University Djilali Bounaâma of Khemis-Miliana
Department of Artificial Intelligence
Master 2 - AIBD
Final Examination – Natural Language Processing/Correction
January 2025

Consider the following sentences as the base text corpus:

(1) المعلم يكتب درسا

(2) يأكل الولد تفاحة

(3) ذهب الولد إلى المدرسة

(4) رجع الأب من السفر

(5) في المدرسة كتب كثيرة متنوعة

Exercise N°1 (09 pts): We want to perform a HMM-based POS tagging. Use the base text corpus as training data:

1) Calculate the emission and the transtion matrices (Provide a detailed description of the calculation process)
2) Draw the appropriate HMM diagram
3) Calculate the probability of generating the sentence: **كتب الولد كثيرة متنوعة**
4) Deduce the appropriate POS tags of each word of the sentence

Notes:

a. Only consider the three POS tags (فعل، إسم، حرف) to tag the words of the text corpus
b. For the verbs, use their lemma instead of their inflected forms throughout the process.
c. Perform all calculations (matrices, diagrams) using a right-to-left reasoning

Solution:

1) Emission and Transition matrices

a. Assign each word in each sentence by the rigth POS tag (1.25 pt)

إسم	فعل	إسم		
المعلم	**يكتب**	**درسا**		
فعل	إسم	إسم		
يأكل	**الولد**	**تفاحة**		
فعل	إسم	حرف	إسم	
ذهب	**الولد**	**إلى**	**المدرسة**	
فعل	إسم	حرف	إسم	
رجع	**الأب**	**من**	**السفر**	
حرف	إسم	إسم	إسم	إسم
في	**المدرسة**	**كتب**	**كثيرة**	**متنوعة**

b. Count the number of appearences of each word by tag , then divide each tag by the total number of their appearences (02 pts):

	المعلم	كتب	درسا	أكل	الولد	تفاحة	ذهب	إلى	المدرسة	رجع	الأب	من	السفر	في	كثيرة	متنوعة
فعل	0	1/4	0	1/4	0	0	1/4	0	0	1/4	0	0	0	0	0	0
إسم	1/12	1/12	1/12	0	2/12	1/12	0	0	2/12	0	1/12	0	1/12	0	1/12	1/12
حرف	0	0	0	0	0	0	0	1/3	0	0	0	1/3	0	1/3	0	0

c. Emission matrix (0.5 pt):

	المعلم	كتب	درسا	أكل	الولد	تفاحة	ذهب	إلى	المدرسة	رجع	الأب	من	السفر	في	كثيرة	متنوعة
فعل	0	0.25	0	0.25	0	0	0.25	0	0	0.25	0	0	0	0	0	0
إسم	0.083	0.083	0.083	0	0.16	0.083	0	0	0.16	0	0.083	0	0.083	0	0.083	0.083
حرف	0	0	0	0	0	0	0	0.33	0	0	0	0.33	0	0.33	0	0

d. Add the special tags (S) and (E) to each sentence (0.5 pt)

