NGUYEN, Réal **COMP 472 SEC F**
ID#275662673 **DUE 04/12/2018**

MINI-PROJECT 3 REPORT

1. BASIC SETUP

Spanish was chosen as the third language because there are plenty of corpora freely available on the Internet, and it is in the Romance language family, like French. The corpora used for Spanish are the first part of *Don Quijote* and the same excerpt of *The Little Prince* as in French and English. For the rest of this report, French will be abbreviated as FR, English as EN, and Spanish as OT.

1.1 Results

This subsection will discuss and analyse the results of a few sentences. There are two for each language: one correctly classified and one incorrectly classified.

In this report, a sentence is considered incorrectly classified when at least one language model has incorrectly predicted the language of a sentence. For example, the sentence “Birds build nests” (EN) is incorrectly classified, because even if the bigram model classified it as EN, the unigram model classified it as FR. Overall, 15 sentences are incorrectly classified: 5 in the handout’s 10 sentences, and the last 10 sentences, which are supposed to be incorrectly classified. Refer to the “sentences.txt” file in the “input” folder to see all 30 sentences.

Please note that the sentence numbers below do not reflect their positions in “sentences.txt.”

Sentence #	Sentence	Language
S_1	Que la lumiere soit, et la lumiere fut.	FR
S_2	Woody Allen parle.	FR
S_3	The weather in Taumata[…] is lovely.	EN
S_4	Numismatic symmetry should not antagonize economic acme.	EN
S_5	Que alcahuete!	OT
S_6	Voy a buscar el kayak.	OT

Table 1

Each odd numbered sentence is correctly classified, and each even numbered sentence is incorrectly classified.

Let us analyse the correctly classified sentences first. S_1 is in proper, literary FR, so it is no surprise that it is correctly classified, as all the language models in the basic setup are trained exclusively on literature. S_3 on the other hand uses a very long foreign place name but is still correctly classified. That place name is abbreviated to Taumata[…]. It is an abbreviation for Taumatawhakatangihangakoauauotamateaturipukakapikimaungahoronukupokaiwhenuakitanatahu, an 85-letter Maori language place name. The reason why it is correctly classified is that Taumata[…] contains many instances of the unigrams “w” and “k”, which appear often in EN but rarely in FR or OT. The bigram “wh” is also much more frequent in EN than in other languages because of functional words like “who”, “where”, “what”, etc. Compare these 0.5-smoothed probabilities, taken from the unigram and bigram output files:

Language	P(k)	P(w)	P(h|w)
FR	2.91E-04	0.004585144	0.015625
EN	0.008458685	0.023369316	0.267180174
OT	6.24E-07	2.14E-04	0.03125

Table 2

S_5 is comparatively short and is spoken language using slang. Nonetheless, S_5 is accurately classified as OT in both language models. For the unigram model, the approximation for OT is quite close to FR, but the bigram model gives a clear advantage to OT because the bigrams “qu” and “ue” are very common because of functional words like “que” and the name “Quijote”, and the bigram “lc” appears more often in lexical words like “alcanzar” and the name “Dulcinea.” The point where “lc” is read is the point where OT gains a clear advantage over the other language models.

Now, let’s analyse the incorrectly classified sentences. It is obvious why S_2 is classified as EN for both models: Woody Allen is an anglophone name, and that is reflected in the unigrams and bigrams used in the name. The unigram “w” rarely appears in FR and OT compared to EN (refer to Table 1 above), and the bigrams “wo” and “oo” are enough to significantly skew the prediction in EN’s favour. “wo” is naturally rare in FR and OT because the unigram “w” is rare, but the unigram “o” appears relatively often in all languages. However, the bigram “oo” is much more common in EN. Compare these 0.5-smoothed values:

Language	P(o)	P(o|o)
FR	0.053168777	0.001830757
EN	0.072814226	0.043833869
OT	0.098156166	0.006135945

Table 3

S_4 exclusively uses words of Greek origin, save for functional words (here, “should” and “not”). The unigram model incorrectly classifies this sentence as OT, but the bigram model correctly classifies it as EN mostly because of these functional words.

Finally, S_6, like S_2, uses uncommon unigrams in its words, and makes it easy for the language models to incorrectly classify it. The letter “k” practically does not exist in OT, except in a few loanwords. In fact, “k” never appears in OT’s training corpus (refer to Table 1 above). The first usage of “k” in the sentence marks an immediate turning point for both language models. Table 4 below shows the total logarithmic probabilities before the first instance of “k” and after it.

Language	Before "k" (U)	After "k" (U)	Before "k" (B)	After "k" (B)
FR	-17.50966943	-21.046264	-12.9831655	-17.158027
EN	-17.13511495	-19.207812	-13.3666063	-15.566456
OT	-16.12367002	-22.328329	-11.8764195	-16.821169

Table 4

The abbreviations (U) and (B) refer to the unigram model and the bigram model, respectively.

