



Contextual grammars and natural languages

Tomáš Mikolov, 2007



Natural signals

- Non-random sequences of symbols
- Small subset of all signals
- Example: speech, written text, image, ... anything



Natural languages

- usually considered as a single natural signal with small discrete alphabet (simplification!)
- language models try to **effectively** capture all regularities within some language
- optimal model needs to understand the language - example: "five and four is nine"
- how to say how well does some model understand a language?



Statistical models

- Information-theoretic based
- Language is viewed as a pseudorandom process
- Surprisingly good performance (even in machine translation)



Example: bigram model

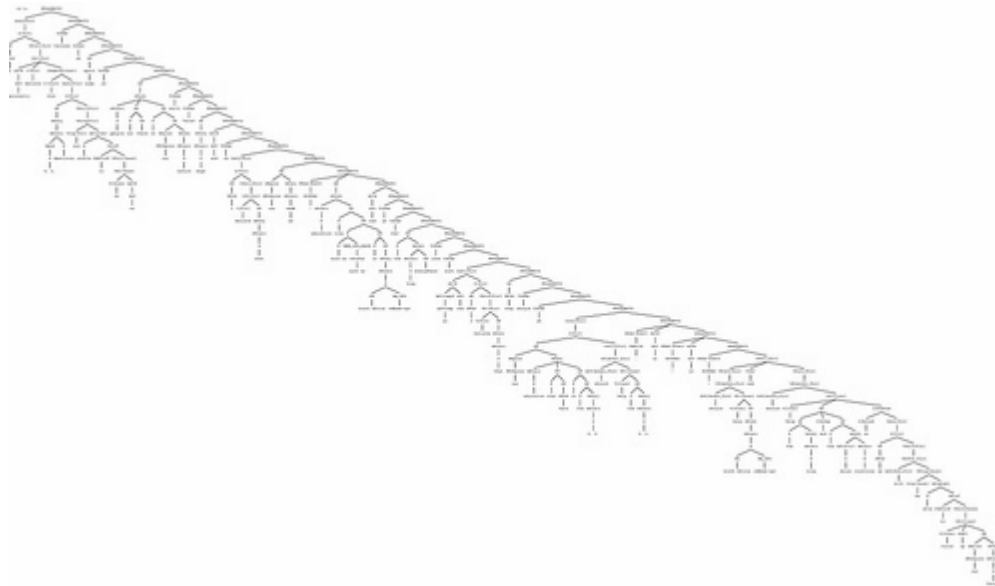
- Consider a sequence 'AABABAABBAAABABAB'; what symbol comes next?
- $P(A|B)=4/5$
- $P(B|B)=1/5$

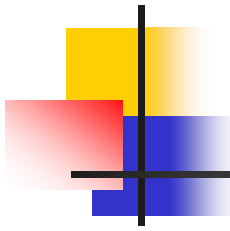
- weakness: can't effectively model long context dependencies
- However, for infinite training data, ngram statistics are optimal!



Structural models

- Linguistically motivated
- Tries to capture structure of a sentence





Example: probabilistic context free grammar (PCFG)

- Consider sequences 'The train was very fast' and 'The train was fast very'; which one is correct?
- In practice, both sentences may appear, but the second one is much less probable
- weakness: computationally expensive, poor performance for spontaneous speech



Contextual grammars in NLP

- Well-formed strings (wfs) & non-well-formed strings (nwfs); fuzzy set?
- Simple contextual grammar $G=(A, L_0, C_0)$, where L_0 is a finite language over alphabet A (finite set of primitive wfs), while C_0 is a finite set of contexts over A



Contextual grammars in NLP

- Language generated by G is a minimal set L of strings over A such that L_0 is contained in L and for any string x in L and any context (u, v) in C_0 the string uxv is in L_0



Evaluation of different natural language modeling approaches

- Information theory: perplexity (average branching factor)

$$PPL = \sqrt[n]{\prod_{i=1}^n \frac{1}{P(w_i | w_{1..i-1})}}$$



Evaluation of different natural language modeling approaches

- Using some existing natural language recognizer (OCR or ASR systems)
- Evaluation by rescoreing hypotheses from recognizers