

Apprentissage de métrique pour la traduction automatique

Guillaume Wisniewski

guillaume.wisniewski@linguist.univ-paris-diderot.fr

novembre 2019

Attention : il s'agit d'une version préliminaire du sujet amenée à évoluer en fonction de vos « découvertes ». N'hésitez pas à consulter régulièrement sur la page du cours pour obtenir la dernière version du sujet ou des informations supplémentaires.

1 Objectif du projet

L'objectif du projet est de développer une métrique pour l'évaluation de la traduction automatique. Cette métrique devra, étant donné deux phrases dans une même langue (typiquement une hypothèse de traduction et une traduction de référence) renvoyer un score entre 0 et 1, ce score étant d'autant plus grand que la similarité entre les deux phrases.

Les performances de cette métrique devront être évaluées en mesurant la corrélation entre ses scores et les scores donnés par les annotateurs humains. Nous utiliserons comme mesure de la corrélation le ρ de Spearman.

Les paramètres de votre métrique et ses performances seront à évaluer sur le corpus disponibles sur le site du cours¹. Ce fichier, au format `json` contient une liste de dictionnaires, chaque dictionnaire décrivant un exemple. Un exemple est décrit par :

- une phrase source ;
- une traduction de référence ;
- une hypothèse de traduction ;
- une évaluation humaine de cette traduction (par un score entre 0 et 1).

Les données peuvent être chargées simplement à l'aide des instructions python

```
import json

corpus = json.load(open("en2fr_manual_evaluation.json"))
```

1. <https://gw17.github.io/>

- ① Comment pouvez-vous utiliser les données disponibles pour développer et tester votre métrique ?
- ② Déterminer la probabilité de chaque hypothèse de traduction selon un modèle de langue estimée sur le corpus Europarl² Est-ce que cette probabilité est un bon indicateur de la probabilité d'une traduction ?
- ③ Déterminer la probabilité de l'alignement IBM 1 entre chaque hypothèse de traduction et la phrase source correspondante. Est-ce que cette probabilité est un bon indicateur de la qualité d'une traduction ?
- ④ Mesurer la corrélation entre les différentes métriques que vous connaissez et les scores humains. Qu'en déduisez-vous ?
- ⑤ Quelles informations (c.-à-d. caractéristiques) pouvez-vous utiliser pour déterminer si deux phrases sont sémantiquement équivalentes ? Vous pourrez, par exemple, vous interroger sur la possibilité d'utiliser des plongements lexicaux contextualisés pré-entraînés (typiquement CamemBERT pour le français³).
- ⑥ Comment combiner ces différentes sources d'information ? Est-ce que les performances que vous obtenez vous paraissent satisfaisantes ?
- ⑦ A-t-on besoin d'utiliser des informations bilingues (c.-à-d. dépendant de la source) pour évaluer la qualité de la traduction ?

2 Travail à effectuer

Le projet est à rendre pour le **10 janvier 2020 à 8h00** en envoyant votre rapport au format pdf ainsi qu'une archive contenant l'ensemble de votre code à Guillaume Wisniewski (guillaume.wisniewski@linguist.univ-paris-diderot.fr).

2. <http://statmt.org/europarl/>

3. <https://camembert-model.fr/>