Clearly, the main problem in the basic setup is that uncommon unigrams skew the results towards classes that contain more instances of that same unigram. Using that same logic, it would be beneficial to add diacritics to the models' character set to immediately skew the probabilities towards languages that contain them. For example, "ç" only appears in FR, and "ñ" only appears in OT, so reading those characters in a sentence are sure-fire signs that the sentence is not in EN.

2. EXPERIMENTAL SETUP

There are 4 experiments in this section:

- EX$_1$: reading sentences in old forms of FR, EN, and OT;
- EX$_2$: reading sentences in languages with high degrees of mutual intelligibility with FR, EN, and OT;
- EX$_3$: training the language models to read Haitian Creole and reading sentences in that same language;
- EX$_4$: re-reading the 30 sentences from the basic setup after having trained the models to read Haitian Creole.

2.1 EX$_1$ Results

Each of the sentences in the table below are pulled from the oldest surviving texts of their respective languages: Chanson de Roland (FR); Beowulf (EN); and Cantar de mio Cid (OT). They are all written in poetry rather than prose, so each "sentence" is a concatenation of lines until a period is reached, with "/" representing a line break in the verse.

Language	Sentence
FR	Carles li reis, nostre emper[er]e magnes / Set anz tuz pleins ad estet en Espaigne: / Tresqu'en la mer cunquist la tere altaigne.
EN	Oft Scyld Scefing sceathena threatum / monegum maegthum meodosetla ofteah, egsode eorl, sythan aerest weard / feasceaft funden.
OT	Fablo mio Cid bien e tan mesurado: / <<grado a ti, Senor Padre, que estas en alto!>> / Esto me an buelto mios enemigos malos.>>

Table 5

Surprisingly, both models accurately predicted the actual language for each sentence. This is because the words in the old forms in FR and OT still look relatively similar to their modern counterparts. Old EN words cannot be understood by a speaker of modern EN. Old EN uses a different character set, and some characters had to be replaced by a modern phonetic equivalent (e.g. sceaþena → sceathena). However, many unigrams and bigrams used in modern EN are still in old EN. For example, the unigrams "t", "h", and the bigram "th" are much more common in EN than in other languages. Refer to the table below:

Language	P(t)	P(h)	P(h|t)
FR	0.068239	0.007841	0.007429
EN	0.092422	0.066028	0.373938
OT	0.038333	0.011787	8.14E-05

Table 6

2.2 EX$_2$ Results

The sentences in this subsection are written in languages very similar to FR, EN, and OT. The languages used are Norman (FR); Scots (EN); and Asturian (OT). Each language has 3 sentences: 1 in a formal context, 1 in a poetic context, and 1 being the translation of the same sentence taken from *The Little Prince*.

Sentence #	Sentence	Language
S_1	S'lon le recensement d'2001 y'avait 2,674 personnes tchi palent l'Jerriais (3.2% d'la population).	FR
S_2	Le jouo va s' dejuqui, le cyil est d'exces rouoge.	FR
S_3	J'i apprins chu morce de ta vie le quatriyime jouo, des petra-jaquet, quaund tu m'as announchi : J'aime byin les couchis de sole !	FR
S_4	We wad like tae mak shair that as mony fowk as possible can get tae speir aboot the Scots Pairlament.	EN
S_5	I amna fou' sae muckle as tired - deid dune.	EN
S_6	I lairnt this new detail on the mornin o the fowert day, whan ye said tae me: I'm awfy fond o doungangs.	EN
S_7	L'oxetu del alcuerdu ye la promocion, l'espardimientu y la normalizacion llinguistica de la llingua asturiana na institucion academica.	OT
S_8	Santa Olaya fo l'abeya / que de Merida ensamo.	OT
S_9	Dime cuenta d'ello al cuartu dia cuandu me dixisti pela mananina: _Comu me presten les atapecies!	OT

Table 7

S_3, S_5, and S_9 are incorrectly classified in the unigram model, with S_3 being classified as OT, and S_5 and S_9 as FR. Like in the basic setup, the main problem is uncommon unigrams. For S_3, it is the unigram “y”, for S_9, the unigram “x”. For S_5, no single unigram is at fault: it was always a close race between FR and EN, and some unigrams towards the end happened to be more common in FR.

2.3 EX$_3$ Results

I have chosen Haitian Creole as an experiment language because it is a language mainly made from mixing words from FR, EN, OT, and other languages, but has significantly different orthography. From this point on, Haitian Creole will be abbreviated to EX.

It is interesting to note that compared to FR, EN, and OT, which are trained exclusively on literature, it is quite difficult to find entire works of EX-language literature online. Because of this, the training corpus is a mixture of poems, online magazine articles, blogs, Wikipedia articles, textbook materials, song lyrics, and religious texts. Because of this, loanwords are much more common.

