Validation of Multiple Path Translation (MTP) for SNOMED CT localisation

Andrea PRUNOTTO[a,1] , Martin BOEKER[b] ,Stefan SCHULZ [c]
[a] *Institute of Medical Biometry and Statistics, University of Freiburg, Germany*
[b] *Institute of Medical Informatics, Statistics and Epidemiology, Technical University of Munich, Germany*
[c] *Institute for Medical Informatics, Statistics and Documentation, Medical University of Graz, Austria*

Abstract. The MTP (multiple translation paths) approach was developed to support human translators in the process of clinical terminology localisation. It exploits the results of free, web-based machine translation tools and generates, for a chosen target language, a scored output of translation candidates for each input term. We here present the first results of a validation of this approach, which makes use of four SNOMED CT benchmarks, and three machine translation engines. We could show a significant advantage of the MTP approach as a generator of plausible translation-candidates lists, and a moderate advantage of the top-ranked MTP translation candidate over single best performing direct-translation approaches. These result are based both on the percentage of exact match with the benchmark terms and on the BLEU metric. From the discussion of our preliminary results we derive a list of suggestions for future work.

Keywords. Machine Translation, SNOMED CT, NLP

1. Introduction

The acceptance of an international terminology standard, like SNOMED CT [1], crucially depends on its localisation, i.e. on its capability to adapt to a given language in such a way that it can both mirror the sublanguage preferences of users and, at the same time, maintain the original meaning of the representational units. In SNOMED CT, each representational unit (SNOMED CT *concept*) is identified by a unique code, together with a unique, maximally self-explaining label (or *fully specified name*, FSN) in English and Spanish, the two official languages of SNOMED CT. For other languages, there are no (or only limited numbers of) term translations. Current localisation projects and released translations are often limited in scope (pre-selected content) and granularity (e.g. translation of FSNs only). Capitalising on free, web-based machine translation tools [2-4], we have proposed an approach named MTP (multiple translation path) aiming at assisting human translation in SNOMED CT localisation projects [5]. For a chosen target language, MTP generates a scored output of translation candidates (TCs) for each SNOMED CT concept. This paper outlines MTP and provides first validation results. MTP uses the joint power of popular machine translation engines and different source/intermediate languages in order to collect a majority-vote based translation-candidates list for each concept into a target language. The ultimate goal of the process is to support and fasten the onerous task of human translators involved in the localisation of SNOMED CT terminology (consisting of thousands of complex medical terms), by providing the former with a shortlist of highly statistical significant possibilities to choose therein.

2. Material and Methods

The currently available web-based translation tools yield exactly one translation per input string. If we feed a single tool with input strings for the same concept from *n* source languages we expect up to *n* distinct TCs per concept. We will refer to those *n* translations as to *direct translations* (DT). The basic element of MTP is the translation path (tp). Translation paths are defined as triples (s, e, t), with “s” being a language from the set of source languages S, “e” a translation engine from the set of engines E and “t” a language from the set of target languages T. Assuming that all engines serve all language, the number of translation paths |tp| in the DT scenario with one output language equals |S| x |E|. Our preliminary work has shown that DTs from different

1 Corresponding Author; Andrea Prunotto, Instituts für Medizinische Biometrie und Statistik, Stefan-Maier Str. 26, 79104 Freiburg im Breisgau, Germany. Email: prunotto@imbi.uni-freiburg.de

sources often coincide, especially for short input terms, often no TC is of the expected quality, and clinically relevant synonyms are more difficult to be generated. An approach that generates more TCs would widen their lexical variability. Therefore, we devised a strategy involving *support translations*. This means translating a source term first into an intermediate language, and then translating the result into the target language. This process can therefore be split in two translation paths, tp_1 and tp_2. This *multiple translation paths* approach capitalises on combinatorial growth by the combination of source languages, intermediate support languages, and translation engines. tp_1 is thus defined as ("s", "e", "I") with "I" being a language out of the set of intermediate support languages I, yielding the number of primary translation paths: $|tp_1| = |S| \times |E| \times |I|$. The secondary translation path tp_2 translates a tp_1 output to the target language T, using again all translation engines. For MPT the number of all paths is therefore $|tp| = |S| \times |E|^2 \times |I|$. In the case of four source languages, three translation engines and five intermediate languages the number of MPTs would therefore be $4 \times 3^2 \times 5 = 180$, along with $4 \times 3 = 12$ for DTs. In practice however, many secondary translation paths coincide because translations from the primary paths will be the same, not all engines support the same language pairs, and paths may end with no translation at all. On the other hand, more paths can be yielded with more than one input string per language, when considering synonyms. We applied DT and MTP to SNOMED CT concepts represented by several benchmarks. English, Spanish and Swedish were taken as source languages for Google Translator, Deep and Systran. Danish, Dutch, Norwegian, Italian, Portuguese, Polish and Russian were defined as support languages. German was the only target language. The following benchmarks were used:

