Impact de la ≪ translationese ≫ sur l'évaluation de 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 Contexte

La manière naturelle d'évaluer un système de traduction automatique, par exemple français-anglais, consiste à collecter un corpus parallèle [1] de tests contenant des phrases françaises et leur traduction en anglais. Il est alors possible de comparer les traductions prédites à ces traductions de référence en utilisant une métrique automatique comme le score BLEU [1] ou la distance d'édition [2]. Même si elle est moins fiable qu'une évaluation manuelle dans laquelle des annotateurs évoluent directement la qualité d'une traduction en lui attribuant une ≪ note ≫ (cette méthode, remise au goût du jour par [3], est appelée DA pour *Direct Assessment*), cette approche est aujourd'hui au cœur de l'apprentissage et de l'évaluation des systèmes de traduction automatique.

Pour réduire le coût de l'annotation, les corpus parallèles sont généralement utilisés dans les deux direction. Ainsi un corpus de test français-anglais (c.-à-d. contenant des phrases françaises traduites en anglais) sera utilisé à la fois pour évaluer des systèmes traduisant du français en anglais mais également pour évaluer des systèmes traduisant de l'anglais en français. Toutefois, cette manière de procéder introduit un biais dans l'évaluation : [4] a montré que les textes traduits diffèrent des textes écrits par des locuteurs natifs selon plusieurs axes (absence de répétition, grammaire plus conventionnelle, ...), un phénomène appelé *translationese* ; plusieurs travaux, comme [5], suggèrent que, pour un système de traduction automatique, il est plus facile de traduire des phrases issues d'une traduction que des phrases écrites par un locuteur natif.

1. De manière plus générale, on appelle ≪ corpus parallèle ≫ un corpus contenant des phrases associée à leur traduction.

L'objectif de ce projet est *i)* de vérifier si cette hypothèse est vraie (*spoiler alert* : c'est bien le cas) et *ii)* mettre en évidence les différences entre les phrases écrites par un locuteur natif et les phrases issues d'une traduction qui pourraient expliquer pourquoi ces dernières sont plus faciles à traduire.

Ces questions ont aujourd'hui un intérêt majeur pour la communauté TAL : en 2018, une équipe de chercheurs de Microsoft a prétendu avoir développé un système de traduction aussi bon que des traducteurs humains — ils ont, selon leurs termes [6], obtenu «*human parity* ». Cependant plusieurs travaux [5, 7] remettent en cause ces conclusions en suggérant que la qualité du système de Microsoft était sur-évaluée à cause de ce phénomène de *translationese*.

2 Données

Pour répondre aux différentes questions de ce projet, nous utiliserons les données collectées dans le cadre de la campagne d'évaluation WMT'16 [2]. Celles-ci sont composées, pour 6 paires de langues (donc 12 directions de traduction) :

- d'une hypothèse de traduction prédite par un systèmes de traduction automatique (différents systèmes ont été utilisés pour générer ces hypothèses) ;
- d'une évaluation de la qualité de cette hypothèse par un score DA ;
- d'une traduction de référence correspondante ;
- de la phrase source ayant été traduite ;
- de langue dans laquelle la phrase source a été initialement écrite (qui permet donc de déterminer s'il s'agit d'un exemple *direct* (c.-à-d. que la phrase source a été écrite par locuteur natif puis traduite) ou *indirect* (c.-à-d. que la phrase source est issue d'une traduction et que c'est la référence qui a été écrite par un locuteur natif)..

Ce corpus est téléchargeable à l'url `https://gw17.github.io/da_newstest2016.json.xz`. Les données sont stockées sous forme d'une liste de dictionnaires et peuvent être chargées dans une structure python grâce aux instructions :

```
import json

corpus = json.load(open("da_newstest2016.json"))
# corpus est une liste de dictionnaires
# les clés du dictionnaires permettent d'identifier la nature de
# chaque information
```

3 Prise en main du corpus

Les questions de cette section ont pour objectif à la fois de vérifier que vous ayez bien compris la manière dont les données sont structurées et que vous vous soyez *familiarisé*

2. `http://www.statmt.org/wmt16/`

avec celles-ci. Dans les expériences de cette section nous ne distingueront pas les exemples directs des exemples indirects.

① De combien d'exemples dispose-t-on pour chaque direction de traduction ?

② La qualité moyenne des systèmes, évaluée par le score DA, est-elle meilleure lorsque l'on traduit depuis ou vers l'anglais ? Est-ce que cette différence est statistiquement significative ? Ce résultat est-il conforme à votre intuition ?

③ Déterminer, pour chaque direction de traduction, le score BLEU et le distance d'édition moyenne obtenus [3]. Laquelle de ces deux métriques offrent la meilleure approximation de l'évaluation manuelle.

