter	.7630 (.005)	**.8437 (.0061)**

5. Experiment 2

Number of subsitutes. We need to show 16 is better than 1 but 16+ is same

5.1. Input Corpora

5.2. Method

5.3. Results

6. Experiment 3

N-gram order 2,3

7. Experiment 4

What happens when we change the data size?
What happens when we change the vocabulary threshold?

7.1. Input Corpora

7.2. Method

7.3. Results

8. Experiment 5

Left/rigth context substitute

8.1. Input Corpora

8.2. Method

8.3. Results

9. Experiment 6

Other languages that we have in CHILDES

10. Experiment 7

What happens if some of the words are given (semi-supervised setting)

10.1. Input Corpora

10.2. Method

10.3. Results

11. General Discussion

References

Bloom, L., Hood, L., & Lightbown, P. (1974). Imitation in language development: If, when, and why. *Cognitive Psychology*, *6*, 380 – 420. doi:10.1016/0010-0285(74)90018-8.

Bloom, L., Lightbown, P., Hood, L., Bowerman, M., Maratsos, M., & Maratsos, M. P. (1975). Structure and variation in child language. *Monographs of the society for Research in Child Development*, (pp. 1–97).

Crystal, D. (1974). Roger brown, a first language: the early stages. cambridge, mass.: Harvard university press, 1973. pp. xi + 437. *Journal of Child Language*, *1*, 289–307. doi:10.1017/S030500090000074X.

MacWhinney, B. (2000). *The CHILDES Project: Tools for Analyzing Talk, Volume II: The Database* volume 2. Lawrence Erlbaum.

Mintz, T. H. (2003). Frequent frames as a cue for grammatical categories in child directed speech. *Cognition*, *90*, 91 – 117. URL: http://www.sciencedirect.com/science/article/pii/S0010027703001409. doi:10.1016/S0010-0277(03)00140-9.

Parisse, C. et al. (2000). Automatic disambiguation of morphosyntax in spoken language corpora. *Behavior Research Methods, Instruments, & Computers*, *32*, 468–481.

Sachs, J. (1983). Talking about the there and then: The emergence of displaced reference in parent-child discourse. *Childrens language*, *4*.

Sagae, K., MacWhinney, B., & Lavie, A. (2004). Automatic parsing of parental verbal input. *Behavior Research Methods, Instruments, & Computers*, *36*, 113–126.

St Clair, M. C., Monaghan, P., & Christiansen, M. H. (2010). Learning grammatical categories from distributional cues: flexible frames for language acquisition. *Cognition*, *116*, 341–60. URL: http://www.biomedsearch.com/nih/Learning-grammatical-categories-from-distributic

Stolcke, A. (2002). Srilm-an extensible language modeling toolkit. In *Proceedings International Conference on Spoken Language Processing* (pp. 257–286).

Suppes, P. (1974). The semantics of childrens language. *American psychologist*, *29*, 103.

Theakston, A. L., Lieven, E. V., Pine, J. M., & Rowland, C. F. (2001). The role of performance limitations in the acquisition of verb-argument structure: An alternative account. *Journal of child language*, *28*, 127–152.