Lemmatisation

La **lemmatisation** désigne l'analyse lexicale du contenu d'un texte regroupant les mots d'une même famille. Cela correspond a la forme canonique du mot, par exemple, l’infinitif pour un verbe, ou la forme masculin singulier pour les autres mots. C’est ce qu’on peut appeler la forme de base d’un mot, celle présente dans le dictionnaire.

German, English, French, Italian, Danish, Dutch, Spanish, Bulgarian, Russian, Portuguese, Galician, Greek, Chinese, Swahili, Slovak, Slovenian, Latin, Estonian, Polish, Romanian, Czech, Coptic and old French texts

Nous ne parlerons que de l’anglais et du français afin de comprendre les exemples et de nous assurer de leur justesse.

Baum-Welch algorithme, modèles caché de Markov.

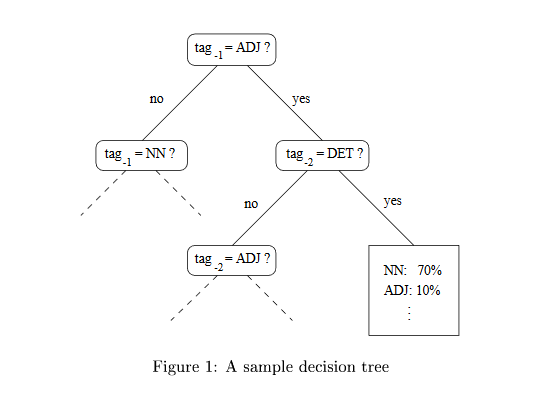
L’algo permet de connaitre les inconnus des modèles de Markov cachés et ainsi de pouvoir l’utiliser.

Pour cela, il utilise l’algorithme de maximisation de l’espérance, ou EM

Problème : a plus de 5000 phrases d’entrainement, la précision de l’algorithme ne s’améliore pas

Les treetaggers ne sont pas seulement des trigrams tagger, ils utilisent également, en plus du contexte classique des trigram AKA les deux mots précédents, mais également ce que ces mots ne sont pas, par exemple : le mot précédent est un adjectif, mais le mot encore avant n’est ni un adjectif, ni un déterminant.

Ngram tagger : ils utilisent les probabilités conjointes du nombres d’occurrences des trigram et bigram tagger correspondants existe, différant sur ce point par rapport aux treetaggers



Afin d’analyser un mot, il suffit de commencer a la racine, et de répondre petit à petit aux questions jusqu’à arriver à une feuille, nous donnant une probabilité pour notre mot, ensuite, il suffit d’appliquer cette probabilité au mot en lui-même et non pas seulement à ceux qui l’entoure

Par exemple : « tu peux le faire »

Ici, le, qui a une forte probabilité d’être un déterminant

tu PRO:PER tu

peux VER:pres pouvoir

le PRO:PER le

faire VER:infi faire

(http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/data/french-tagset.html)

le treetagger reconnait néanmoins qu’il s’agit bien d’un pronom personnel et non d’un déterminant, grâce à l’arbre de décision