Traitement du langage
Approches linguistiques et empiriques
Exemples de questions d'examen

A.A. 2019-2020

January 10, 2020

1 Language Models

- The derivation of language model probabilities is in three steps. Expand the formula for a trigram language model, discussing
  - the chain rule to expand the joint probability
  - the independence assumption
  - smoothing
- We are given the following corpus:

  <s> I am Sam </s>

  <s> Sam I am </s>

  <s> I am Sam </s>

  <s> I do not like green eggs and Sam </s>

  If we use back-off smoothing what is P(Sam|am)? Include <s> and </s> in your counts just like any other token.

2 Mesures d'évaluation

1. Considérez la matrice de confusion suivante qui décrit la performance d'un classifieur ternaire:

		Vrai label t		
		$+1$	-1	0
Label	$+1$	a	b	c
prédit y	-1	d	e	f
	0	g	h	i

a) Quelle est l'exactitude (*accuracy*) de ce classifieur?

b) Quelle est la précision de ce classifieur pour le label 0?

c) Quel est le rappel (*recall*) de ce classifieur pour le label $+1$?

d) Quelles est la mesure d'évaluation qui correspond à la probabilité $P(y = +1 \mid t = +1)$?

2. Considérez un classifieur A dont la précision vaut 86.1% et le rappel vaut 80.9%. Quelles est la mesure F de ce classifieur? Ce score est-il meilleur que la mesure F d'un classifieur dont la précision vaut 72% et le rappel 88%? Justifiez votre réponse.

3 Modèles de Markov cachés (HMMs)

Un HMM tri-gramme suppose que la probabilité d'un tag ne dépend que des deux tags qui le précèdent, commme discuté en classe. Un HMM bi-gramme impose des hypothèses d'indépendance encore plus fortes. En particulier, un HMM bi-gramme suppose qu'un tag ne dépend que du tag qui le précède immédiatement.

Comme décrit en classe, l'**algorithme de Viterbi** calcule la probabilité de la sequence de n tags la plus probable associée à une séquence de n mots. Pour un HMM bi-gramme dont le tag initial est $*$, cet algorithme implémente la récursion suivante (très proche de celle associée aux HMM tri-grammes):

- Base:
$$\pi(0, *) = 1$$

- Induction: pour tous les tags v and tout $1 \geq k \leq n$
$$\pi(k, v) = \max_u(\pi(k-1, u) \times q(v \mid u) \times e(x_k \mid v))$$

Dans ces clauses, $q(v \mid u)$ est la probabilité de passer du tag u au tag v. La probabilité d'émettre le k-ième mot étant donné le tag v est notée $e(x_k \mid v)$.

Considérez the paramètres suivants d'un HMM bi-gramme et résolvez les problèmes ci-dessous:

	Paramètres de transition			Paramètres d'émission		
	N	V	$STOP$	*love*	*models*	*scientists*
$*$	0.8	0.2	0.0	0.0	0.0	0.0
N	0.2	0.4	0.4	0.3	0.4	0.3
V	0.7	0.2	0.1	0.7	0.2	0.1
$STOP$	0.0	0.0	0.0	0.0	0.0	0.0

1. Etant donné ces paramètres et la séquence de mots *scientists love models*, complètez la table suivante des valeurs $\pi(k, v)$ selon l'algorithme de Viterbi pour un HMM bigramme donné plus haut:

	$_0$	scientists$_1$	love$_2$	models$_3$
$*$	1	0	0	0
N	0	$0.8 \times 0.3 = 0.24$		
V	0	$0.2 \times 0.1 = 0.02$		

2. Pour la séquence d'entrée *scientists love models*, quelle est la séquence de tags la plus probable dans laquelle le tag associé avec le mot *models* est N? Justifiez votre réponse.

