# NLP Homework 4 (MVA 2017/2018)

Rémi Lespinet

I did my experiments on the Google translation system and used [2] to comment. The sections 1 and 2 focus on french ↔ english tranlsations, and the section 3 describe translations to other languages (especially rare languages).

To summarize quickly, Google translation system use wordpieces as input (subwords), it consists of an encoder with 8 stack of LSTM (only the bottom one is bidirectional) with residual connections, and a decoder, also with 8 stacks of LSTM and residual connections. The encoder and decoder are connected through an attention mechanism (AM) which weights , at a given time $i$ of the decoding, AM tries to select the words of the input sentence that are the most useful for this step (by computing a weight for each input word using a feed forward network).

## 1 Note on successful cases

I was impressed about the performances of Google translation system. For simple constructions for english ↔ french, it had the right translations, even with non trivial vocabulary. I also found long sentences to work relatively well (I tested it on some sentences from Proust (about 250 words), and some long english sentences found on this site (about 120) words).

## 2 Presentation of the failure cases

1. **Idiomatic expressions**

   For the most simple cases, the translation is correct (for example “Il pleut des cordes”), but the translation fails for for less common expressions, such as :

> I hate going to the dentist, but I'll just have **to bite the bullet**
>
> Je déteste aller chez le dentiste, mais je vais devoir **mordre la balle**

> Il voulait résoudre le problème P=NP mais il est rapidement **tombé sur un os**
>
> He wanted to solve the problem P = NP but he quickly **fell on a bone**.

   These expressions are hard to translate because the meaning is unrelated to the words taken separately, it completly changes when a group of word is present. This is a many to many dependency that is probably rare in the corpus and hence hard to represent. Moreover it's also possible that in some cases, the meaning is really this one (it is possible to really fall on a bone).

2. **New words and rare words**

   Regardless of the size of the training corpus, there will always be words that do not occur, and even words that are created, the first examples uses a rare french word, and the second the word ghost has a verb which is a (relatively) new usage of the word.

> Plusieurs passages nous ont conservé l'allégorie de l'**ophiomaque**.
>
> Several passages have preserved the allegory of the **ophiomaque**.

> I've texted him a couple of times since then and **he's been ghosting me** ...
>
> Je lui ai envoyé un texto à plusieurs reprises depuis et il **m'a fait des fantômes** ...

   Thanks to the *Wordpiece Model* (WPM) (they feed the bi-LSTM encoder with sub-words instead of words), the system effectively produces likely translations for rare and OOV words, for example if I create the word “ambirateur” in french, it is translated to “ambirator” in english, which is nice, as rare words can often be translated in an automatic way. It allows to translate words that – I think – are not in the corpus such as “flexitarien” (in french) to “flexitarian” in english (which is the correct translation)

3. **Uncommon/Difficult constructions**

   The construction of the sentence makes it hard to understand. The sentence “The rat the cat killed, ate the malt” is correctly translated. Adding another information (about the cat being chased by the dog) makes a highly uncommon structure which is probably not represented that much in the corpus.

   The rat the cat the dog chased killed ate the malt
   Le rat le chat que le chien a chassé a mangé le malt

4. **Syntactic ambiguity**

   Some words have multiples meaning, and understanding the sentence can only be done by catching its underlying meaning.

   Elle est sortie en pleurant **du café**
   She came out **crying coffee**

   The blind man picked up the hammer and **saw**
   L’homme aveugle a ramassé le marteau et **a vu**

5. **Too much unknown**

   The following sentence is left untouched, my guess is that the word smith is not very used in the corpus as a verb, and is probably also present as a proper noun (NNP), therefore there is not much the decoder knows for sure in this sentence and translating is is probably risky.

   Will Will Smith smith?
   Will Will Smith Smith?

   The phrase is funny because words are repeated twice, but we could replace Will Smith by *John Lenon* for example and the problem stays the same. If we replace the verb “smith” by another verb such as “cook”, or if we add an adverb such as in the sentence “Will Will Smith **really** smith ?”, then the sentence is correcly translated.

6. **Context information**

   A mouse appeared. **It** looked hungry.
   Une souris est apparue. **Il** avait l’air affamé.

   We see that the translation fails to translate “it” correcltly. If we replace the point by a comma, making the whole input one sentence, the translation is correct. This is probably because the translation system as exposed in [2] processes sentences separately which causes information loss accross sentences boundaries.