(S) إسم فعل إسم (E)
المعلم يكتب درسا

(S) فعل إسم إسم (E)
يأكل الولد تفاحة

(S) فعل إسم حرف إسم (E)
ذهب الولد إلى المدرسة

(S) فعل إسم حرف إسم (E)
رجع الأب من السفر

(S) حرف إسم إسم إسم إسم (E)
في المدرسة كتب كثيرة متنوعة

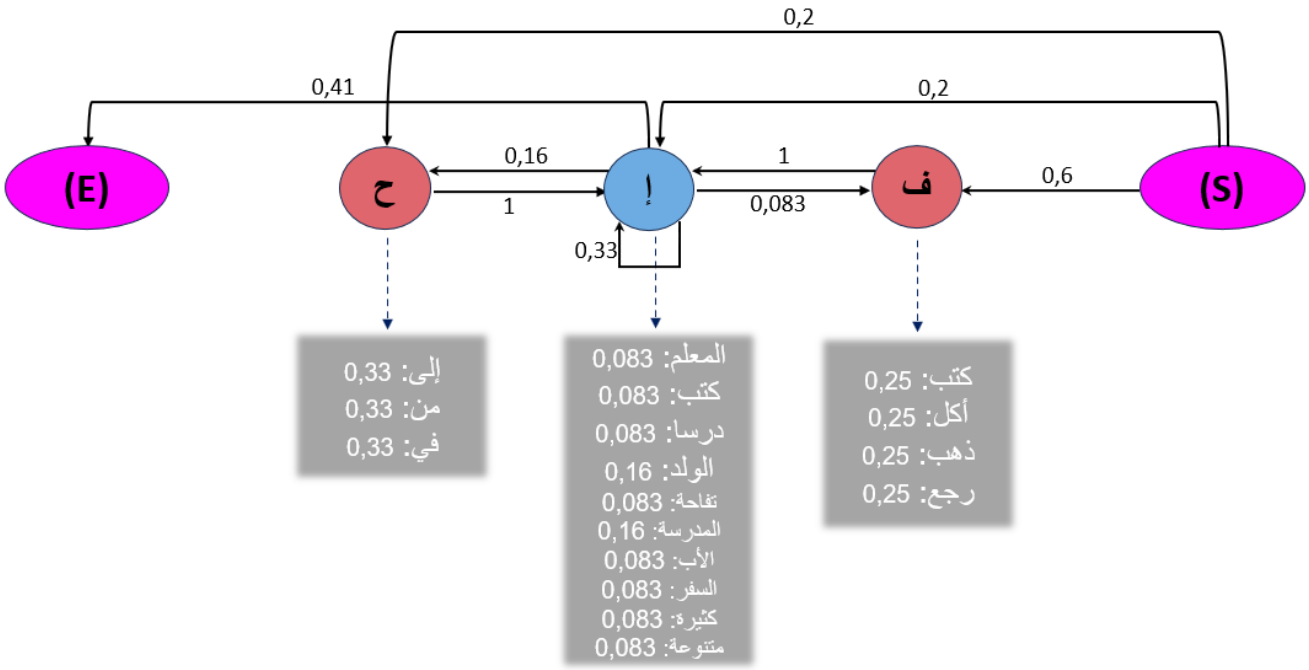
e. Count the co-occurences of each tag (number of times each tag is followed by another), then divide each term in a row by the total number of co-occurences of the tag (01 pt)

	فعل	إسم	حرف	(E)
(S)	3/5	1/5	1/5	0
فعل	0	4/4	0	0
إسم	1/12	4/12	2/12	5/12
حرف	0	3/3	0	0

f. Transition matrix (0.5 pt)

	فعل	إسم	حرف	(E)
(S)	0.6	0.2	0.2	0
فعل	0	1	0	0
إسم	0.083	0.33	0.16	0.41
حرف	0	1	0	0

2) **HMM diagram (1.5 pt)**



3) Calculate the probability of generating the sentence: **كتب الولد كثيرة متنوعة (1.25 pt)**

Path 1: (E) - فعل – إسم – إسم – إسم - (S) ←

Path 2 : (E) - إسم - إسم – إسم – إسم - (S) ←

$P(Path\ 1) = 0.6\ \times 0.25 \times 1 \times 0.16 \times 0.33 \times 0.083 \times 0.33 \times 0.083 \times 0.41 = \ 0.00000738 = 7.38 \times 10^{-6}$

$P(Path\ 2) = 0.2 \times 0.083 \times 0.33 \times 0.16 \times 0.33 \times 0.083 \times 0.33 \times 0.083 \times 0.41 = 0.000000269 = 0.26 \times 10^{-6}$

$P(Sentence) = P(Path\ 1) + P(Path\ 2) = 7.38 \times 10^{-6} + 0.26 \times 10^{-6} = 7.64 \times 10^{-6}$

4) Deduce the appropriate POS tags of each word of the previous sentence **(0.5 pt)**

Path 1 is the **most probable** path, we say then that the appropriate POS tagging (according to our HMM) will be as follows:

فعل	**إسم**	**إسم**	**إسم**
كتب	الولد	كثيرة	متنوعة

Exercise N°2 (04 pts): We want to perform a **3-gram language model** with a **smoothing rate** of **0.2**. Additionally to the base text corpus we include the two following sentences as training data:

(6) يأكل الولد خبزا

(7) ذهب الولد غاضبا

1) Calculate the probability of generating **"خبزا"** given the context **"يأكل الولد"**

