Machine Translation of OHDSI Standard Vocabulary Terms

Erik M. van Mulligen, PhD[1], Peter R. Rijnbeek, PhD[1], and Jan A. Kors, PhD[1]

[1]Dept of Medical Informatics, Erasmus University Medical Center, Rotterdam, The Netherlands

Background

The OHDSI vocabularies only contain English terms, which cannot be used to annotate free-text electronic patient records in non-English languages. We explored the possibilities to translate one of the OHDSI standardized vocabularies for the conditions domain, SNOMED-CT, from English into Dutch. We compared three state-of-the-art machine translation systems and evaluated the quality of the translations to assess whether automatic machine translation can effectively be used to generate a vocabulary in another language.

Methods

From the OHDSI vocabulary, 104,214 preferred English terms of SNOMED-CT condition concepts were extracted early February 2018. Each term was fed into three machine translation systems: Google Translate, Microsoft Translator, and DeepL. For each term, the Dutch translations by the machine translation services were compared case insensitive. If two or three translations matched exactly, this translation was taken as the majority translation. For a random subset of 500 terms, the individual translations and the majority translation (if present) were independently assessed by two reviewers, with a final session to reach consensus.

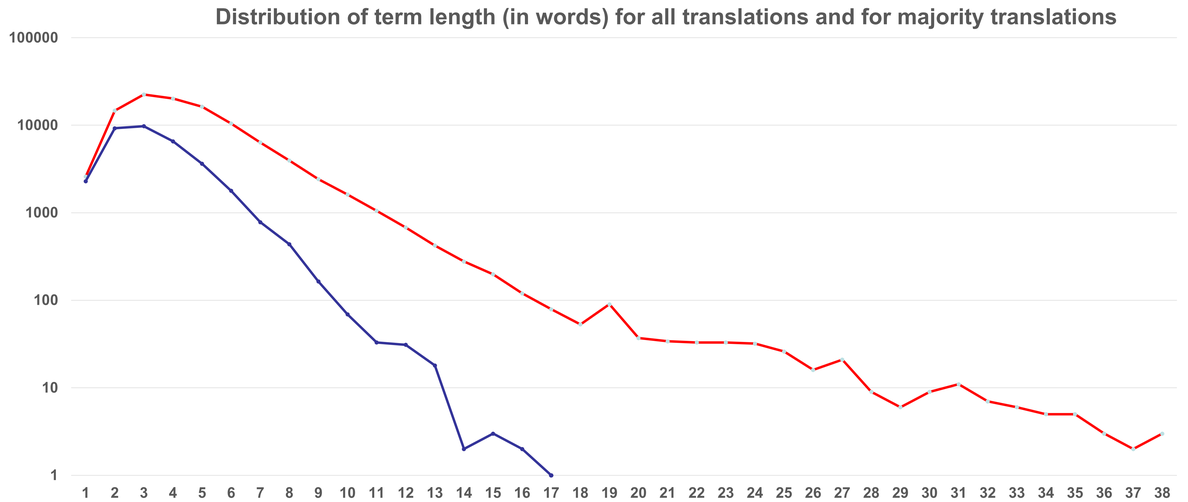
Results

The translations of the 104,214 terms from SNOMED-CT yielded 34,777 (33.4%) majority translations. Figure 1 shows the distribution of the term length for all translated terms and for the majority translations, indicating that term length has a strong impact on the presence of majority translations. The translations of the 500 randomly selected terms resulted in 172 (34.4%) majority translations, of which 154 (89.5%) were correct. Of the translations that resulted in a majority translation, Google Translate contributed 147 (85.5%), MS Translator 121 (70.3%), and DeepL 120 (69.8%). Of the correct majority translations, Google Translate contributed 131 (85.1%), MS Translator 107 (69.5%), and DeepL 111 (72.1%). For the 328 English terms that had no majority translation, Google Translate provided 247 (75.3%) correct translations, MS Translator 180 (54.9%), and DeepL 170 (51.8%). Of all 500 terms, Google correctly translated 394 (78.8%), MS Translator 301 (60.2%), and DeepL 290 (58.0%). Figure 2 shows the accuracy (proportion of correct translations) as a function of term length, per translation system.

Discussion

We present the results for automatic translation of English terms from a subset of the OHDSI standard vocabulary into Dutch. The results show that state-of-the-art machine translation systems can correctly translate a significant part of the terms. Of the three systems, Google Translate is the most accurate, showing good performance for terms up to 9 words (Figure 2). Construction of a majority translation by combining systems can further improve performance, but majority translations are only available for a third of the terms. This is partially due to our strict matching algorithm in constructing the majority translation. Manual inspection indicated that unimportant variations in word order or articles sometimes prevent construction of a majority translation. We also noticed that punctuation impacts the translation accuracy (e.g., Google Translate ignores anything after a slash and all systems have issues with hyphens). Future research will investigate the occurrence of translated terms in a Dutch electronic patient record system (IPCI).

Contact:e.vanmulligen@erasmusmc.nl



	Contribution to majority	Majority correct	Non-majority correct	Overall correct
Google Translate	147 (85.5%)	131 (85.1%)	247 (75.3%)	394 (78.8%)
MS Translator	121 (70.3%)	107 (69.5%)	180 (54.9%)	301 (60.2%)
DeepL	120 (69.8%)	111 (72.1%)	170 (51.8%)	290 (58.0%)
Majority	172 (100.0%)	154 (89.5%)