P. Thomas Barthelemy
Jireh Tan

1. THE CHOSEN F-LANGUAGE AND GENERAL OBSERVATIONS

We have chosen the canonical F-language in machine translation, French. We chose French because the both of us know its grammatical structure reasonably well, and therefore it is easier for us to implement rules after we word-for-word translation.

French is a Romance language. Some of its *grammatical features* are shared with other Romance languages:

a. Two grammatical genders (masculin, féminin).
b. Tenses formed from auxiliaries.
c. The use of pronominal verbs. In general, pronominal verbs in French are either reflexive, reciprocal or idiomatic. As the term suggests, idiomatic pronominal verbs are quite idiosyncratic and the pronoun may modify the sense of the verb in unexpected ways.

More *quirky grammatical features* include:

a. Declarative word order is SVO, except that pronominal objects precede verbs ("Tu parles français"; "tu le parles").
b. Verb-subject inversion to denote questions ("parles-tu français?"), although this is not always required ("tu parles français?").
c. Existence of simple past, principally used in poetic diction.
d. Split negation, typically of the form "ne (VERB) pas". Other forms of negation include: "ne (VERB) point", "ne (VERB) plus".

Some *vocabulary features*:

a. The majority of French words come from Vulgar Latin. The Roman conquest in England, as well as the French influence over the English court, explain the large overlap of its vocabulary with that of English.

Some phonological features:

a. The key feature to note is elision; this is otherwise unproblematic except that certain forms of elision are carried over into the orthography, for example, "j'ai": "je ai"; "l'homme qu'il a vu": "le homme que il a vu".

We proceed to discuss the impact that these features have on the translation:

PREPROCESSING: Elision meant that we needed to be more sensitive to tokenization, so we couldn't simply remove every single punctuation mark. However, splitting on punctuation marks could lead to confusion too. Take, for example, "je l'ai vu". In our word-by-word translation model, there is no straightforward way of distinguishing between "l'" as an article ("the"), and "l'" as a pronoun ("he" or "she" or "it). Thankfully our training set did not have very many of these ambiguities; it would be less easy in this deterministic setup to deal with such problems.

TRANSLATING: We were thankful that most of these French words had a pretty straightforward counterpart in English, due to the cultural (and geographic!) proximity of English to French. It would have been far harder translating a different language, say, Indonesian, in which many words would escape straightforward translation because many words would be dependent on some basis of cultural understanding. This meant that we didn't have to worry too much about fidelity.

We were also glad that French and English are hot languages, so we do not have to deal with inferring agents, as they are explicitly marked. However, a problem that comes up as a result is that we often have to deal with pronouns, and there is no straightforward way of dealing with this, without further context: more specifically, in English, one refers to any entity (physical or not) as "it", as long as this entity is not a person. In French, however, any entity is gendered, and this affects the pronoun that is used. For example, in English one says "This is my passport. It is blue." In French, one says "C'est mon passeport. Il est bleu."

Another thing that proves problematic is the existence of pronominal verbs. It is not immediately obvious how to deal with them, since the pronoun "se" is typically translated as "himself", and in some cases it makes sense (particularly for reflexive verbs). However, for other cases it is far less obvious what we should do with the pronoun. For example, it would take a more sophisticated language model to tell that "il se peut" ought to be translated as "it is possible" or "it appears that" or "it could be", rather than "he itself can".

2. THE TEST DATA AND RESULTING TRANSLATIONS

1. **Original:** L'ambassade de Belgique à Washington a l'honneur de vous informer de la mise en place d'une nouvelle mesure européenne en matière de délivrance de passeports.

   Direct: the embassy of belgium at washington has the honour of you to tell of the putting in place of a new measure european in matter of release of passports .

   Cleaned: the *belgium embassy* at washington has the honour of you to tell of the putting in place of a new *european measure* in *release passports matter* .

   Google: The Belgian embassy in Washington has the honor to inform you of the introduction of a new measure for European passports.

2. **Original:** Depuis le 15 février 2012, les passeports délivrés par cette Ambassade contiennent les empreintes digitales de leur demandeur et ce dès l'âge de 12 ans.

   Direct: since the 15 february 2012 , the passports delivered by that embassy contain the prints digital of their enquirer and this as soon as the age of 12 years .

Cleaned: *since february* 15 2012 , the passports delivered by that embassy contain the *digital prints* of their enquirer and this as soon as the age of 12 years .

Google: Since February 15, 2012, passports issued by the Embassy contain the fingerprints of their applicant and at the age of 12 years.

3. **Original:** Depuis cette date, la comparution personnelle de tout demandeur de passeports inscrits ou non dans les registres consulaires de cette Ambassade est obligatoire.

   Direct: since that date , the appearance personal of all enquirer of passports registered persons or no in the register consulars of that embassy is compulsory .

   Cleaned: since that date , the *personal appearance* of all *passports enquirer* registered persons or no in the register consulars of that embassy is compulsory .

   Google: Since that time, the personal appearance of an applicant passports or not registered in the records of the Embassy Consular is mandatory.

4. **Original:** Cette Ambassade n'est plus en mesure d"accepter les demandes introduites par courrier postal.

