Traitement automatique du langage
TP 5 — Exercise: Language Modelling, PoS Tagging, Syntax

Haozhou Wang
Exercises prepared by Tanja Samardžić and Asheesh Gulati
14.11.2019

Submit by 20.11.2019 midnight.

Consider the following annotated corpus (1, 2) and the syntactic analysis of the first part (3).

1.	[We/PRP identify/VB remaining/VBG gaps/NNS in/IN knowledge/NN ./. We/PRP want/VB to/TO boost/VB their/PRP knowledge/NN level/NN ,/, get/VB feedback/NN on/IN the/DT gaps/NNS remaining/VBG in/IN their/PRP knowledge/NN ./.
2.	We/PRP want/VB to/TO get/VB feedback/NN on/IN their/PRP knowledge/NN ./.

3.

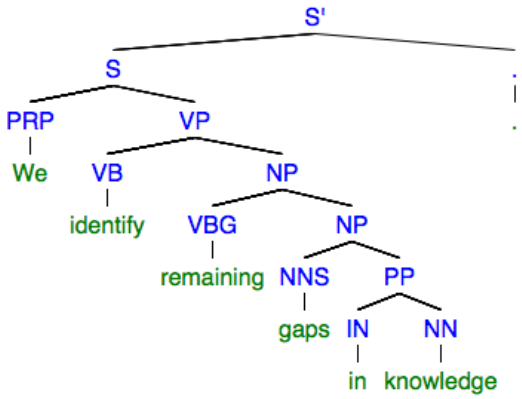

Figure 1: Parse tree of the first sentence in (1)

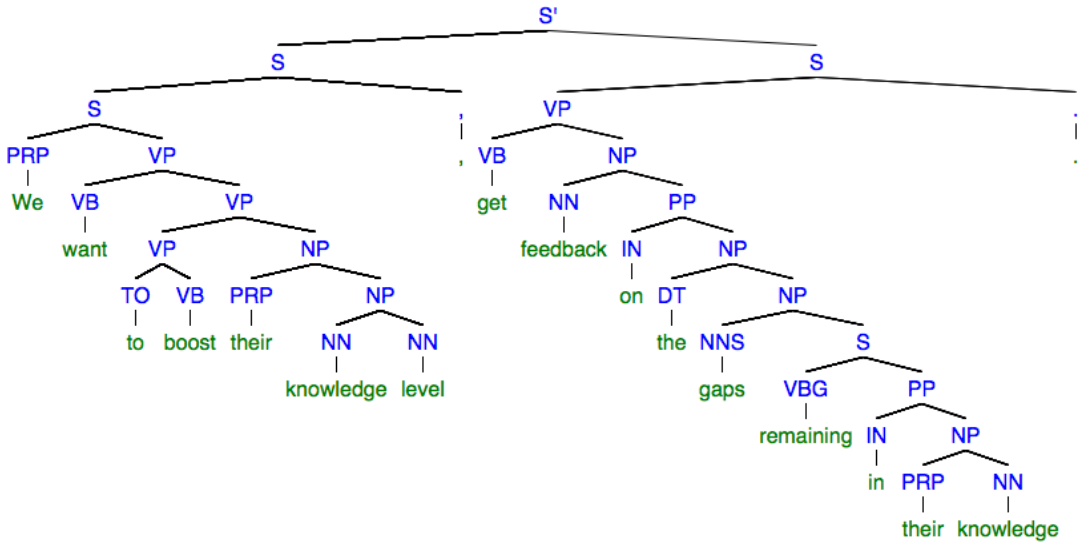

Figure 2: Parse tree of the second sentence in (1)

1 Language Modelling

1. Formulate the language model problem for the sentence in (2).
2. Decompose the language model for the sentence in (2) using the chain rule.
3. Decompose the language model for the sentence in (2) using the Markov assumption.
4. Estimate the probability of the sentence in (2) using the Markov decomposition, maximum likelihood estimate and the corpus in (1) for training.
5. Estimate the probability of the sentence in (2) using the Markov decomposition, maximum likelihood estimate with Jelinek-Mercer smoothing (assume $\lambda = \frac{1}{2}$) and the corpus in (1) for training.

2 PoS tagging

1. Formulate the PoS tagging model problem for the sentence in (2).
2. Decompose the tagging model for the sentence in (2) applying Hidden Markov Model.
3. Estimate the tagging probability of the sentence in (2) using Hidden Markov Model, maximum likelihood estimate and the corpus in (1) for training.

3 Syntax

1. Define a grammar that generates the trees in (1).
2. Draw a tree for the sentence in (2) using the same grammar as in (1).
3. Estimate the probability of the tree in (2) using maximum likelihood estimate and the corpus in (1) for training.