2) Calculate the probability of generating **"فرحا"** given the context **"ذهب الولد"**

Provide a detailed description of the calculation process

Solution:

1) **Divide the training text corpus into 3-gram pieces:**

(1) المعلم يكتب درسا
(2) يأكل الولد تفاحة
(3) ذهب الولد إلى المدرسة
(4) رجع الأب من السفر
(5) في المدرسة كتب كثيرة متنوعة
(6) يأكل الولد خبزا
(7) ذهب الولد غاضبا

المعلم يكتب درسا	**يكتب درسا يأكل**	**درسا يأكل الولد**	**يأكل الولد تفاحة**	**الولد تفاحة ذهب**
تفاحة ذهب الولد	**ذهب الولد إلى**	**الولد إلى المدرسة**	**إلى المدرسة رجع**	**المدرسة رجع الأب**
رجع الأب من	**الأب من السفر**	**من السفر في**	**السفر في المدرسة**	**في المدرسة كتب**
المدرسة كتب كثيرة	**كتب كثيرة متنوعة**	**كثيرة متنوعة يأكل**	**متنوعة يأكل الولد**	**يأكل الولد خبزا**
الولد خبزا ذهب	**خبزا ذهب الولد**	**ذهب الولد غاضبا**		

- The text corpus consists of **23** 3-grams **(1.5 pt)**
- **|V| = 19 unique words** (Vocabulary size) **(01 pt)**

المعلم	يكتب	درسا	يأكل	الولد	تفاحة
ذهب	إلى	المدرسة	رجع	الأب	من
السفر	في	كتب	كثيرة	متنوعة	خبزا
غاضبا	19 unique words				

- $\delta = 0.2$ (Smoothing rate)

$$P\left(w_i \middle| w_{i-(n-1)}, \dots w_{i-1}\right) = \frac{\delta + Count(w_{i-(n-1)}, \dots w_{i-1}, w_i)}{\delta(|V|+1) + Count(w_{i-(n-1)}, \dots w_{i-1})}$$ **(0.5 pt)**

2) Calculate the probability of generating **"خبزا"** given the context **"يأكل الولد"**

$$P(\text{"خبزا"} \mid \text{"يأكل الولد"}) = \frac{0.2 + Count(\text{"يأكل الولد خبزا"})}{0.2 \times (19+1) + Count(\text{"يأكل الولد"})} = \frac{0.2+1}{0.2 \times 20 + 2} = \frac{1.2}{6} = 0.2$$ **(0.5 pt)**

3) Calculate the probability of generating **"فرحا"** given the context **"ذهب الولد"**

$$P(\text{"فرحا"} \mid \text{"ذهب الولد"}) = \frac{0.2 + Count(\text{"ذهب الولد فرحا"})}{0.2 \times (19+1) + Count(\text{"ذهب الولد"})} = \frac{0.2+0}{0.2 \times 20 + 2} = \frac{0.2}{6} = 0.033$$ **(0.5 pt)**

Exercise N°3 (07 pts): Focus on sentence (5) in the base text corpus

1) Assign the correct Arabic POS tags to each word in the sentence
2) Define the Arabic syntactic role of each word in the sentence
3) Generate the appropriate Arabic syntax tree for the sentence
4) Cunstruct the Arabic dependecy graph for the sentence

5) We want to perform a TF-IDF encoding. Use the stemmed version of the base text corpus (including sentence 6 and 7, without stopwords) to calculate the importance of each word in sentence (5)

Solution:

