Automatic Generation of German Translation Candidates for SNOMED CT Textual Descriptions

Andrea PRUNOTTO[a,1] Stefan SCHULZ[b] and Martin BOEKER[a]

[a]*Institute of Medical Biometry and Statistics,*
University of Freiburg, Germany
[b]*Institute for Medical Informatics, Statistics and Documentation,*
Medical University of Graz, Austria

Abstract. We present an approach called MTP (multiple translation paths) aiming at assisting human translation in SNOMED CT localisation projects based on free, web-based machine translation tools. For a chosen target language, MTP generates a scored output of translation candidates (TCs) for each input concept. This paper describes the basic idea of MTP, the distribution of its output TCs and discusses typical examples with German as target language. The MTP approach capitalises on combinatorial growth by the combination of input languages, support languages, and translation engines. We applied MTP on the SNOMED CT Starter Set, using Google Translator, DeepL and Systran, together with the four source languages English, Spanish, Swedish and French, and Danish, Dutch, Norwegian, Italian, Portuguese, Polish and Russian as support languages. The descriptive assessment of TC variety, together with an analysis of typical results is the focus of this paper. MTP defines, for each input concept, TPs by the combination of input languages, support languages and translation engines, resulting in 91 translation results with various degrees of co-incidence (cardinality). The most configurations produce an average number of TCs indicating that the same TC is often derived via different translation paths. Combinations of translation engines result in distributions with a higher number of distinct TCs per concept. We present work in progress on using machine translation (MT) for terminology translation, by leveraging several free MT tools fed by different languages and language combinations. A first qualitative analysis was promising and supports our hypothesis that a majority voting applied to many translation candidates yields higher quality results than from one single engine and input language.

Keywords. Machine Translation, SNOMED CT, NLP

1. Introduction

SNOMED CT is a huge international clinical terminology standard, which provides codes and terms for describing all aspects of electronic health records [1]. Representational units, known as SNOMED CT concepts, are characterised by numeric codes, by unique labels in English and Spanish, known as *fully specified names* (FSNs), varying numbers of synonyms and ontology-like axioms. The July 2020 international

---

[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

release includes 354,384 active SNOMED CT concepts. FSNs are devised as maximally unambiguous and explanatory, e.g., "Primary malignant neoplasm of the lung (disorder)", whereas non-preferential synonyms better align with the clinical jargon ("lung cancer"). For most 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). The size of SNOMED CT renders manual translation extremely onerous. In the light of the popularisation of free, web-based machine translation (MT) tools [2-6], we are developing an approach called MTP aiming at assisting human translation in SNOMED CT localisation projects. For a chosen target language, MTP generates a scored output of *translation candidates* (TCs) for each input FSN. This paper describes the basic idea of MTP, the distribution of its output TCs and discusses typical examples with German as target language. A systematic validation, currently in process, will be published elsewhere.

2. Material and Methods

This work is based on a small subset of SNOMED CT, *viz.* the English, Spanish, Swedish and French versions of the SNOMED CT Starter Set [7], consisting of one FSN per language for 6006 concepts, selected according to the frequency of use in clinical data across disciplines.

If we use a single MT tool and feed it with the FSNs from *n* source languages we expect up to *n* distinct TCs per concept. We refer to this as *direct translation* (DT). Because DTs from different sources often coincide, especially for short input terms, often no TCs is of the expected quality and clinically relevant synonyms are not generated. An approach that generates more TCs would widen their lexical variability. Therefore, we devised a strategy involving: (i) *more* independent MT engines, and (ii) *support translations* (ST). The latter means translating a source term first into an intermediate language, and then translating the result into the target language. This MTP (*multiple translation paths*) approach capitalises on combinatorial growth by the combination of input languages, support languages, and translation engines. The descriptive assessment of this growth in terms of TC variety, together with an analysis of typical results is the focus of this paper.

We applied MTP on the Starter Set, using Google Translator [8] DeepL [9] and Systran [10], together with the four source languages (see above) and Danish, Dutch, Norwegian, Italian, Portuguese, Polish and Russian as support languages. Danish, Dutch, Norwegian were chosen for their similarity with German and Swedish, whereas Italian and Portuguese for their similarity with Spanish and French. Polish and Russian were selected to increase translation flexibility (Table 1).

