Master MVA, 2017/2018

Algorithms for Speech and Natural Language Processing - TD 5

Sébastien Ohleyer

The goal of this assignment is to understand the current state of the art in machine translation. We used Google Translate as machine translation system.

1 Syntactic constructions

First we investigate which syntactic constructions are correctly or uncorrectly translated. As an example, we try to translate in French the assignment question "Which syntactic constructions are correctly translated? Which ones are not? Can you guess why?". Google translate as follows: "Quelles constructions syntaxiques sont incorrectes traduites? Lesquels ne sont pas? Pouvez-vous deviner pourquoi?". The first question is not correctly translate, the translated sentence has the opposite meaning of the original sentence. It is quite suprising because intuitively the system should translate word by word and translate "correctly" by "correctement" and not by "incorrectes". Here the system gives an adjective instead of an adverb and of the opposite meaning of the original word. Note that users can improve Google Translate by validating each sentences one by one through different propositions. The second proposition for this first question is "Quelles constructions syntaxiques sont traduites correctement?", which is the correct translation.

There is also a mistake in the second question translation. However in this case, words are correctly translated one by one but the system does not succeed in adding the object "le" which is absent in English. The last question (quite easy) is correctly translated. Nevertheless we note that the system choose "vous" instead of "tu" for the translation which is probably a default choice of Google Translate.

Concerning idioms, we tried two examples which seems to raise the same error. "Vendre la peau de l'ours avant de l'avoir tué" is translated in English by "To sell the skin of the bear before killing it" when the corresponding English idiom is "Don't count your chickens before they hatch". Identicaly, the French phrase "Se taper le cul par terre" means to laugh uproariously. According to Google Translate, though, it means "Ass banging on the floor". And when we translate "Don't count your chickens before they hatch" in French, we obtain "Ne comptez pas vos poulets avant qu'ils éclosent"... It seems that Google Translate cannot deal with idioms and choose to return the direct translation.

An overall remark can be made following these tries. The Google Translate system appears to mostly translate directly, with a word by word approach. When it has to take into account language particularities (e.g idioms or syntax), it fails to return the appropriate translation. We will now investigate the influence of two parameters: the language pair and the context.

2 Influences

2.1 Language pair

From the Google Translate website itself we can read that Google Translate's quality varies with the language pair. Indeed, the tool may produce a usable gist for English-Spanish or English-French translations, as billions of words for each language can easily be found on the web.

However, for rarer language such as Danish and Romanian, or Turkish and Thai, correspondences are harder to find. Besides, that is why Google proposes to its users to improve the translator by taking part of the Google Translate Community. On this platform you can check and validate some translation or even propose yours.

As an example, we tried to translate some words from French to Igbo (a language proposed by Google Translate) which is "an ethnic group native to the present-day south-central and southeastern Nigeria" according to Wikipedia. It is quite difficult for us to judge the translation - we do not really speak fluently this language - but we probe some clear errors of the system. We simply ask the translation of "Bonjour", and Google Translate returns "Hello". Here, the system does not have the translation of "Bonjour" in memory - after some research it appears that the correct translation is "kedu!" or "!bola ch!" with "!" representing a strange letter like a "ị" upside down.

If we change French-Igbo to English-Igbo and we ask for the translation of "Hello", we obtain "Nnoo" – with "o" denoting a letter close to a "o" but with a kind of a point below it. In this case, the system returns something, but the translation is apparently wrong according to our research. Something weird is that when Google Translate does not have a translation in memory in French-Igbo for example, it does not try to translate in English – which can be done with a good accuracy – and then to Igbo. This process could avoid some lack of translation in many cases.

2.2 Context

Investigating the examples proposed by the assignment, we ask Google Translate to translate "A mouse appeared. It looked hungry." to French. It returns "Une souris est apparue. Il avait l'air affamé.". On this example, the system makes a mistake on the pronoun "It". Translating this pronoun is hard for the system because in English we do not use different pronoun depending on the gender of the subject. Here the system must consider the context, and this context is given by the first sentence. Unfortunately, Google Translate seems to clearly separates the two sentence when it translates and forget about the first sentence context.

With "Il aime bien le mouton. Surtout les côtelettes de mouton." translated to English, we obtain "He likes the sheep. Especially the mutton chops.". Here again, the system translates "mouton" by "sheep" instead of "mutton" because it does not take into account the context of these sentences. The word "mouton" refers to the meat and not to the animal so "mutton" is a more appropriate translation than "sheep".

With Google Translate the meaning can be *lost in translation* because there is no way to incorporate context. The complexity of the text, as well as any context which cannot be interpreted without a true knowledge of the language bring a lot of translation errors. Google Translate often translates directly and gives nonsensical literal results while a correct translation would require well-established online glossaries or back translation methods.