03.04.2017 – 08:00 Uhr

Maschinelles Übersetzen über die Satzgrenze hinaus

Bern (ots) -

Die Algorithmen der maschinellen Übersetzung verarbeiten Texte Satz für Satz. So entgeht ihnen ein Grossteil des Kontexts, was zu Übersetzungsfehlern führt. Ein vom SNF unterstütztes Projekt hat nun einen Ansatz entwickelt, der es möglich macht, Texte stärker als Ganzes zu erfassen.

Die vom Schweizerischen Nationalfonds (SNF) unterstützten Forschenden haben einen neuen Weg zur Verbesserung maschineller Übersetzungstools gefunden. Ein solches Tool ist auch die berühmte Software Google Translate, die täglich rund 100 Milliarden Wörter von einer Sprache in eine andere überträgt. Was die Informatiker und Sprachwissenschaftler, die an diesem Projekt mitarbeiten, erstmals zeigen konnten: Übersetzungstools werden besser, wenn man die künstliche Intelligenz dazu bringt, über die satzweise Verarbeitung hinaus Informationen zu berücksichtigen, die an anderen Stellen im Text stehen. Ihr Ansatz findet inzwischen weltweit Beachtung. Am 3. April stellen die Wissenschaftler ihre jüngsten Ergebnisse (*) im Rahmen einer Konferenz der Association for Computational Linguistics in Valencia (Spanien) vor.

Übersetzen ohne Textverständnis

«Maschinelle Übersetzungstools verstehen nicht wirklich den Sinn der Texte, die sie verarbeiten», erklärt Andrei Popescu-Belis, Projektleiter und Leiter der Natural Language Processing Group im Forschungsinstitut Idiap in Martigny (Wallis). Sie wenden statistische Regeln an, um Inhalte von einer Sprache in eine andere zu übertragen. Dabei gehen sie Satz für Satz vor. Allerdings fehlen den einzelnen Sätzen oft Informationen, die für ihre korrekte Übertragung wichtig sind. Die Tools müssten daher auch Dinge berücksichtigen können, die an anderen Stellen im Text stehen.»

Um ihre Annahme zu belegen, haben sich die Forschenden insbesondere mit den Pronomen beschäftigt - kleinen Wörtern, wie «er» oder «diese», die auf andere Textteile verweisen. Da diese Bezugswörter oft ausserhalb des zu übersetzenden Satzes stehen, machen die Übersetzungstools viele Fehler. Popescu-Belis nennt ein einfaches Beispiel aus dem Französischen, das sogar ausgeklügelte Tools in die Irre führt: «Meine Tante hat eine tolle Limousine gekauft. Sie ist aber nicht so schön.» Google Translate macht daraus im Englischen: «My aunt has bought a great sedan. But she is not so beautiful.» Das Tool übersetzt «sie» mit «she». Da sich dieses Pronomen aber nur auf Personen weiblichen Geschlechts bezieht, versteht der englische Leser, dass «meine Tante» «nicht sehr hübsch» ist.

Die Fallen der Statistik

Das Tool wird in die Irre geführt, weil es weiss, dass das Attribut «nicht sehr hübsch» sich häufiger auf Personen als auf Gegenstände bezieht. Stünde an seiner Stelle «rostig» oder «defekt» - also ein Begriff, der sich in der Regel auf Gegenstände bezieht, wären die Chancen für die korrekte Übersetzung «it» grösser.

Um ein passendes Ergebnis zu erhalten, hätte das maschinelle Übersetzungstool Informationen heranziehen müssen, die im ersten Satz enthalten sind. Das ist grob, was das Tool der Forschenden des Idiap leistet, das sie in Zusammenarbeit mit den sprachwissenschaftlichen Fachbereichen der Universitäten Genf und Utrecht (Niederlande) sowie dem Institut für Computerlinguistik der Universität Zürich entwickelt haben.

Die Wissenschaftler setzen in erster Linie selbstlernende («machine learning») Techniken ein. Bei jedem Versuch lassen sie die Algorithmen Hunderte von Parametern abgleichen, die hinzugefügt oder entfernt werden, bis sich das Ergebnis verbessert. «Im Prinzip geben wir dem System an, wie viele der voranstehenden Sätze es in welcher Weise analysieren muss. Dann testen wir es unter realen Bedingungen.»

Google rekrutiert Mitarbeitende des Projekts

Laut Popescu-Belis sind die Ergebnisse vielversprechend. Bei Sprachpaarungen wie Französisch-Englisch oder Spanisch-Englisch führen Pronomen die maschinellen Übersetzungstools in rund der Hälfte aller Fälle in die Irre. «Indem wir das Tool zwingen, auch Informationen zu berücksichtigen, die ausserhalb des gerade übersetzten Satzes stehen, können wir die Fehlerquote inzwischen auf 30 Prozent senken», sagt der Wissenschaftler.

Für die Forschenden geht die Herausforderung weit über die Problematik der Pronomen hinaus: Weitere Übersetzungsprobleme, die sich im Wesentlichen nur lösen lassen, wenn der Text nicht in einzelnen Sätzen, sondern in seiner Gesamtheit betrachtet wird, sind beispielsweise die Zeitenfolge, die Auswahl der passenden Terminologie und die richtige Sprachebene.

Auch wenn die von Popescu-Belis und seinen Kolleginnen und Kollegen entwickelten Techniken noch nicht für die breite Anwendung ausgereift sind, haben sie doch das Interesse der Akteure in diesem Bereich geweckt. «Unsere Arbeit hat gezeigt, dass das maschinelle Übersetzen sich von der reinen Satz-für-Satz-Übertragung lösen muss. Was uns besonders freut: Drei an diesem Projekt beteiligte Nachwuchsforschende arbeiten nun bei Google in Zürich zu diesem Thema. Das zeigt, wie gross das Interesse an unserem Ansatz ist.»

LINKS

Projekt "MODERN" in der Projekt-Datenbank des SNF 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. 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. 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. 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. http://www.zora.uzh.ch/136447/1/CoReferenceAwareMT.pdf

Kontakt:

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