Algorithms for Speech and NLP TD 5 : Etat de l'art de la traduction automatique

Pirashanth RATNAMOGAN
MVA ENS Paris-Saclay
pirashanth.ratnamogan@ens-paris-saclay.fr

1. But du TD

Le but de ce TD est d'explorer et de comprendre l'état de l'art de la traduction automatique. La plupart des gens ont pu observer une amélioration significative des outils de traduction automatique ces dernières années : on est passé des modèles de statistique axée sur les syntagmes à la traduction automatique neuronale. Il est donc intéressant de comprendre quelles sont les limites aujourd'hui atteinte.

2. Quelles constructions syntaxiques sont correctement traduites ? Lesquels ne sont pas ? Pouvez-vous deviner pourquoi ?

J'utilise pour cette question Google Translate en faisant une traduction du français vers l'anglais. Les phrases simples et extrêmement courantes sont parfaitement traduites ('Il est allé à l'école'). Les phrases avec des subordonnées relatives et circonstancielles sont aussi bien traduites. Le discours rapporté est également bien traduit. Là ou le traducteur a du mal c'est sur les syntaxes très particulières au Français par exemple les phrases infinitifs 'Pourquoi refuser une telle offre ?' traduit en 'Why refuse such an offer ?' de l'anglais vers le francais.

3. Qu'en est-il des questions, des phrases longues et / ou complexes, des phrases à la première personne, des idiomes (kick the bucket vs. casser sa pipe), entités nommées (London)...

Les phrases longues sont en général relativement bien traduite. Bien sûr lors des traductions entre anglais et français on peut perdre de l'information. Il est assez drôle de voir que beaucoup d'idiomes sont parfaitement traduites du français vers l'anglais mais sont mal traduites dans l'autre sens. "Il pleut des cordes" est bien traduite en "Raining cats and dogs" oubliant juste les termes "it's" mais lorsque l'on veut faire la traduction dans l'autre sens à partir de cette expression on obtient "Pleut des chats et des chiens". Il est incroyable que les transformations français $\rightarrow$ anglais $\rightarrow$ français ne sont pas parfaite. On observe exactement le même phénomène avec l'expression "kick the bucket" qui est bien traduite en "casser sa pipe" mais inversement l'expression "casser sa pipe" est traduite littéralement en "break his pipe".
Concernant les entités nommées on observe que "London" est bien traduit en Londres dans la plupart des cas. Le problème que l'on peut avoir c'est que si par exemple nous avons une personne nommée "John London" on ne veut pas traduire London en Londres, on voit que ce cas est bien géré par l'outil de traduction de Google. Si par contre j'utilise un nom moins commun par exemple "Pirashanth London" on voit que l'outil de traduction traduit directement en "Pirashanth Londres". On voit donc que la traduction des entités nommées est également géré automatiquement. D'après [3] les entités nommées et mots en dehors du vocabulaire sont gérés par l'utilisation de "wordpieces" (on apprend la découpe grâce à un modèle) une représentation intermédiaire des mots entre le caractère et les mots entiers qui permet de limiter la taille du vocabulaire.

4. Quelle est l'influence de la paire de langues ? Cf. Anglais $\leftarrow\rightarrow$ Français vs. Anglais $\leftarrow\rightarrow$ Chinois ou Anglais $\leftarrow\rightarrow$ (a rare language)

Ayant des origines Sri Lankaises, j'ai pu demander à mes parents d'évaluer la traduction Anglais vers Tamoul et Français vers Tamoul. Ils étaient impréssionnés par la qualité de la traduction produite. Ce qui a été le plus perturbant est surement la traduction des idiomes Français vers le Tamoul, l'expression "il pleut des cordes" est traduites étrangement en "il pleut des chats et des chiens" ce qui n'a aucun sens. Ce que j'ai pu comprendre c'est que pour faire sa traduction Google passe systématiquement par l'anglais comme intermédiaire ce qui provoque cette mauvaise traduction assez farfelue. Le traducteur fonctionne plutot très bien pour la traduction du Tamoul surement parce qu'à la fois le Tamoul a des expressions proche de l'anglais, et parce que le Sri Lanka et l'Inde sont d'anciennes colonies britanniques et l'on doit trouver une grande quantités de documents pour l'entrainement des modèles utilisés.
Ne parlant pas chinois je peux difficilement juger mais en lisant [3] on voit que le chinois est la langue pour laquelle l'outil de traduction a le plus de mal, c'est aussi la langue pour laquelle la traduction par les humains est la moins

bien notés. C'est une langue difficile à traduire, ces difficultés se répercutent sur les modèles de traduction automatique.