- BfArM Catalogue: A semiautomatic translation of value sets for different clinical use cases into German, collected by the German BfArM, annotated with SNOMED CT codes. Synonyms were added using the SNOMED CT interface terminology [6] created by the first author and reviewed by experts
- SNOMED-CT (2003): unofficial German translation of an early version of SNOMED CT, provided by SNOMED International for experimental purposes
- SNOMED-CT (2021): a random subset of the current (2021) SNOMED CT translated by medical students under supervision of the authors
- SNOMED-CT (Starter-set): a subset of 6006 popular SNOMED CT concepts, distributed by SNOMED International as the SNOMED CT Starter-set [7]

From each benchmark, a random sample of 500 SNOMED CT concepts that existed in the current version was extracted. For each concept, the FSN (without the hierarchy tag) was fed into the process.

3. Results

For each input SNOMED CT concept, the MTP scenario resulted in 91 translation paths with various degrees of coincidence (cardinality), resulting in ranked lists of TCs. Ranking by cardinality is based on the assumption that those TCs that resulted from many paths independently, were more trustworthy than those generated by just one path. Compared to the maximal number, most configurations produced an average number of TCs indicating that the same candidate is often derived via different translation paths. As expected, combinations of translation engines result in distributions with a higher number of distinct TCs per concept: MTP yields on average 28.7 distinct translations per concept [5]. Tab. 1 presents the comparison between MTP and DT in their capability to target the above-mentioned four human-translated benchmarks. This capability is measured both by the percentage of exact match (PEM) and by the 1-gram BLEU metric [8], excluding capitalization and hyphenation. All values strongly vary between the four benchmarks. This is explained by different support of synonyms therein, with the SNOMED-CT (2003) translation having the least synonyms. MTP at rank 1 slightly outperforms any MT tool (around 1-3%). By including candidates at rank 1 to 2, the MTP performance rises up to 20%. Including further candidates at lower rank improves only to 1-3%. This suggests that, beyond the first 3 candidates, MTP produces translations which are unlikely to be human plausible (this trend appears similar both according to PEM and BLEU metric). However, restricting the comparison between DT and MTP at rank 1 (i.e. considering MTP as an autonomous machine translator), only the first scenario (BfArM) 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 Starter-set 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. This result, shown by means of an UpSet plot [9], is illustrated in Fig. 1. More details are available at [10].

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

Table 1: PEM (Left) and 1-gram BLEU metric (Right) related to the four benchmarks for DT translations and MTP at various rank-range choices. Note that the four benchmarks are very different regarding the selection of concepts and the availability of synonyms. Uncertainties on these values are 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.

