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

Tanja Samardžić, Asheesh Gulati

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.

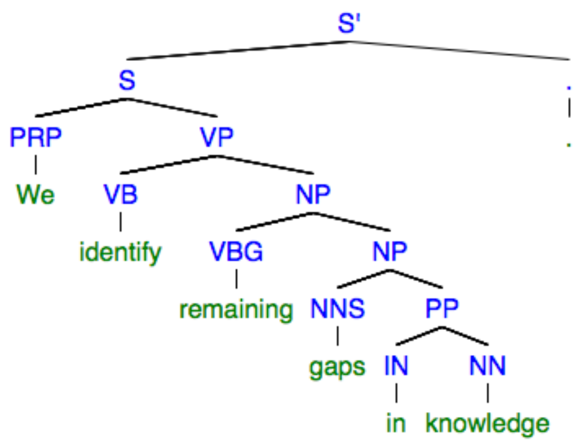

Figure 1: Parse tree of the first sentence in (1)

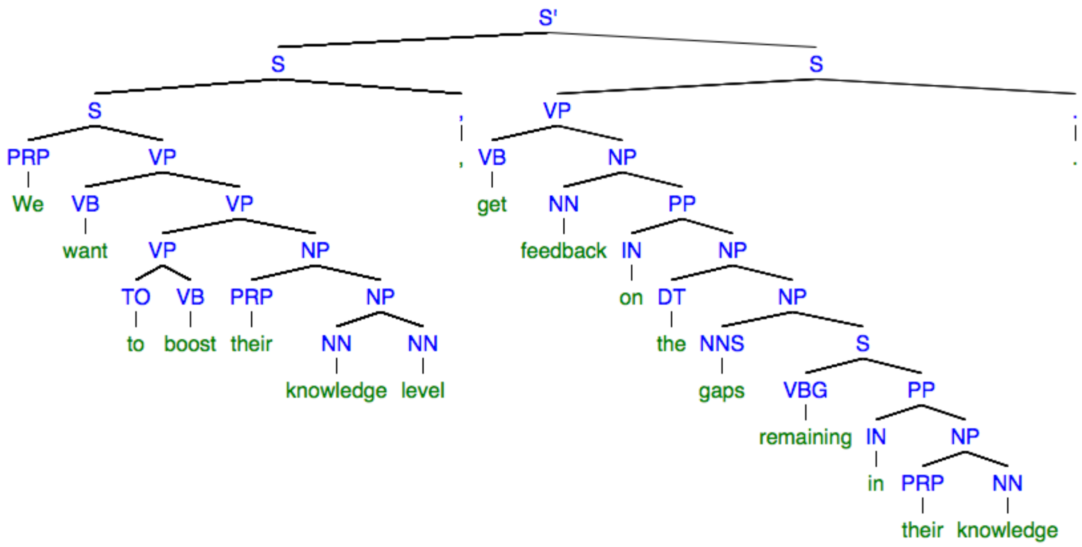

Figure 2: Parse tree of the second sentence in (1)

1 Language Modelling

1. Formulate the language model problem for the sentence in (2).

   $p(*, we, want, to, get, feedback, on, their, knowledge, ., STOP)$

2. Decompose the language model for the sentence in (2) using the chain rule.

   $$p(*, we, want, to, get, feedback, on, their, knowledge, ., STOP) =$$
   $$= p(STOP|*, we, want, to, get, feedback, on, their, knowledge, .)$$
   $$\cdot\, p(.|*, we, want, to, get, feedback, on, their, knowledge)$$
   $$\cdot\, p(knowledge|*, we, want, to, get, feedback, on, their)$$
   $$\cdot\, p(their|*, we, want, to, get, feedback, on)$$
   $$\cdot\, p(on|*, we, want, to, get, feedback)$$
   $$\cdot\, p(feedback|*, we, want, to, get)$$
   $$\cdot\, p(get|*, we, want, to)$$
   $$\cdot\, p(to|*, we, want)$$
   $$\cdot\, p(want|*, we)$$
   $$\cdot\, p(we|*)$$

