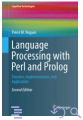
Language Processing with Perl and Prolog

Chapter 8: Part-of-Speech Tagging Using Stochastic Techniques

Pierre Nugues

Lund University
Pierre.Nugues@cs.lth.se
http://cs.lth.se/pierre_nugues/



POS Annotation with Statistical Methods

Modeling the problem:

$$t_1, t_2, t_3, \ldots, t_n \rightarrow \text{noisy channel} \rightarrow w_1, w_2, w_3, \ldots, w_n.$$

The optimal part of speech sequence is

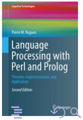
$$\hat{T} = \underset{t_1, t_2, t_3, \ldots, t_n}{\arg\max} P(t_1, t_2, t_3, \ldots, t_n | w_1, w_2, w_3, \ldots, w_n),$$

The Bayes' rule on conditional probabilities:

$$P(A|B)P(B) = P(B|A)P(A).$$

$$\hat{T} = \underset{T}{\arg\max} P(T)P(W|T).$$

$P(T)$ and $P(W|T)$ are simplified and estimated on hand-annotated corpora, the "gold standard".

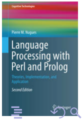


The First Term: *N*-Gram Approximation

$$P(T) = P(t_1, t_2, t_3, ..., t_n) \approx P(t_1)P(t_2|t_1)\prod_{i=3}^{n} P(t_i|t_{i-2}, t_{i-1}).$$

If we use a start-of-sentence delimiter <s>, the two first terms of the product, $P(t_1)P(t_2|t_1)$, are rewritten as
$P(< s >)P(t_1| < s >)P(t_2| < s >, t_1)$, where $P(< s >) = 1$.
We estimate the probabilities with the maximum likelihood, P_{MLE}:

$$P_{MLE}(t_i|t_{i-2}, t_{i-1}) = \frac{C(t_{i-2}, t_{i-1}, t_i)}{C(t_{i-2}, t_{i-1})}.$$



Sparse Data

If N_p is the number of the different parts-of-speech tags, there are $N_p \times N_p \times N_p$ values to estimate.
If data is missing, we can back off to bigrams:

$$P(T) = P(t_1, t_2, t_3, ..., t_n) \approx P(t_1) \prod_{i=2}^{n} P(t_i | t_{i-1}).$$

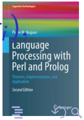
Or to unigrams:

$$P(T) = P(t_1, t_2, t_3, ..., t_n) \approx \prod_{i=1}^{n} P(t_i).$$

And finally, we can combine linearly these approximations:

$$P_{LinearInter}(t_i | t_{i-2} t_{i-1}) = \lambda_1 P(t_i | t_{i-2} t_{i-1}) + \lambda_2 P(t_i | t_{i-1}) + \lambda_3 P(t_i),$$

with $\lambda_1 + \lambda_2 + \lambda_3 = 1$, for example, $\lambda_1 = 0.6$, $\lambda_2 = 0.3$, $\lambda_3 = 0.1$.



The Second Term

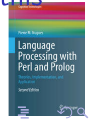
The complete word sequence knowing the part-of-speech sequence is usually approximated as:

$$P(W|T) = P(w_1, w_2, w_3, ..., w_n|t_1, t_2, t_3, ..., t_n) \approx \prod_{i=1}^{n} P(w_i|t_i).$$

Like the previous probabilities, $P(w_i|t_i)$ is estimated from hand-annotated corpora using the maximum likelihood:

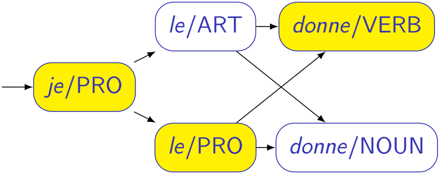
$$P_{MLE}(w_i|t_i) = \frac{C(w_i, t_i)}{C(t_i)}.$$

For N_w different words, there are $N_p \times N_w$ values to obtain. But in this case, many of the estimates will be 0.

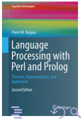


An Example

Je le donne 'I give it'

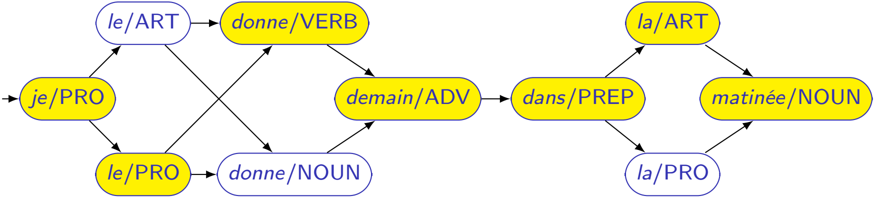


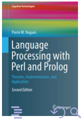
1. $P(pro|\emptyset) \times P(art|\emptyset, pro) \times P(verb|pro, art) \times P(je|pro) \times P(le|art) \times P(donne|verb)$
2. $P(pro|\emptyset) \times P(art|\emptyset, pro) \times P(noun|pro, art) \times P(je|pro) \times P(le|art) \times P(donne|noun)$
3. $P(pro|\emptyset) \times P(pro|\emptyset, pro) \times P(verb|pro, pro) \times P(je|pro) \times P(le|pro) \times P(donne|verb)$
4. $P(pro|\emptyset) \times P(pro|\emptyset, pro) \times P(noun|pro, pro) \times P(je|pro) \times P(le|pro) \times P(donne|noun)$



Viterbi (Informal)

Je le donne demain dans la matinée 'I give it tomorrow in the morning'





Viterbi (Informal)

The term brought by the word *demain* has still the memory of the ambiguity of *donne*: $P(adv|verb) \times P(demain|adv)$ and $P(adv|noun) \times P(demain|adv)$.

This is no longer the case with *dans*.

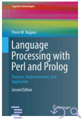
According to the noisy channel model and the bigram assumption, the term brought by the word *dans* is $P(dans|prep) \times P(prep|adv)$.

It does not show the ambiguity of *le* and *donne*.

The subsequent terms will ignore it as well.

We can discard the corresponding paths.

The optimal path does not contain nonoptimal subpaths.



Supervised Learning: A Summary

- Needs a manually annotated corpus called the **Gold Standard**
- The Gold Standard may contain errors (*errare humanum est*) that we ignore
- A classifier is trained on a part of the corpus, the **training set**, and evaluated on another part, the **test set**, where automatic annotation is compared with the "Gold Standard"
- **N-fold cross validation** is used avoid the influence of a particular division
- Some algorithms may require additional optimization on a development set
- Classifiers can use statistical or symbolic methods