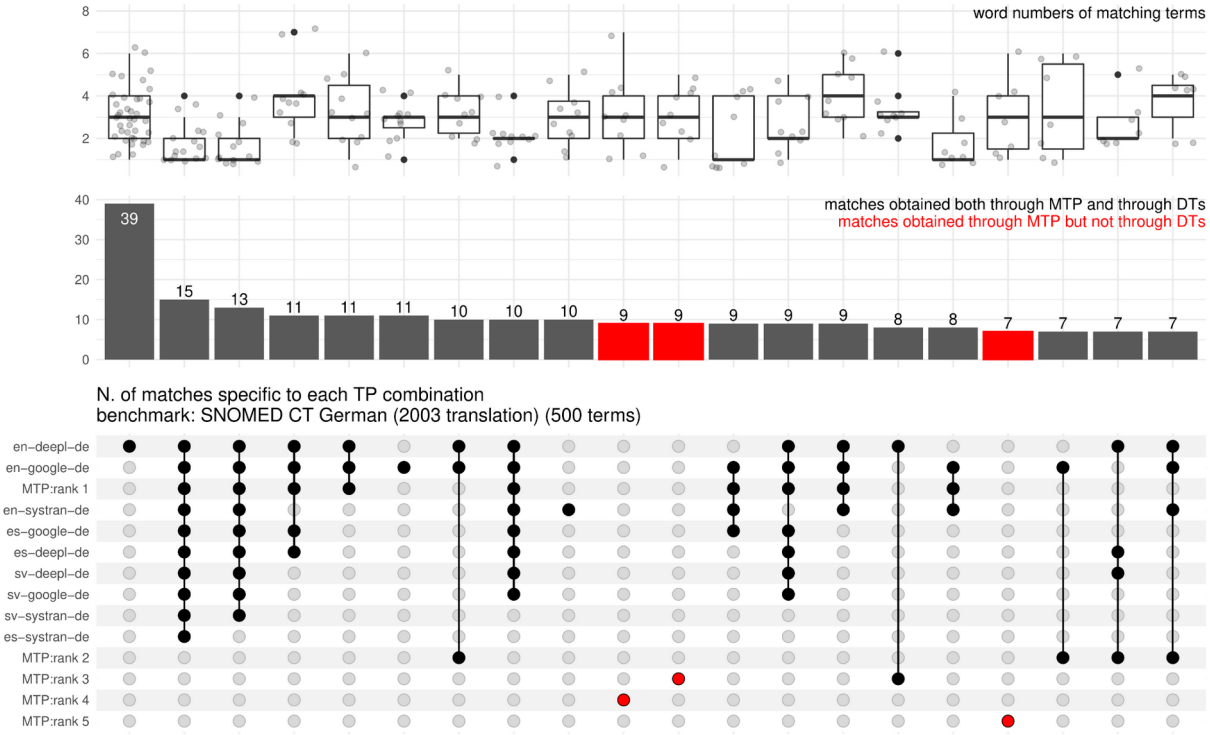

Figure 1: UpSet Plot showing the number of matching translations found by MTP at various ranks and by DTs (benchmark: SNOMED CT 2003). The bottom panel shows all the possible combinations in which a matching translation can be found: e.g. only by en-deepl-de, or by MTP rank 1 and en-google-de, etc. Terms that fall in one column do not fall in any other column (exclusive intersection). The central histogram displays the number of matches found in each combination (red bars indicate that those terms were found only by MTP at the related rank, but not by any DT). Finally, the top boxplot displays the number of words composing the terms that were found by that combination of TPs. Remarkably, MTP ar rank 3,4,5 detect 8,9,7 terms, respectively, which are not detected by and DT path, and targeting concepts whose textual description reaches 7 words.

4. Discussion

The main feature of the MTP method is to create *lists* of TCs. Conversely, the direct translations approach - using one single web-based translation engine - produces one single TC. However, the choice of using one specific engine among the many currently available (e.g. Google Translate, Deepl, Systran), as well as the choice of one of the available source languages (English, Spanish, Swedish), is actually arbitrary. Moreover, the coupled use of different engines and source languages often results in distinct translations, especially for complex entries (this suggests that also the direct translation approach is intrinsically MTP one). The evaluation of the results in the first row in Tab. 1 gives a hint to the usefulness of the method in a scenario in which MTP is used to provide shortlists of candidates (here, up to five TCs) for downstream human selection. The highest performance is shown by the SNOMED CT (2021) selection with a likelihood of 80% to have a correct