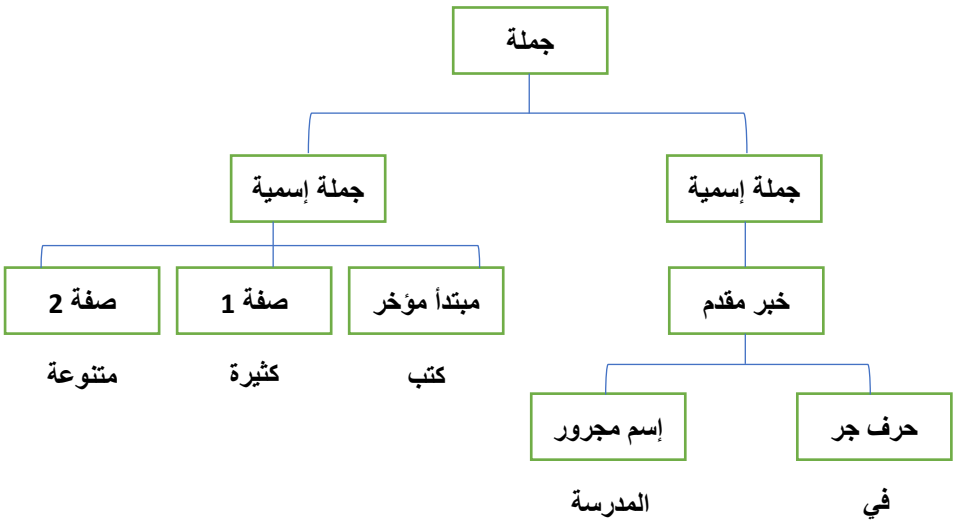
1) Assign the correct Arabic POS tags to each word in the sentence (01 pt)

حرف جر	إسم مجرور	إسم	صفة	صفة
في	**المدرسة**	**كتب**	**كثيرة**	**متنوعة**
Preposition	Noun	Noun	Adjective	Adjective

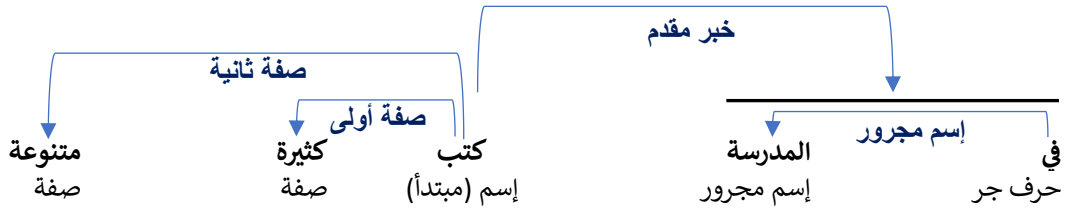
2) Define the Arabic syntactic role of each word in the sentence (01 pt)

خبر مقدم		**مبتدأ مؤخر**	**صفة أولى**	**صفة ثانية**
حرف جر	إسم مجرور	إسم	صفة	صفة
في	**المدرسة**	**كتب**	**كثيرة**	**متنوعة**
Preposition	Noun	Noun	Adjective	Adjective

3) Generate the appropriate Arabic syntax tree for the sentence (1.25 pt)



4) Cunstruct the Arabic dependecy graph for the sentence (1.25 pt)



5) We want to perform a TF-IDF encoding. Use the stemmed version of the base text corpus (including sentence 6 and 7, without stopwords) to calculate the importance of each word in sentence (5) : **(2.5 pts)**

a.Preprocessing: (Stopword removal, Stemming) (0.5 pt)

Word	غاضبا	خبزا	متنوعة	كثيرة	كتب	السفر	الأب	رجع	المدرسة	ذهب	تفاحة	الولد	يأكل	درسا	يكتب	المعلم
Stem	غضب	خبز	نوع	كثر	كتب	سفر	أب	رجع	درس	ذهب	تفح	ولد	أكل	درس	كتب	علم

b. TF-IDF (Sentence 5):

$TF(Term\ Frequency) = \frac{Number\ of\ occurences\ of\ the\ word\ (stem)\ in\ sentence\ 5}{Number\ of\ words\ in\ sentence\ 5}$ (0.25 pt)

$IDF(Inverse\ Document\ Frequency) = \log\left(\frac{Number\ of\ sentences\ in\ the\ corpus}{Number\ of\ sentences\ that\ include\ the\ word\ (stem)}\right)$ (0.25 pt)

TF	متنوعة	كثيرة	كتب	المدرسة	في
(0.5 pt)	1/4=0.25	1/4=0.25	1/4=0.25	1/4=0.25	0

×

IDF	متنوعة	كثيرة	كتب	المدرسة	في
(0.5 pt)	log(7/1)=0.84	log(7/1)=0.84	log(7/2)=0.54	log(7/2)=0.54	0

=

TF×IDF	متنوعة	كثيرة	كتب	المدرسة	في
(0.5 pt)	0.21	0.21	0.13	0.13	0