Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220641

Validation of Multiple Path Translation for SNOMED CT Localisation

Stefan SCHULZ[a,1], Martin BOEKER[b] and Andrea PRUNOTTO[b]
[a] *IMI, Medical University of Graz, Austria*
[b] *Institute for AI in Healthcare, Technical University of Munich, Germany*
[c] *IMBI, University of Freiburg, Germany*

Abstract. The MTP (multiple translation paths) approach supports human translators in clinical terminology localization. It exploits the results of web-based machine translation tools and generates, for a chosen target language, a scored output of translation candidates for each input terminology code. We present first results of a validation, using four SNOMED CT benchmarks and three translation engines. For German as target language, there was a significant advantage of MTP as a generator of plausible translation candidate lists, and a moderate advantage of the top-ranked MTP translation candidate over single best performing direct-translation approaches.

Keywords. Machine Translation, SNOMED CT, NLP

1. Introduction

SNOMED CT's [1] acceptance depends on localization, i.e. its adaptation to the end users' language. SNOMED CT is distributed in English and Spanish, with official translations available for some other languages. Translations provide a unique label (*fully specified name*, FSN) for each localized SNOMED CT code. As terminology translation is expensive and time-consuming, free machine translation tools bear the promise to accelerate this process. The MTP (multiple translation path) approach [2] was developed to generate, for a chosen target language, a scored output of translation candidates (TCs) for each SNOMED CT code and the related source FSNs. Besides direct translations, e.g. from Spanish to German, additional paths are created via support languages (e.g. Spanish via English to German) in order to collect a majority-vote TC list per code.

2. Material and Methods

We compared direct with MTP translations to German using four benchmarks: (i) value sets of the German BfArM Catalogue linked to SNOMED CT, enriched with synonyms provided by [3]; (ii) the unofficial translation of an early SNOMED CT version, (iii) a random subset of the 2021 SNOMED CT release, translated by medical students; and (iv) the SNOMED CT Starter Set collection [4], enriched by German synonyms. English, Spanish and Swedish were taken as source languages for Google Translator, DeepL and Systran. Danish, Dutch, Norwegian, Italian, Portuguese, Polish and Russian were chosen

[1] Corresponding Author, Stefan SCHULZ; E-mail: stefan.schulz@medunigraz.at

as support languages. From each benchmark, a random sample of 500 SNOMED CT codes that existed in the 2021 SNOMED CT release was extracted. For each code and language, the FSN (without the hierarchy tag) was fed into the process.

3. Results and Discussion

Table 1. Exact match and 1-gram BLEU metric related to benchmarks for direct and MTP translations at various rank-range choices. Uncertainties on these values were obtained by performing 10 different random choices of 500 terms (among a total of 1000 MTP-translated terms) for each benchmark, and evaluating the standard deviation both on PEM and BLEU metric.

Exact match	BfArM Catalogue	SNOMED-CT (2003)	SNOMED-CT (2021)	SNOMED-CT (Starter-set)
MTP (rank 1-5)	73.90 +/- 1.49	29.96 +/- 0.95	77.91 +/- 1.26	55.35 +/- 1.27
MTP (rank 1-4)	72.40 +/- 1.41	28.71 +/- 0.91	75.49 +/- 1.35	51.98 +/- 1.46
MTP (rank 1-3)	69.57 +/- 1.43	27.15 +/- 0.98	72.27 +/- 1.44	49.54 +/- 1.61
MTP (rank 1-2)	65.26 +/- 1.44	23.88 +/- 0.80	65.98 +/- 1.59	44.71 +/- 1.53
MTP (rank 1)	54.23 +/- 1.56	17.60 +/- 0.88	51.99 +/- 1.67	35.29 +/- 1.63
en--google--de	50.11 +/- 1.71	17.07 +/- 1.33	45.84 +/- 1.60	32.91 +/- 1.53
en--deepl--de	43.21 +/- 1.01	19.41 +/- 1.12	41.18 +/- 1.68	33.35 +/- 1.39
es--google--de	37.92 +/- 1.63	13.21 +/- 1.17	40.56 +/- 1.44	28.95 +/- 1.37
es--deepl--de	37.85 +/- 1.31	14.17 +/- 0.97	35.64 +/- 2.09	29.63 +/- 1.61
en--systran--de	37.32 +/- 1.54	14.57 +/- 1.30	29.51 +/- 1.56	29.7 +/- 1.57
sv--google--de	33.48 +/- 1.40	10.39 +/- 1.02	28.66 +/- 1.49	25.77 +/- 1.38
sv--systran--de	23.72 +/- 1.42	7.38 +/- 0.87	16.07 +/- 1.41	20.68 +/- 1.17
es--systran--de	22.31 +/- 1.23	7.72 +/- 0.94	19.10 +/- 1.31	16.03 +/- 0.82

