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

Original Idea

3.1. Input Corpora
3.2. Method
3.3. Results

4. Experiment 2

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

Table 1: The 70% of the sentences in the union of 6 child corpora are used as the training set while the remaining 30% sentences of each corpus are used as the test set. Average classification accuracy (10 runs) of the supervised connectionist model for the standard labelling on each child corpus after 10K words of training are summarized and the corresponding standard errors are reported in parentheses.

	Frames				Number of Substitute Words	
Child	aX	Xb	aXb	aX + Xb	1	16
Anne	.5317 (.0374)	.5065 (.0214)	.3788 (.0279)	.6190 (.0253)	.6221 (.0333)	.7829 (.0059)
Aran	.5098 (.0344)	.4779 (.0171)	.5098 (.0289)	.5863 (.0227)	.5916 (.0371)	.7597 (.007)
Eve	.5408 (.036)	.4905 (.0183)	.5408 (.0186)	.6125 (.0229)	.6189 (.0287)	.7862 (.0058)
Naomi	.5250 (.0269)	.4922 (.0205)	.3860 (.028)	.6019 (.0218)	.6021 (.0322)	.7672 (.0071)
Nina	.5412 (.0304)	.5032 (.0213)	.3950 (.0223)	.6308 (.0277)	.6474 (.0331)	.8089 (.0089)
Peter	.5359 (.036)	.5092 (.0216)	.5359 (.0188)	.6250 (.0252)	.6315 (.0263)	.7974 (.0082)

4.1. Input Corpora

4.2. Method

4.3. Results

5. Experiment 3

N-gram order 2,3

6. Experiment 4

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

6.1. Input Corpora

6.2. Method

6.3. Results

7. Experiment 5

Left/rigth context substitute

References

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

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).