5. Quelle est l'influence du contexte. Par exemple, comment sont traduits les discours suivants ? (Anglais $\longleftrightarrow$ Français) : - (en$\longleftrightarrow$ fr) A mouse appeared. It looked hungry. - (fr$\longleftrightarrow$ en) Il aime bien le mouton. Surtout les côtelettes de mouton.

La phrase "A mouse appeared. It looked hungry." est traduite en "Une souris est apparue. Il avait l'air affamé.", le contexte est mal géré par l'outil. Si on transforme le point en virgule la phrase est alors bien traduite puisque la deuxième partie de la phrase est bien traduite au féminin. Ce problème est un problème récurrent dans la traduction automatique, (différents noms qui ont une classe grammaticale) évoqué dans [1]. Ce problème semble bien géré par d'autres traducteurs que celui de Google par exemple SYSTRAN. L'article [1] évoque également les problèmes dans la traduction de l'anglais au Français lorsque le genre est porté par le nom (nom ambigenre par exemple les métiers danseur/danseuse) en français et en anglais on perd cette information ainsi à la phrase "The danser starts the show" on peut associer plusieurs traductions et il est impossible de trouver la bonne à partir de cette phrase seule. On peut remarquer que des cas difficiles évoquées dans [1] qui date de 2009 sont bien traités par Google Translate par exemple "We went to see the baby and we admired him. Her name was Mary.". On peut remarquer que l'algorithme a parfois des difficultés à utiliser le contexte, par exemple la phrase "je sale le plat" est mal traduite en "I dirty the dish". Pour la phrase "Il aime bien le mouton. Surtout les côtelettes de mouton." elle est bien traduite en "He likes the sheep. Especially the mutton chops." Dans le cadre de la traduction automatique neuronale le contexte est pris en compte par l'utilisation de réseaux récurrent particuliers qui permettent de prendre en compte le contexte.

L'architecture classique des NMT est basée sur un encoder et un decoder. L'encoder en entrée est constitué d'un LSTM (un type de réseau récurent qui a été crée pour capturer les dépendences à long termes). Il vient encoder l'entrée que l'on donne à notre algorithme et un decoder qui est également constitué d'un réseau de type LSTM qui doit renvoyer la phrase dans la langue cible. Dans [3] on utilise en plus une architecture de type "Bidirectional LSTM" pour prendre en compte les dépendances dans les deux sens et les "Attention Networks" [2] qui permettent au mot cible à une position i fixé de s'appuyer sur toutes les sorties du LSTM avec un certain poids (permet de cibler la zone du texte source qui permet de traduire le mot cible, c'est ce qui permet de gérer les longues phrases). Comme on peut le comprendre l'architecture du réseau est vraiment pensé pour que le contexte soit pris en compte lors de la traduction.

6. Conclusion

L'utilisation de la traduction automatique neuronale ainsi que de jeux d'entrainement de plus en plus grand a vraiment permis d'obtenir d'excellents résultats dans la tâche de traduction automatique. Comme on peut le voir à partir des résultats obtenues par Google dans ses différents articles la traduction automatique neuronale a de loin dépassé les modèles statistiques autrefois utilisés (basé sur le score BLEU) et approche en performance la performance humaine lors des tests sur des humains. J'ai moi-même été impressioné lorsque je faisais quelques tests pour ce TD.

Références

[1] C.-B. Antonia. Les erreurs dans la traduction automatique du genre dans les couples français-anglais et anglais-français : typologie, causes linguistiques et solutions. *Revue française de linguistique appliquée, 2009/1 (Vol. XIV), p. 93-108*, 2009.

[2] D. Bahdanau, K. Cho, and Y. Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv :1409.0473*, 2014.

[3] 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.