## 3 Translation to rare languages

Here’s a french sentence that I’ve traduced in all possible languages (and then traduced back). The output for 2 “rare” languages is shown below (appendix A shows the output for many languages) :

Je n’ai jamais pu refuser quoi que ce soit à une brune aux yeux marrons
Je n’aurais jamais bloqué une caméra avec un café. (Tadjik)
Je ne pourrais jamais laisser la moustache verte. (Persan)

As we can see this is not perfect. I think the reason is the following : First Google does not train a model for each pair of languages : there is 104 languages so this would mean 5356 models to train, each model requiring to have an annotated corpus, this is impossible in practice. Instead Google uses Zero-Shot translation [1], which basically allow to produce any language from any language with only one model. This is done by adding an artificial token before the sentence to specify the target language. The system find the input language by itself, which explains why we can input sentences with multiple languages such as :

> *I don't know* si je *take* mon costume bleu ou *grey* ...
>
> Je ne sais pas si je prends mon costume bleu ou gris ... (Francais)
>
> I do not know if I take my blue or gray suit ... (Anglais)

Obviously the corpus models for the rare languages are not as big as the most used languages in Google translate, which probably explains these poor results. For the Swahili language, we can even observe this output :

> Je n'ai jamais pu refuser quoi que ce soit à une brune aux yeux marrons
>
> Je ne peux jamais interdire quoi que ce soit pour les couleurs de couleur marron de couleur marron, de couleur, de c ouleur, de couleur, de couleur, de couleur, de couleur, de couleur, de couleur, de couleur, de couleur, de couleur, de couleurs, de couleurs colorées. (Swahili)

Which corresponds to the decoder LSTM failing to output the end of sentence token !

## A Translation of a french sentence to all other languages

I have translated the following sentence to all available languages and translated back in french and observed the languages that gave the “most wrong” results :

Je n’ai jamais pu refuser quoi que ce soit à une brune aux yeux marrons

Je ne veux rien écrire sur les yeux bruns. (Amharique)

Je n’ai jamais rien nié de brun avec des yeux bruns. (Basque)

Je ne pourrais jamais rien brune aux yeux bruns refusent. (Bielorusse)

J’ai les yeux bruns et les cheveux noirs ne pourrait jamais rien refuser. (Birman)

Je ne renierais jamais une brune aux yeux bruns. (Cebuano)

Avec les yeux bruns j’ai refusé n’importe quoi à Breadtee. (Chichewa)

Je ne pourrais jamais rien avoir à refuser des yeux nuptiaux à brunir. (Espéranto)

Je n’ai jamais rien su d’une brune aux oeufs bruns. (Frison)

Je ne pouvais rien dire à un fantôme aux yeux bruns. (Gaelique)

Je ne peux pas te refuser comme les yeux des ombres. (Georgien)

Je ne peux rien laisser marron avec du brun. (Haoussa)

Je ne pouvais rien faire pour blanc avec des yeux bruns (Hindi)

Je n’ai jamais rien refusé à une brune avec un œil jaune. (Hmong)

Je ne pourrais jamais rien acheter avec une brune brune. (Hongrois)

Je ne peux pas nier tout ce qui est blanc et blanc. (Igbo)

Je ne pourrais jamais détourner mes yeux bruns avec les yeux bruns. (Kazakh)

J’ai les yeux bruns et brune n’abandonnèrent. (Khmer)

Je n’ai jamais rien à voir avec les yeux rouges de mes frères. (Kurde)

Je ne peux rien refuser aux aveugles avec les yeux bruns (Laotien)

Je ne peux rien dire avec des yeux noirs marron foncé. (Malayalam)

Je ne peux tout transformer en brune aux yeux bruns. (Maltais)

Je ne peux rien interdire à une prostituée aux yeux bruns (Marathi)

Je ne pouvais rien faire avec les yeux noirs. (Népalais)

Je ne pourrais jamais rien refuser aux yeux noirs. (Ouzbek)

Je n’ai jamais rien eu à faire avec un œil méchant. (Pachtô)

Je ne peux jamais nier un œil noir vilain. (Panjabi)

Je ne pourrais jamais laisser la moustache verte. (Persan)

Je ne peux rien nier dans une couleur bleue et brune. (Samoan)

Je ne peux rien nier de l’église aux yeux noirs (Sindhi)

Je ne peux jamais le nier avec des yeux chauds (Somali)

Je ne peux jamais interdire quoi que ce soit pour les couleurs de couleur marron de couleur marron, de couleur, de c ouleur, de couleur, de couleur, de couleur, de couleur, de couleur, de couleur, de couleur, de couleur, de couleur, de couleurs, de couleurs colorées. (Swahili)

Je n’aurais jamais bloqué une caméra avec un café. (Tadjik)

Je ne peux pas rejeter quiconque a les cheveux noirs aux yeux noirs. (Telugu)

Je ne peux rien écrire à un brun avec des couleurs brunes. (Yorouba)

## References

[1] M. Johnson, M. Schuster, Q. V. Le, M. Krikun, Y. Wu, Z. Chen, N. Thorat, F. Viégas, M. Wattenberg, G. Corrado, et al. Google's multilingual neural machine translation system: enabling zero-shot translation. *arXiv preprint arXiv:1611.04558*, 2016.

[2] Y. Wu, M. Schuster, Z. Chen, Q. V. Le, M. Norouzi, W. Macherey, M. Krikun, Y. Cao, Q. Gao, K. Macherey, et al. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*, 2016.