3. Decompose the language model for the sentence in (2) using the Markov assumption.

   Bigram language model:

   $p(*, we, want, to, get, feedback, on, their, knowledge, ., STOP) =$
   $= p(STOP|.)$
   $\cdot\, p(.|knowledge)$
   $\cdot\, p(knowledge|their)$
   $\cdot\, p(their|on)$
   $\cdot\, p(on|feedback)$
   $\cdot\, p(feedback|get)$
   $\cdot\, p(get|to)$
   $\cdot\, p(to|want)$
   $\cdot\, p(want|we)$
   $\cdot\, p(we|*)$

4. Estimate the probability of the sentence in (2) using the Markov decomposition, maximum likelihood estimate and the corpus in (1) for training.

   $p(*, we, want, to, get, feedback, on, their, knowledge, ., STOP) =$
   $= 1$
   $\cdot\, \frac{2}{3}$
   $\cdot\, 1$
   $\cdot\, 0$
   $\cdot\, 1$
   $\cdot\, 1$
   $\cdot\, 0$
   $\cdot\, 1$
   $\cdot\, \frac{1}{2}$
   $\cdot\, 1$
   $= 0$

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.

   $p(*, we, want, to, get, feedback, on, their, knowledge, ., STOP) =$
   $= (\frac{1}{2} + \frac{2}{58})$
   $\cdot (\frac{2}{6} + \frac{2}{58})$
   $\cdot (\frac{1}{2} + \frac{3}{58})$
   $\cdot (0 + \frac{2}{58})$
   $\cdot (\frac{1}{2} + \frac{1}{58})$
   $\cdot (\frac{1}{2} + \frac{1}{58})$
   $\cdot (0 + \frac{1}{58})$
   $\cdot (\frac{1}{2} + \frac{1}{58})$
   $\cdot (\frac{1}{4} + \frac{1}{58})$
   $\cdot (\frac{1}{2} + \frac{2}{58})$
   $\simeq 0.00000127$

2 PoS Tagging

1. Formulate the PoS tagging model problem for the sentence in (2).

   $p(we, want, to, get, feedback, on, their, knowledge, .,$
   $*, PRP, VB, TO, VB, NN, IN, PRP, NN, ., STOP)$

2. Decompose the tagging model for the sentence in (2) applying Hidden Markov Model.

   Bigram HMM:

   $p(we, want, to, get, feedback, on, their, knowledge, .,$
   $*, PRP, VB, TO, VB, NN, IN, PRP, NN, ., STOP) =$
   $= p(STOP|.)$
   $\cdot p(.|NN) \cdot p(.|.)$
   $\cdot p(NN|PRP) \cdot p(knowlegde|NN)$
   $\cdot p(PRP|IN) \cdot p(their|PRP)$
   $\cdot p(IN|NN) \cdot p(on|IN)$
   $\cdot p(NN|VB) \cdot p(feedback|NN)$
   $\cdot p(VB|TO) \cdot p(get|VB)$
   $\cdot p(TO|VB) \cdot p(to|TO)$
   $\cdot p(VB|PRP) \cdot p(want|VB)$
   $\cdot p(PRP|*) \cdot p(we|PRP)$

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

$$p(we, want, to, get, feedback, on, their, knowledge, .,$$
$$*, PRP, VB, TO, VB, NN, IN, PRP, NN, ., STOP) =$$
$$= 1$$
$$\cdot \frac{2}{5} \cdot 1$$
$$\cdot \frac{2}{4} \cdot \frac{3}{5}$$
$$\cdot \frac{1}{3} \cdot \frac{2}{4}$$
$$\cdot \frac{1}{5} \cdot \frac{2}{3}$$
$$\cdot \frac{1}{4} \cdot \frac{1}{5}$$
$$\cdot 1 \cdot \frac{1}{4}$$
$$\cdot \frac{1}{4} \cdot 1$$
$$\cdot \frac{2}{4} \cdot \frac{1}{4}$$
$$\cdot 1 \cdot \frac{2}{4}$$
$$\simeq 0.0000005105$$