BLEU metric	BfArM Catalogue	SNOMED-CT (2003)	SNOMED-CT (2021)	SNOMED-CT (Starter-set)
MTP (rank 1-5)	0.93 +/- 0.01	0.50 +/- 0.01	0.80 +/- 0.01	0.62 +/- 0.01
MTP (rank 1-4)	0.91 +/- 0.01	0.48 +/- 0.01	0.78 +/- 0.01	0.60 +/- 0.01
MTP (rank 1-3)	0.89 +/- 0.01	0.44 +/- 0.01	0.75 +/- 0.01	0.56 +/- 0.01
MTP (rank 1-2)	0.86 +/- 0.01	0.38 +/- 0.01	0.68 +/- 0.01	0.49 +/- 0.01
MTP (rank 1)	0.75 +/- 0.01	0.28 +/- 0.01	0.53 +/- 0.01	0.35 +/- 0.01
en--google--de	0.69 +/- 0.01	0.30 +/- 0.01	0.51 +/- 0.01	0.34 +/- 0.01
en--deepl--de	0.69 +/- 0.01	0.32 +/- 0.01	0.47 +/- 0.01	0.34 +/- 0.01
sv--google--de	0.51 +/- 0.01	0.31 +/- 0.01	0.55 +/- 0.01	0.41 +/- 0.02
en--systran--de	0.58 +/- 0.01	0.26 +/- 0.01	0.43 +/- 0.01	0.33 +/- 0.01
es--google--de	0.55 +/- 0.01	0.23 +/- 0.01	0.48 +/- 0.01	0.31 +/- 0.01
es--deepl--de	0.53 +/- 0.01	0.26 +/- 0.01	0.45 +/- 0.01	0.31 +/- 0.01
sv--systran--de	0.41 +/- 0.01	0.25 +/- 0.01	0.41 +/- 0.01	0.35 +/- 0.01
es--systran--de	0.39 +/- 0.01	0.18 +/- 0.01	0.34 +/- 0.01	0.22 +/- 0.01

For each input code, MTP resulted in 91 translation paths with various degrees of coincidence, with an average 28.7 distinct translations per code. Tab. 1 shows the comparison between MTP and direct translation (DT) in their capability to target the benchmarks, measured by % exact match (PEM) and the 1-gram BLEU metric [5]. Values strongly vary, which is explainable by different support of synonyms. MTP at rank 1 slightly outperforms any MT tool (around 1-3%). Extending to rank 1 and 2, MTP performance rises up to 20%. Including candidates at lower rank improves only to 1-3%. Restricting the comparison to rank 1 only (i.e. considering MTP a translator that produces only one translation like in DT), only the first scenario showed an advantage of MTP, whereas MTP was outperformed in the other cases particularly by the Swedish-to-German Google translation scenario by up to .06 (BLEU metric) in the fourth scenario. Finally, it is to be noted that a fraction of exact matches with the human translation is found only by MTP but not by any DT. This means that some human-like translations are found only by means of the use of intermediate languages. Our work is encouraging insofar that it suggests that for terminology translation the combination of web-based translation engines produces higher translation quality and coverage The effect size is remarkable for MTP as a shortlist creator, but still moderate when considering only top-ranked TCs.

Acknowledgements: BMBF MIRACUM, FKZ: 01ZZ1801A.

References

[1] Bodenreider O, Cornet R, Vreeman DJ. Recent Developments in Clinical Terminologies - SNOMED CT, LOINC, and RxNorm. Yearb Med Inform. 2018 Aug;27(1):129-139.

[2] Prunotto A, Schulz S, Boeker M. Automatic Generation of German Translation Candidates for SNOMED CT Textual Descriptions. Stud Health Technol Inform. 2021 May 27;281:178-182.

[3] Hashemian Nik D, Kasáč Z, Goda Z, Semlitsch A, Schulz S. Building an Experimental German User Interface Terminology Linked to SNOMED CT. Stud Health Technol Inform. 2019 Aug 21;264:153-157.

[4] SNOMED CT Starter-set. Available at: https://www.snomed.org/news-and-events/articles/snomed-ct-starter-set-translation-rfp. Accessed March 30, 2022.

[5] Papineni K, Roukos S, Ward T, Zhu WJ. Bleu: a method for automatic evaluation of machine translation. Proceedings of the 40th annual meeting of the Association for Computational Linguistics, 2002; 311-318.