For EX_3, the sentences to be read by the language models are exclusively in EX. There are 13 sentences: the 10 default sentences included in the handout but translated into EX, and 3 more sentences taken from an online article, the Bible, and the same sentence from *The Little Prince* as in EX_2.

Sentence #	Sentence
S_1	Ki sa ki pral vini ekonomi an Japone ane ca?
S_2	Li te mande l si li te yon elev nan lekol sa.
S_3	Mwen OK.
S_4	Zwazo bati nich.
S_5	Zwazo vole.
S_6	Mwen rayi AI.
S_7	Woody Allen pale.
S_8	Eske abite la?
S_9	Fraz sa a se nan angle.
S_{10}	Mwen renmen AI.
S_{11}	AYITI: NOUVO GOUVENMAN AN AP KONFWONTE DEFI SOU KESYON DWA MOUN
S_{12}	Se pou limye fet. Epi limye te fet.
S_{13}	M aprann sa nan maten katriyem jou a, le w di m : Mwen renmen we soley kouche.

Table 8

The only sentence to be incorrectly classified is S_9, which was classified as OT in the unigram model. The last three unigrams of the sentence “g”, “l”, and “e” gave a slight advantage to OT.

2.4 EX_4 Results

EX_4 is like EX_3, but instead of reading EX-language sentences, the language models read the 30 sentences from the basic setup. These sentences contain no words in EX. This experiment is done to see how the addition of a language affects the predictions in the basic setup.

7 sentences were incorrectly classified as EX in this experiment.

Sentence #	Sentence	Language
S_1	Woody Allen parle.	FR
S_2	Sea la luz, y fue la luz.	OT
S_3	The weather in Taumata[…] is lovely.	EN
S_4	Tabarnak!	FR
S_5	Voy a buscar el kayak.	OT
S_6	Dale botjia!	OT
S_7	I'm OK.	EN

Table 9

Sentences S_1 to S_4 are incorrectly classified in the unigram model; S_6 is incorrectly classified in the bigram model; and S_7 in both.

As expected, the explanation behind this is the same as in the basic setup and all the other experiments: there are unigrams that are much more common in EX than in other languages. It is no coincidence that 4 of the 7 incorrectly classified sentences contain one or more instances of the unigram “k”. “w” and “z” also appear more frequently, but for S_1, it was unigrams like “n” and “p” that misled the unigram model into thinking that it is in EX.

Altogether, the experiments suffer from the same issues as the main setup, because only the inputs of the models changed. However, these experiments were done out of curiosity to see how the language models would react to different lexicons and syntaxes given the same limitations as the basic setup, rather than to see how the models could be improved.

3. REFERENCES

General resources
https://pteo.paranoiaworks.mobi/diacriticsremover/
https://translate.google.ca/
http://www.petit-prince.at/

Old languages resources
https://en.wikipedia.org/wiki/Cantar_de_Mio_Cid
https://www.hs-augsburg.de/~harsch/gallica/Chronologie/11siecle/Roland/rol_ch01.html
http://ebeowulf.uky.edu/ebeo4.0/start.html

Norman resources
https://en.wikipedia.org/wiki/Norman_language
https://fr.wikipedia.org/wiki/Jersiais
http://members.societe-jersiaise.org/geraint/jerriais.html
http://magene.pagesperso-orange.fr/allain.html

Scots resources
https://en.wikipedia.org/wiki/Scots_language
https://en.wikipedia.org/wiki/A_Drunk_Man_Looks_at_the_Thistle
http://www.parliament.scot/help/79056.aspx

Asturian resources
https://en.wikipedia.org/wiki/Asturleonese_language
https://en.wikipedia.org/wiki/Asturian_language
https://ast.wikipedia.org/wiki/Dominiu_lling%C3%BC%C3%ADsticu_%C3%A1stur
http://www.academiadelallingua.com/
https://web.archive.org/web/20100924153252/http://www.asterionxxi.com.ar/numero7/asturias.htm

Haitian Creole resources
https://www.wikimizik.com
http://woymagazine.com/cr
https://medium.com/@gaetantguevara/
https://haitiglobal.com/
https://www.madansarafilm.com/
https://www.bookrix.com/_ebook-roselaure-beton-ayiti-cheri/
http://www.fouyebible.com/
https://ht.wikipedia.org/wiki/Tousen_Louv%C3%A8ti
http://www.potomitan.info/ayiti/ayiti12.php
http://www.potomitan.info/ayiti/baron.php
http://www.potomitan.info/vedrine/web.php
https://web.archive.org/web/20080517032011/http://hometown.aol.com/mit2haiti/WojeDezi-Alfabetizasyon.htm
http://www.tanbou.com/