translation among the five most frequent TCs, all the more that this gold standard does not contain synonyms. However, this set includes 107 concepts consisting of one word and 269 of two words (many of them being the Latin version of disorders/findings). On the other hand, the higher the length of input terms, the likelier a translation, even if valid, is not contained in the benchmark, since the number of translational permutations increases with the number of words. The results related to the SNOMED CT (2003) translations are quite different. This can be explained on the one hand, by the fact that this corpus only covers parts of the terminology and, on the other hand, by observing that many FSN translations are overly synthetic and rarely used in practice, so that automated translations often prefer other ones. The difference between the BfArM (first column) and Starter-set (last column) translations, both of which include more "popular" concepts can be explained by the higher coverage of synonyms of the former. A direct comparison between MTP and DT with a baseline is however possible only for the top-ranked term of the MTP pipeline, i.e. the one with the highest cardinality regarding the MTP output. We can state a moderate advantage of this "majority vote" when comparing it to single translation paths, where English as source language - not surprisingly - outperformed Spanish, which still performed better than English. Across the scenarios without Swedish, Deepl and Google Translate showed comparable behaviour. These are preliminary results, and several limitations have to be addressed. The main limitation is the error analysis. Apart from a few qualitative analyses in our first study [5] we have not yet systematically investigated the non-matching results, of which we assume that only a part is plainly wrong, whereas the other part did not match just because of too few synonyms and variations in the gold standard. Another limitation is that we did not comparatively assess the quality of the output of the translation paths. Particularly the paths that use small languages as intermediate languages could be of little help or could even support the accumulation of wrong TCs. It is known that when MT systems translate smaller languages internally use a pivot language, normally English. This may introduce additional noise due to word sense ambiguity. As an example, Google Translate wrongly translates the Danish word "ryg" ("back" in the sense of "back pain") to the German "zurück", whereas the correct translation would be "Rücken". This strongly suggests that the engine uses English as a pivot language, and the error is due to the ambiguity of the English word "back", which can mean either the anatomical region or a spatial or temporal direction such as in "back to school". The rationale of using the BLEU metric as an alternative to PEM also requires more in-depth error analysis. At this point we can only speculate why the BLEU performed better in three scenarios based on direct translations starting with Swedish. Maybe the benefit came from untranslated content and the proximity of Swedish and German words, such as in "familjeanamnes på demens" vs. "Familienanamnese einer Demenz" (English: "family history of dementia").

5. Conclusions and Further Work

Our work is encouraging insofar that it suggests that the combination of web-based translation engines produces higher translation quality and coverage when using them for medical terminology translation. The effect size is remarkable if MTP is used as a shortlist creator, but still moderate when considering only the top-ranked TCs. However, we see considerable potential for improvement. In particular we want to highlight the following points:

- Revisiting and manual enrichment of benchmark translations by more TCs
- Addition of new translation paths with new intermediate languages, assessing the relative gain of new languages and language combinations
- Assessment of structural features of input terms regarding the preference of translation paths. We hypothesise that flat nominal sequences in English terms such as "community outreach worker services surveillance" are more prone to translation errors compared to nested prepositional expressions like in Spanish "vigilancia de servicios del programa de atención comunitaria".
- Addition of synonyms from the English and Spanish description table as input terms, which however requires the recognition and exclusion of ambiguous synonyms beforehand
- Including new input languages, as SNOMED CT translation activities evolve
- Extending the methodology to new output languages, particularly minor ones, e.g. Eastern European languages
- Considering new translation engines and new functionalities of existing translation engines, particularly the possibility of producing more than one translation output
- Considering online and offline dictionaries
- Considering open community-built resources like Wikidata, DBpedia etc.
- Clarifying legal requirements for large-scale use of translation engines and dictionaries

Acknowledgements: The first author is funded by the German Federal Ministry of Education and Research (BMBF) within the Medical Informatics Initiative (MIRACUM Consortium) under the Funding Number FKZ: 01ZZ1801A.

6. 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] Google Translate. Available at: https://translate.google.com. Accessed January 22, 2022.
[3] Deepl Translator. Available at: https://www.deepl.com/translator. Accessed January 22, 2022.
[4] Systran Translator. Available at: https://translate.systran.net. Accessed January 22, 2022.
[5] 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.
[6] 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.
[7] SNOMED CT Starter-set. Available at: https://www.snomed.org/news-and-events/articles/snomed-ct-starter-set-translation-rfp. Accessed January 22, 2022.
[8] 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.
[9] Lex A, Gehlenborg N, Strobelt H, Vuillemot R, Pfister H: UpSet: Visualization of Intersecting Sets. IEEE Transactions on Visualization and Computer Graphics, vol. 20, no. 12, pp. 1983-1992, 31 Dec. 2014.
[10] SNOMED CT MTP Validation github repository. Available at: https://github.com/andreaprunotto/SNOMED-CT-MTP-Validation. Accessed January 22, 2022.