   Direct: that embassy not is more in measure of to take the applications introduced by mail mail .

   Cleaned: that embassy is *no longer* in *measure to* take the applications introduced by *mail* .

   Google: The Embassy is no longer able to accept applications by mail.

5. **Original:** Veuillez également noter qu'à partir de l"âge de 6 ans, les enfants doivent se présenter à l'Ambassade pour la prise de photo et la signature du passeport.

   Direct: hope for equally to write down that at to leave of the age of 6 years , the children must self put at the embassy for the taking of photo and the signature from passport .

   Cleaned: *please* equally to write down that *from* the age of 6 years , the children must self put at the embassy for the taking of photo and the signature from passport .

   Google: Please also note that from the age of 6, children must come to the Embassy for taking photo and signature of the passport.

6. **Original:** La même procédure est d'application auprès des Consulats Généraux d"Atlanta, Los Angeles et New York depuis le 1er octobre 2012.

   Direct: the same procedure is of application near some consulates general of atlanta , los angeles and new york since the 1st october 2012 .

   Cleaned: the same procedure is of application near some *general atlanta consulates* , los angeles and new york *since october 1st* , 2012 .

   Google: The same procedure applies to Consulates General in Atlanta, Los Angeles and New York since 1 October 2012.

7. **Original:** Pour de plus amples informations, veuillez consulter les liens suivants

Direct: for of more loose news , hope for to consult the affiliation following

Cleaned: for of more loose news , *please* to consult the *following affiliation*

Google: For more information, please visit the following links

8. **Original:** La section passeport précise que ce procédé d"identification n'a aucune autre incidence sur l'introduction, le traitement, et la délivrance des passeports.

   Direct: the section passport exactly that this behavior of identification not has no other repercussion on the introduction , the treatment , and the release some passports .

   Cleaned: the section passport exactly that this *identification behavior has* no other repercussion on the introduction , the treatment , and the release some passports .

   Google: The section specifies that the passport identification process has no further impact on the introduction, processing, and issuing passports.

9. **Original:** Cette nouvelle mesure est imposée par un règlement Européen du 13.12.2004 et s'inscrit dans la lutte contre la fraude d'identité.

   Direct: that new measure is required by a regulation european from 13 . 12 . 2004 and self registered in the struggle against the fraud of identity .

   Cleaned: that new measure is required by a *european regulation* from 13 . 12 . 2004 and self registered in the struggle against the *identity fraud* .

   Google: This new measure is imposed by European Regulation of 13.12.2004, and is in the fight against identity fraud.

10. **Original:** Dans le futur, elle permettra aux détenteurs de passeports d"Etats membres de l'Union Européenne de bonne foi, de bénéficier de contrôles de frontière plus aisés.

    Direct: in the future , she will allow to the sharers of passports of states members of the union european of good faith , of to receive of checks of border more comfortable .

    Cleaned: in the future , she will allow to the *passports states sharers* members of the *european union* of good faith *, to* receive of *border checks* more comfortable .

    Google: In the future, it will allow passport holders Member States of the European Union in good faith to benefit from border controls easier.

3. THE RULES USED TO REORDER

First, as instructed, we ran the output through a POS tagger. Then we applied the 10 rules described below.

1. **NN JJ → JJ NN**

   Whenever we encountered a noun followed by an adjective, we switched it such that the adjective comes before the noun. This is perhaps the most obvious difference between English and French (in that most new students of French will learn it within the first two weeks of class) and that is why we had this rule first. One example of the result of this transformation comes from the first sentence: "measure european" is reordered to make "european measure."

2. **NN VBG → VBG NN**
   Implemented to invert gerund constructions: in French probably more grammatical to say something along the lines of "l'homme gagnant", in English "the winning man". Rarer than the previous case, but some cases persisted in our training set, so we inverted gerund constructions.
3. **NN1 "OF" NN2 → NN2 NN1**
   Implemented to catch compound nouns in French. For example, the phrase "fraud d'identité" should translate as "identity fraud" instead of "fraud of identity". This is simply the way the French compound nouns, whereas in English we can simply concatenate them.
4. **"HOPE FOR" → "PLEASE"**
   Implemented to dramatically improve fluency. The French "veuillez" is literally translated as "hope for", but this translation is inadequate, because it is used as a marker of politeness, usually closer to "please" instead.
5. **DELETING ARTICLES FROM DATES**
   French represents dates as "le 12ème janvier", i.e. they append a definite article at the front of dates. English never does this, so we remove the article.
6. **DELETING REPEATED WORDS**
   Sometimes languages will translate word-for-word extremely badly, because noun phrases in one language can be translated into a single noun in another language. For example, "courrier" translates as "mail" and "postal" translates as "mail", but "courrier postal" should translate as "mail", not "mail mail". If we see two same words in a row we should be able to delete one of them without much loss of fidelity, and probably gain some fluency.
7. **"NOT" VB "NO" → VB "NO" [DELETE THE "NOT"]**
   This has to do with French having a split negation, so verbs are negated with as "ne" + VB + PARTICLE, with the particle denoting the richness of negation. So whenever we see a verb surrounded by negation, we need to make a choice. In this case, it comes from French "ne... aucune" which means "no".
8. **"NOT" VB "MORE" → VB "NO LONGER" [DELETE THE "NOT", "MORE", INSERT "NO LONGER"]**
   Refer to rule 7 for split negation. In this case, the split negation arose from French "ne... plus" which means "no longer".
9. **"OF TO" → "TO" [DELETE "OF"]**
   French often requires the preposition "de" in front of an infinitive in order to be grammatical, e.g. "permettre ... de bénéficier". This does not hold in English, one would never

say "allow ... of to benefit", one would simply say "allow to benefit".

