03.04.2017 – 08:00 Uhr

Traduction automatique: au-delà du phrase par phrase

Bern (ots) -

En travaillant phrase par phrase, les algorithmes de traduction omettent une grande partie du contexte et font des erreurs. Un projet soutenu par le FNS a développé de nouvelles techniques afin qu'ils considèrent mieux l'ensemble du texte.

Des scientifiques financés par le Fonds national suisse (FNS) ont ouvert une nouvelle voie pour améliorer les outils de traduction automatique, comme le célèbre Google Translate, qui traite quotidiennement quelque 100 milliards de mots. Les informaticiens et linguistes ont été les premiers à montrer qu'il était possible d'améliorer les systèmes de traduction en forçant l'intelligence artificielle à dépasser le simple «phrase à phrase», et à croiser des informations contenues ailleurs dans le texte, une démarche qui fait aujourd'hui l'objet de travaux dans le monde entier. Les scientifiques dévoilent leurs derniers résultats (*) le 3 avril 2017 lors d'une conférence de l'Association for Computational Linguistics à Valence (Espagne).

Traduire sans comprendre

«Les systèmes de traduction automatique ne comprennent pas vraiment le sens des textes, explique Andrei Popescu-Belis, responsable du projet ainsi que du Natural Language Processing Group à l'Institut de recherche Idiap, situé à Martigny (VS). Ils opèrent un rendu d'une langue vers une autre, en suivant des règles statistiques. Et surtout, ils travaillent phrase par phrase. Or une phrase isolée ne comporte souvent pas assez d'informations sur le contexte pour pouvoir être traduite correctement. Les systèmes devraient pouvoir prendre en compte des données situées ailleurs dans le texte.»

Pour démontrer leur approche, les chercheurs se sont notamment penchés sur la question des pronoms - des mots tels que «lui» ou «celle-ci», qui se substituent à d'autres éléments du texte. Souvent, ces derniers se trouvent hors de la phrase à traduire, d'où le nombre important d'erreurs commises par les systèmes automatiques. Andrei Popescu-Belis donne un exemple simple, mais qui trompe aisément les systèmes les plus sophistiqués: «Ma tante a acheté une excellente voiture. Elle n'est pas très jolie.» En anglais, Google Translate la traduit en «My aunt bought an excellent car. But she is not very pretty.» L'outil a traduit «elle» par «she». Comme ce pronom est réservé aux personnes de genre féminin, le lecteur anglophone lira que c'est «ma tante» qui «n'est pas très jolie».

Le piège de la statistique

Le système est induit en erreur, car il sait que le qualificatif «pas très jolie» s'applique plus souvent à des personnes qu'à des objets. Si on le substitue par «rouillée» ou «en panne», plus fréquemment appliqués aux objets, le pronom aura plus de chances d'être correctement traduit par «it».

Pour obtenir un résultat pertinent, le traducteur automatique aurait dû considérer les informations contenues dans la première phrase. C'est dans les grandes lignes ce que fait le système mis au point par les chercheurs de l'Idiap en collaboration avec les Départements de linguistique des universités de Genève et d'Utrecht (Pays-Bas) ainsi que l'Institut de linguistique computationnelle de l'Université de Zurich.

Les chercheurs utilisent essentiellement des outils d'apprentissage automatique (ou «machine learning»). A chaque essai, ils introduisent ou retirent des centaines de paramètres, que les algorithmes ajustent, jusqu'à constater une amélioration. «Dans les grandes lignes, nous indiquons au système le nombre de phrases précédentes qu'il doit analyser, comment il doit les analyser, puis nous procédons à des tests en conditions réelles.»

Google recrute au sein du projet

Les résultats sont encourageants, selon Andrei Popescu-Belis. Dans des couples de langues comme français-anglais ou espagnol-anglais, les pronoms induisent en erreur les traducteurs automatiques dans environ la moitié des cas. «En forçant le système à considérer des informations externes à la phrase, nous sommes parvenus à réduire le taux d'erreur à 30%», note le chercheur.

L'enjeu de ces travaux va bien au-delà de la seule question des pronoms: la cohérence des temps verbaux, le choix de la terminologie ou encore les niveaux de politesse constituent autant de problématiques qui dépendent largement du texte dans son ensemble, plutôt que d'une phrase prise isolément.

Les techniques développées par Andrei Popescu-Belis et ses collègues ne sont pas encore mûres pour des outils grand public, mais elles intéressent les acteurs du domaine. «Ce sont nos travaux qui ont fait connaître la nécessité de dépasser la traduction automatique phrase à phrase. Mais surtout, trois jeunes chercheurs impliqués dans le projet travaillent maintenant dans ce domaine chez Google Zurich. Cela montre bien l'intérêt suscité par notre approche.»

LINKS

Projet "MODERN" dans la base de données des projets du FNS http://p3.snf.ch/project-147653

(*) N. Q. Luong and A. Popescu-Belis: Machine translation of Spanish personal and possessive pronouns using anaphora

probabilities. Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL), Valencia, 5-7 April 2017. pdf http://publications.idiap.ch/downloads/papers/2017/Luong_EACL_2017.pdf

(*) X. Pu, L. Mascarell and A. Popescu-Belis: Consistent Translation of Repeated Nouns using Syntactic and Semantic Cues. Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL), Valencia, 5-7 April 2017. pdf http://publications.idiap.ch/downloads/papers/2017/Pu_EACL_2017.pdf

(*) L. Miculicich Werlen and A. Popescu-Belis: Using Coreference Links to Improve Spanish-to-English Machine Translation. Proceedings of the EACL Workshop on Coreference Resolution beyond OntoNotes (CORBON), Valencia, 4 April 2017. pdf http://publications.idiap.ch/downloads/papers/2017/Werlen_CORBON_2017.pdf

(*) A. R. Gonzales and D. Tuggener: Co-reference Resolution of Elided Subjects and Possessive Pronouns in Spanish-English Statistical Machine Translation. Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL), Valencia, 5-7 April 2017. pdf http://www.zora.uzh.ch/136447/1/CoReferenceAwareMT.pdf

Contact:

Andrei Popescu-Belis
Idiap Research Institute
Centre du Parc, CP 592
1920 Martigny
Tel.: +41 (0)27 721 77 29
E-mail: andrei.popescu-belis@idiap.ch