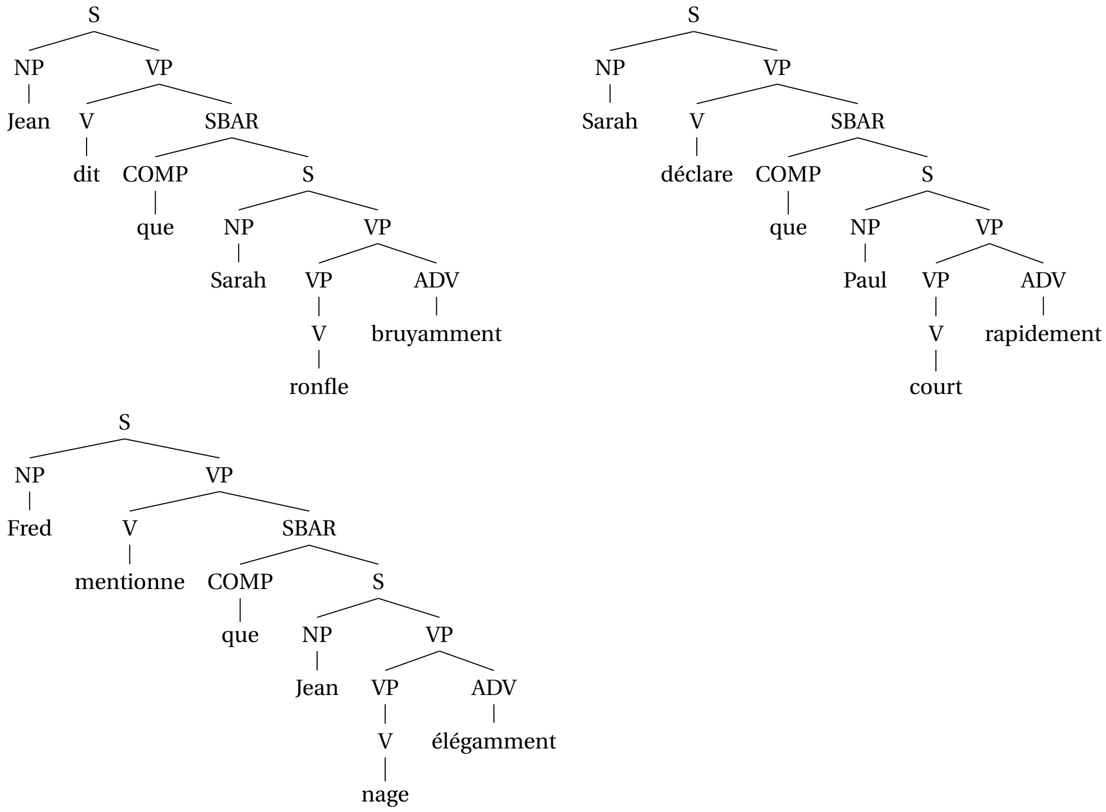
3. Pour la séquence d'entrée *scientists love models*, quelle est la séquence de tags la plus probable dans laquelle le tag associé avec le mot *models* est V? Justifiez votre réponse.

4 PCFGs

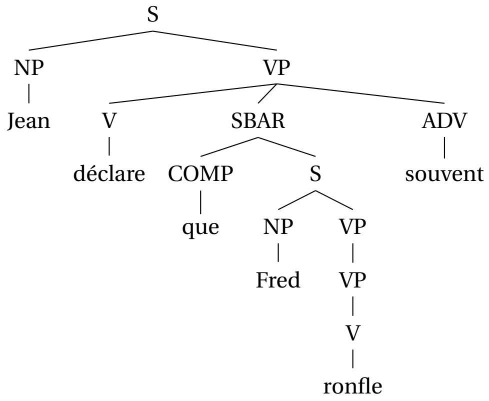
Les PCFG constituent le modèle le plus simple de parsing statistique, mais leur performance est généralement considérée comme insuffisante. Expliquez les raisons de cette inadéquation.

5 Extraction de grammaires

Soit un treebank constitué des trois arbres syntaxiques suivants.



1. Décrivez une grammaire probabiliste de ce corpus, c'est-à-dire notez les règles de grammaire et calculez leurs probabilité.
2. Générez tous les arbres syntaxiques possibles pour la phrase *Jean déclare que Fred ronfle souvent*, *souvent* est un adverbe (ADV), et calculez leurs probabilités selon la grammaire.



Une des analyses possible pour la phrase *Jean déclare que Fred ronfle souvent* attache l'adverbe *souvent* très haut, au niveau du verbe *déclare*, comme dans l'arbre ci-haut, qui décrit la situation où c'est Jean qui déclare souvent quelque chose.

3 Ce type d'attachement n'a jamais été vu dans le corpus. Afin d'éviter ce genre d'attachements, modifiez les étiquettes des non-terminaux dans le corpus. Votre solution devrait introduire de nouveaux symboles non-terminaux qui permettent à la grammaire de capturer la distinction entre les attachements hauts et bas. La grammaire résultante devrait donner une probabilité de 0 aux arbres avec des attachements hauts.

6 Semantics

1. Explain the main components of the Word2Vec algorithm.
2. What are the lexical relations coverd by WordNet? Give examples.