|  |  |
| --- | --- |
| **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 Quixote* and the same excerpt of *The Little Prince* as in French and English. For the sake of brevity, French will be abbreviated as FR, English as EN, and Spanish as OT for the rest of the report.

**1.1 Results**

This subsection will discuss and analyse the results of a subset of all the 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 the unigram model classified it as FR and the bigram model classified it as EN. 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 as per the handout. Refer to the “sentences.txt” file in the “input” folder to see all 30 sentences.

Any reference to sentences in this subsection will now use the form Sn where n is the sentence number. Please note that n does not reflect the number of the sentence in “sentences.txt.”

|  |  |  |
| --- | --- | --- |
| **Sentence #** | **Sentence** | **Language** |
| S1 | Que la lumiere soit, et la lumiere fut. | FR |
| S2 | Woody Allen parle. | FR |
| S3 | The weather in Taumata[…] is lovely. | EN |
| S4 | Numismatic symmetry should not antagonize economic acme. | EN |
| S5 | Que alcahuete! | OT |
| S6 | Voy a buscar el kayak. | OT |

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

Let us analyse the correctly classified sentences first. S1 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. S3 on the other hand uses a very long foreign place name but is still correctly classified. In this report, 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 letters ‘w’ and ‘k’, which appear often in EN but rarely in FR or OT. The bigram (w|h) is also much more frequent in EN than in other languages because of functional words like “who”, “where”, “what”, etc. Compare these smoothed probabilities, taken from the language model 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 |

S5 is comparatively short and is spoken language using slang. Nonetheless, S5 is accurately classified as OT in both the unigram and bigram 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 (q|u) and (u|e) are much more common in functional words, and the bigram (l|c) appears more often in lexical words. The point where (l|c) is read is the point where OT has a clear advantage over the other language models and keeps that advantage. *(insert comparison table)*

Now, let’s analyse the incorrectly classified sentences. It is obvious why S2 is classified as EN for both the unigram and bigram models: Woody Allen is a rather anglophone-sounding name, and that is reflected in the letters and bigrams used in the name. The letter ‘w’ rarely appears in FR and OT compared to EN. For the bigram model, the bigrams (w|o) and (o|o) are enough to significantly skew the prediction in EN’s favour. *(insert comparison table)*

S4 is unique among other EN sentences, in that it exclusively uses words of Greek origin, save for functional words (in this case, “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, S6, like S2, uses uncommon letters in its words, and makes it easy for the language models to incorrectly classify it. Although all the language models use the same character set, the 26 letters of the Latin alphabet, the letter ‘k’ practically does not exist in the Spanish language, except for a few loanwords (like kayak). In fact, ‘k’ never appears in the corpus on which the OT models are trained. Compare these smoothed probabilities, from the unigram outputs: *(insert comparison table)*

Both the unigram and bigram models identify the sentence as EN. Compare these logarithmic probabilities: *(insert table)*

**2. EXPERIMENTAL SETUP**

There are 4 experiments (EX) in this section:

EX1: reading old forms of FR, EN, and OT;

EX2: reading sentences in languages with high degrees of mutual intelligibility with FR, EN, and OT;

EX3: training the unigram and bigram models on a Haitian Creole corpus and reading sentences written in that same language;

EX4: reading the 30 sentences from the basic setup after having trained the models to read Haitian Creole.

**2.1 EX1 Results**

Each of the sentences in the table below are pulled from the oldest surviving texts from 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 the forward slash 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.>> |

Surprisingly, all language models accurately predicted the actual language for each sentence. This is because the words in the old forms still look relatively similar to their modern counterparts. For EN, the old form’s words cannot be understood by a modern speaker. Old EN uses a different character set, and some letters had to be replaced by modern equivalents (e.g. sceaþena -> sceathena). However, many letters and bigrams used in modern EN are still in old EN, for example the letters ‘t’, ‘h’, and the bigram (t|h) are much more common in EN than in other languages, and the language models see exactly that in the old EN sentence, making them predict that it is in EN.