④ Quel est l'impact de la longueur de la phrase source sur le score DA ?

4 Impact de la translationese

L'objectif de cette série d'expériences est de montrer que le phénomène de *translationese*, tel que défini dans la section 1, a un impact significatif sur l'évaluation de la traduction.

⑤ Comparer la distribution des scores des évaluations humaines pour les deux types de phrases sources. Que pouvez-vous en conclure sur l'impact de ce facteur sur l'évaluation de la traduction automatique ?

⑥ De manière similaire, comparer la distribution des scores BLEU et des distance d'édition. Que pouvez-vous en conclure ?

5 Différences entre exemples directs et indirects

Nous allons maintenant essayer de mettre en évidence les différences entre les deux types de phrases sources qui pourraient expliquer pourquoi les phrases issues d'une traduction sont plus faciles à traduire.

⑦ À l'aide de la bibliothèque `Spacy` [4], faite l'analyse en dépendances des différentes phrases du corpus.

⑧ Imaginer différents indicateurs permettant de comparer des phrases et pouvant permettre de distinguer une phrase écrite par un locuteur natif d'une phrase issue d'une traduction (p.ex. la longueur moyenne des phrases, la profondeur de l'arbre des dépendances, leur probabilité selon un modèle de langue appris sur le corpus d'apprentissage [5]...). Observer la valeur de vos indicateurs pour les différents types de phrase.

3. Le score BLEU se calcule normalement au niveau d'un corpus ou d'un ensemble de phrases. Il est intéressant pour cette question de comparer à la fois les scores estimés au niveau du corpus et ceux estimés au niveau des phrases.

4. `https://spacy.io/`

5. Les données sont accessibles à l'URL `http://www.statmt.org/wmt17/translation-task.html`.

⑨ Pour vérifier la qualité des indicateurs de la question précédente, nous considérons le problème de classification suivant : une paire (phrase source, référence) donnée est-elle un exemple directe ou un exemple indirecte ? Comment peut-on apprendre et évaluer un tel classifieur ? Quelle performance obtient-il ?

6 Travail à effectuer

Le projet est à effectuer en binôme.

Vous devrez rendre un rapport d'une quinzaine de pages décrivant le travail effectué et les résultats obtenus. Vous veillerez, en particulier, à faire apparaitre :

— les questions et problématiques que vous abordez ;

— les difficultés rencontrées ;

— les solutions apportées et les choix qui ont été faits ;

— les résultats obtenus et leur interprétation.

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 et des ressources utilisées à Guillaume Wisniewski (`guillaume.wisniewski@linguist.univ-paris-diderot.fr`) et Vincent Segonne (`vs_teaching@hotmail.com`).

7 Références

[1] K. Papineni, S. Roukos, T. Ward et W.-J. Zhu. "Bleu : a Method for Automatic Evaluation of Machine Translation". In : *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. Philadelphia, Pennsylvania, USA : Association for Computational Linguistics, juil. 2002, p. 311–318.

[2] M. Snover, N. Madnani, B. Dorr et R. Schwartz. "Fluency, Adequacy, or HTER ? Exploring Different Human Judgments with a Tunable MT Metric". In : *Proceedings of the Fourth Workshop on Statistical Machine Translation*. Athens, Greece : Association for Computational Linguistics, mar. 2009, p. 259–268.

[3] Y. Graham, T. Baldwin, A. Moffat et J. Zobel. "Can machine translation systems be evaluated by the crowd alone". In : *Natural Language Engineering* 23.1 (2017), p. 3–30.

[4] "Corpus linguistics and translation studies : Implications and applications". In : *Text and Technology : In Honour of John Sinclair*. Netherlands : John Benjamins Publishing Company, 1993.

[5] Y. Graham, B. Haddow et P. Koehn. "Translationese in Machine Translation Evaluation". In : *CoRR* abs/1906.09833 (2019).

[6] H. Hassan, A. Aue, C. Chen, V. Chowdhary, J. Clark, C. Federmann, X. Huang, M. Junczys-Dowmunt, W. Lewis, M. Li, S. Liu, T. Liu, R. Luo, A. Menezes, T. Qin, F. Seide, X. Tan, F. Tian, L. Wu, S. Wu, Y. Xia, D. Zhang, Z. Zhang et M. Zhou. "Achieving Human Parity on Automatic Chinese to English News Translation". In : *CoRR* abs/1803.05567 (2018).

[7] A. Toral, S. Castilho, K. Hu et A. Way. "Attaining the Unattainable ? Reassessing Claims of Human Parity in Neural Machine Translation". In : *CoRR* abs/1808.10432 (2018).