Einfluss der Variation in Referenzübersetzungen auf die Bewertung von NMT-Ergebnissen durch BLEU

Miriam Gerken, Jule Isbrandt, Ulrich Heid
Universität Hildesheim
mail@miriam-gerken.de, jule@isbrandt-calle.de, heid@uni-hildesheim.de

In der Übersetzungswissenschaft existieren mehrere Kriterien, um zu überprüfen, wann eine Übersetzung geglückt ist. Statt einer manuellen Überprüfung maschineller Übersetzungen (MÜ) bezüglich solcher Kriterien wird bei der Entwicklung von MÜ-Systemen wie SMT oder NMT oft auf automatische Evaluation zurückgegriffen. BLEU ist dabei das Standardverfahren.[1] Es vergleicht MÜ-Ergebnisse mit manuell erstellten Referenzübersetzungen (RÜ). Je höher die Übereinstimmung zwischen MÜ und RÜ, desto besser wird die MÜ durch BLEU bewertet. Wir haben untersucht, welche Eigenschaften der RÜ Einfluss auf eine positive BLEU-Bewertung nehmen und wie nahe die BLEU-Bewertung von NMT-Ergebnissen einer übersetzungswissenschaftlichen Evaluation kommt.

Für diese Untersuchung wurde ein Korpus von RÜ je eines journalistischen Textes zur Brexit-Thematik in den Sprachrichtungen EN-DE (587 W.) und DE-EN (425 W.) aus Übersetzungskursen der Universität Hildesheim gewonnen. Das Korpus beinhaltet je drei Gold-Lösungen sowie je elf Studierendenlösungen, die z.T. erhebliche lexikalische, morphosyntaktische bzw. syntaktische Variation aufweisen. Die RÜ wurden danach klassifiziert, ob sie besonders ausgangstextnah oder besonders zielsprachlich orientiert sind und ob sie sich bezüglich der obigen Variationen markant von den Gold-Lösungen unterscheiden. Mit den 14 RÜ in unterschiedlichen Konstellationen wurden je eine Google Translate- sowie eine DeepL-Übersetzung durch BLEU bewertet.

Dabei wurde zunächst untersucht, wie sich die Anzahl von RÜ sowie unterschiedliche Reihenfolgen der RÜ auf die BLEU-Bewertung auswirken.

Wie erwartet steigen die BLEU-Werte mit der Anzahl der RÜ. Je unterschiedlicher dabei die RÜ sind, desto stärker ist die Zunahme des BLEU-Werts. In unserem Setup nähert sich der BLEU-Wert ab acht RÜ einer Obergrenze an. Ausgangstextorientierte RÜ erzielten außerdem höhere BLEU-Werte als zielsprachlich orientierte RÜ: Die MÜ-Systeme übersetzen eher ausgangstextnah. Bei der BLEU-Bewertung EN-DE gegen zielsprachlich orientierte RÜ erzielt Google Translate höhere Werte als DeepL. Dies widerspricht der übersetzungswissenschaftlichen Bewertung der Ergebnisse, die DeepL als stärker zielsprachlich orientiert klassifiziert. Aus übersetzungswissenschaftlicher Sicht relevante lexikalische, morphosyntaktische oder syntaktische Variationen gegenüber den Gold-Lösungen scheinen dagegen nur einen geringen Einfluss auf die BLEU-Bewertung zu haben.