3 Syntax

1. Define a grammar that generates the trees in (1).

   S' → S .
   S' → S S
   S → PRP VP
   S → S ,
   S → VP .
   S → VBG PP
   VP → VB NP
   VP → VB VP
   VP → VP NP
   VP → TO VB
   NP → VBG NP
   NP → NNS PP
   NP → PRP NN
   NP → PRP NP
   NP → NNS S
   NP → DT NP
   NP → NN PP
   NP → NN NN
   PP → IN NN
   PP → IN NP
   PRP → We
   PRP → their
   VB → identity
   VB → get
   VB → boost
   VB → want
   VBG → remaining
   IN → in
   IN → on
   NN → knowledge
   NN → feedback
   NN → level
   NNS → gaps
   TO → to
   DT → the
   , → ,
   . → .

2. Draw a tree for the sentence in (2) using the same grammar as in (1).

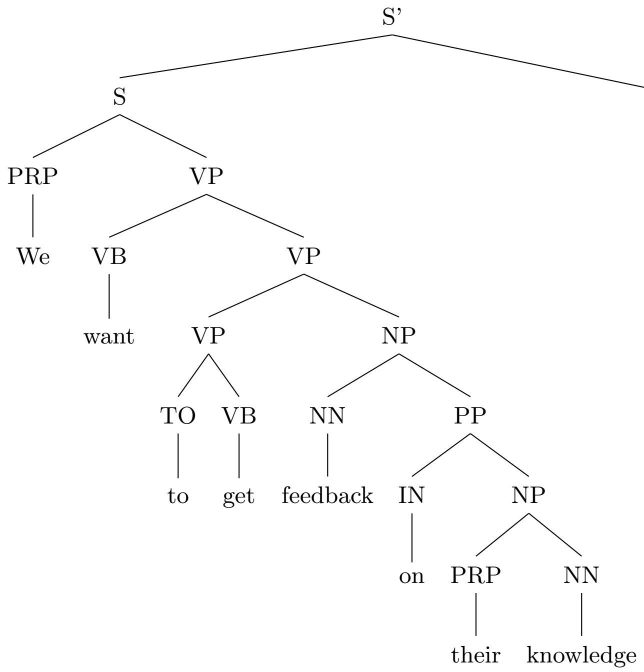

Figure 3: Parse tree of the sentence in the annotated corpus (3)

3. Estimate the probability of the tree in (2) using maximum likelihood estimate and the corpus in (1) for training.

$$
\begin{aligned}
&p(tree) = \\
&= p(S' \to S\ .) \\
&\cdot p(S \to PRP\ VP) \\
&\cdot p(VP \to VB\ VP) \\
&\cdot p(VP \to VP\ NP) \\
&\cdot p(VP \to TO\ VB) \\
&\cdot p(NP \to NN\ PP) \\
&\cdot p(PP \to IN\ NP) \\
&\cdot p(NP \to PRP\ NN) \\
&\cdot p(PRP \to We) \\
&\cdot p(VB \to want) \\
&\cdot p(TO \to to) \\
&\cdot p(VB \to get) \\
&\cdot p(NN \to feedback) \\
&\cdot p(IN \to on)
\end{aligned}
$$

$\cdot\, p(PRP \rightarrow their)$
$\cdot\, p(NN \rightarrow knowledge)$
$\cdot\, p(. \rightarrow .)$

$= \frac{1}{2}$
$\cdot\, \frac{2}{5}$
$\cdot\, \frac{1}{5}$
$\cdot\, \frac{1}{5}$
$\cdot\, \frac{1}{5}$
$\cdot\, \frac{1}{8}$
$\cdot\, \frac{2}{3}$
$\cdot\, \frac{1}{8}$
$\cdot\, \frac{2}{4}$
$\cdot\, \frac{1}{4}$
$\cdot\, 1$
$\cdot\, \frac{1}{4}$
$\cdot\, \frac{1}{5}$
$\cdot\, \frac{1}{3}$
$\cdot\, \frac{2}{4}$
$\cdot\, \frac{3}{5}$
$\cdot\, 1$
$\simeq 0.0000000102$