10. **"AT TO LEAVE OF" → "FROM"**

    In our translation model, "à partir de" is translated as "at" + "to leave" + "of", this is idiomatic in French but hardly comprehensible in English. One is better off translating this as "from".

4. ERROR ANALYSIS

We observe that our translation is far from ideal, though it is in the general ballpark of comprehensibility. The reader of our translations could conceivably understand them, given some context, and might think that the sentences were translated by a native speaker of French attempting to do an English assignment. Some errors are reminiscent of those made by French English-as-a-second-language (ESL) speakers, so from a heuristic perspective, our translation is not terrible.

However, our translation clearly does not achieve full fidelity or fluency. In particular, we noticed the following:

1. *Homograph resolution*: language is full of ambiguities, and one of these ambiguities pertains specifically to orthography. A word that is spelled in one manner can have several different meanings. For example, in French, “des” can mean “some” or “of the” (pl.), compare the senses “il y a des chats” (“there are some cats”) and “la délivrance des passeports” (“the release of the passports”). Given that our translation rule was to pick the most frequent English sense of the word, our translation was not able achieve great fidelity. Clearly, a statistical model which takes into account more features would be able to disambiguate better.
2. *Incorrect translation of noun phrases*: The translation model is word-for-word, so instead of translating entire noun phrases from French into noun phrases in English, we translate at the level of the words. This leads to some bizarre results. Take, for example, the french “mise en place”. Our model translates this as “putting in place”, which is not an unfaithful translation, but it is a less fluent one. A faithful and fluent translation of this would be “introduction”. We observe that one instance of a non-fluent noun-phrase translation would not seriously affect comprehension, but consistent non-fluent noun-phrase translations are difficult for the reader to navigate.
3. *Incorrect handling of pronominal verbs*: This was especially problematic because the most frequent representation of these pronominal pronoun of these verbs is the reflexive pronoun in English, so our translation contained “self”s. There was no general deterministic way to deal with these verbs, and so we had to leave them alone. French pronominal verbs fall into three broad categories, reflexive (subject does the action to the subject it/him/herself), reciprocal (subjects do the action to each other), and idiomatic. In general, fidelity was not

compromised, because the pronominal verbs we encountered (e.g. "se présenter") were not idiomatic. However, idiomatic pronominal verbs would be far harder to translate in our model, because the pronoun interacts with the verb in a way that is not predictable.

5. COMPARISON BETWEEN GOOGLE TRANSLATION AND OUR TRANSLATION

The Google translation provides a much more fluent translation than our system. Most notably, context is used to translate words with ambiguous direct translations. For example, "liens" is properly translated as "links", while our translation used the more corporate interpretation, "affiliation". Further, there is much better understanding in the Google translation of word part of speech than our system, which tagged words individually. A useful rule to have applied would move the pronoun immediately preceding an infinitive after the infinitive, as in the direct translation "... washington has the honour of you to tell...". However, in parsing the words, our system interpreted "tell" as a noun, and thus would not recognize these last two words as an infinitive verb form.

Google is also adept at identifying idiomatic French expressions, like "mise en place", (previously discussed) and "empreintes digitales" meaning "fingerprints", not "digital prints". Naturally, this was an expected disadvantage of our system as the assignment stipulated word-for-word translation. Otherwise, we may have used a dictionary that attempted to maximize the number of words translated at once.

Also notable is the difficulty that both systems had in translating a sequence of nouns joined by "of", as in Sentence 10 ("... détenteurs **de** passeports **d'**Etats membres **de** l'Union Européenne **de** bonne foi..."). Our system translated in order, switching the order of first sharers and passports, and next of sharers and states, giving the awkward phrase "... passports states sharers members..." Google translate had a different strategy–"détenteurs de passeports" is translated directly to "passport holders" and the remaining text is translated as "Member States of the European Union"; this probably relates to Google employing some probabilistic parser to obtain the structure of the French sentence before they translate into English. This allows them to translate entire verb and noun phrases, producing much more intelligible sentences, especially when presented with longer sentences with more than one clause. However, we observed that Google's translation did not understand that the relationship between the states and the holders exists *through* the passport; this relationship is lost in translation. A better, though still somewhat literal, translation would be "holders of passports of European Union states", which could could be shortened to "European Union passport holders", as we can build compound nouns simply by concatenating them in English.

6. CONTRIBUTIONS

P. Thomas Barthelemy coded the assignment and contributed to the writing and data analysis. Jireh Tan did most of the writing and the data analysis.