**2.2 EX2 Results**

The sentences in this subsection are written in languages very close 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 by Antoine de Saint-Exupéry.

|  |  |  |
| --- | --- | --- |
| **Sentence #** | **Sentence** | **Language** |
| S1 | S'lon le recensement d'2001 y'avait 2,674 personnes tchi palent l'Jerriais (3.2% d'la population). | FR |
| S2 | Le jouo va s' dejuqui, le cyil est d'exces rouoge. | FR |
| S3 | 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 |
| S4 | We wad like tae mak shair that as mony fowk as possible can get tae speir aboot the Scots Pairlament. | EN |
| S5 | I amna fou' sae muckle as tired - deid dune. | EN |
| S6 | I lairnt this new detail on the mornin o the fowert day, whan ye said tae me: I'm awfy fond o doungangs. | EN |
| S7 | L'oxetu del alcuerdu ye la promocion, l'espardimientu y la normalizacion llinguistica de la llingua asturiana na institucion academica. | OT |
| S8 | Santa Olaya fo l'abeya / que de Merida ensamo. | OT |
| S9 | Dime cuenta d'ello al cuartu dia cuandu me dixisti pela mananina: \_Comu me presten les atapecies! | OT |

The only incorrectly classified sentences are S3, S5, and S9. All of them are only incorrectly classified in the unigram model, with S3 being classified as S3, S5 as FR, and S9 as FR. All the predicted approximations are relatively close to their actual approximations. *(insert log prob table)*

The only exception is S9, where the usage of the letter ‘x’ significantly skewed the prediction to FR. *(insert table)*

**2.3 EX3 Results**

EX3 and EX4 add a new language: Haitian Creole. I have chosen Haitian Creole as an experiment language because it is a language born out of mixing words from FR, EN, OT, and other languages, but has a significantly different orthography. From this point on, Haitian Creole will be abbreviated to EX. A part of EX3 is feeding a training corpus in EX to the language models, and as such, EX4 is dependent on EX3.

It is interesting to note that compared to FR, EN, and OT, which are trained exclusively on literary works, it is quite difficult to find entire works of EX-language literature online. Because of this, the corpus is an amalgamation of poems, online magazine articles, blogs, Wikipedia articles, textbook material, song lyrics, and religious texts.

For EX3, 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 Little Prince.

|  |  |
| --- | --- |
| **Sentence #** | **Sentence** |
| S1 | Ki sa ki pral vini ekonomi an Japone ane ca? |
| S2 | Li te mande l si li te yon elev nan lekol sa. |
| S3 | Mwen OK. |
| S4 | Zwazo bati nich. |
| S5 | Zwazo vole. |
| S6 | Mwen rayi AI. |
| S7 | Woody Allen pale. |
| S8 | Eske abite la? |
| S9 | Fraz sa a se nan angle. |
| S10 | Mwen renmen AI. |
| S11 | AYITI: NOUVO GOUVENMAN AN AP KONFWONTE DEFI SOU KESYON DWA MOUN |
| S12 | Se pou limye fet. Epi limye te fet. |
| S13 | M aprann sa nan maten katriyem jou a, le w di m : Mwen renmen we soley kouche. |

The only sentence to be incorrectly classified is S9, which was classified as OT in the unigram model. The last three letters of the sentence (‘g’, ‘l’, ‘e’) gave a slight advantage to OT, and in the end OT and EX have a very small difference in their approximations. *(insert table)*

**2.4 EX4 Results**

EX4 is a bit of a continuation of EX3 – it uses the same four languages, but this time, the sentences read by them are the 30 sentences read in the basic setup, without any sentences written in EX. This experiment is done to see how the addition of a language would affect the language models in the basic setup.

As it turns out, 7 sentences were incorrectly classified as EX in this experiment.

|  |  |  |
| --- | --- | --- |
| **Sentence #** | **Sentence** | **Language** |
| S1 | Woody Allen parle. | FR |
| S2 | Sea la luz, y fue la luz. | OT |
| S3 | The weather in Taumata[…] is lovely. | EN |
| S4 | Tabarnak! | FR |
| S5 | Voy a buscar el kayak. | OT |
| S6 | Dale botjia! | OT |
| S7 | I'm OK. | EN |

Sentences S1 to S4 are incorrectly classified in the unigram model; sentence S6 incorrectly classified in the bigram model; and S7 in both.

As expected, the explanation behind this is the same as in the basic setup and all the other experiments: there are letters that are much more prevalent 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 letter ‘k’. The letters ‘w’ and ‘z’ are also appear more frequently, but in the case of S1, it was letters like ‘n’ and ‘p’ that gave EX the advantage.

**3. REFERENCES**