Table 1. List of languages used to form TPs with Google, DeepL and Systran. Out of these selected languages (and starting from English, Spanish, Swedish and French as language sources), up to 91 different TPs can be formed. (DeepL does not provide translations from Swedish).

	en	fr	es	it	pt	pl	ru	sv	no	da	de	nl
Google	√	√	√	√	√	√	√	√	√	√	√	√
DeepL	√	√	√	√	√	√	√	×	×	×	√	√
Systran	√	√	√	√	×	×	√	√	√	√	√	√

3. Results

MTP defines, for each input concept, TPs by the combination of input languages, support languages and translation engines, resulting in 91 translation results with various degrees of co-incidence (cardinality). Fig. 1 visualises the distribution of the cardinality of distinct TCs, in contrast to configurations with less combinations, such as single-engine and direct translations only.

Each of the resulting graphs is characteristic for each of the configurations (Fig. 1). In relation to the maximally possible number, the most configurations produce an average number of TCs indicating that the same TC 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 [11].
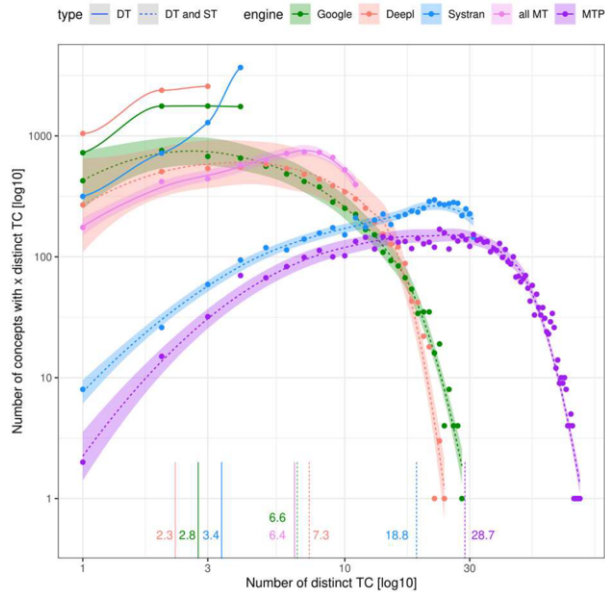
Figure 1. Distribution of (distinct) German TCs per concept obtained by different methods (Data: SNOMED CT Starter Set in four source languages, colour = MT engines and MTP combination of them, line type =direct/support translations). One MT and DTs, generate at most 4 distinct TCs per concept, with an average of 2.3 with DeepL (red), 2.8 with Google (green) and 3.4 with Systran (blue). Pooling all MT engines together, producing DTs only, we obtain at most 11 candidates (not 12, since DeepL supports only 3 source languages), with an average of 7.3 with DeepL (red), 6.6 with Google (green) and 18.8 with Systran (blue). MTP yield up to 79 distinct candidates (28.7 on average) per concept.

For each concept, we ordered the TCs by decreasing cardinality (i.e., how often a distinct TP was derived in all 91 possible translations). This does not necessarily mean that this ranking, based on "majority vote", actually reflects translation quality, but it seems reasonable to assume that those TCs which were produced by many independent paths, are more trustworthy than those generated by just one TP. Finally, such a hypothesis can only be validated against a gold standard. What we can currently report of, is case evidence, such as demonstrated in Table 2, with ranked TC lists for selected SNOMED CT concepts.

Table 2. Examples of shortlists of distinct TCs for selected SNOMED concepts (English FSNs without hierarch tag, in *italics*), ranked according to their TP cardinality. The total number of distinct TCs is given in square brackets. TCs with cardinality = 1 are removed for better readability.

70342003: *Cholelithiasis without obstruction*	[43]
Cholelithiasis ohne Obstruktion	29
Gallensteinkrankheit ohne Obstruktion	8
Cholelithiasis	5
Freie Cholelithiasis	4
Colelitiase ohne Obstruktion	2
freie Cholelithiasis	2
Nicht-obstruierte Cholelithiasis	2
unbehinderte Cholelithiasis	2
ungehinderte Cholelithiasis	2
Ungehinderte Cholelithiasis	2
Unverschobene Cholelithiasis	2

75137002: *Closed traumatic dislocation of joint of finger*	[53]
Geschlossene traumatische Dislokation des Fingergelenks	7
Geschlossene traumatische Luxation eines Fingergelenks	7
Geschlossene traumatische Luxation des Fingergelenks	6
geschlossene Luxation des Fingers	5
unbelichtete traumatische Dislokation eines Fingergelenks	5
Geschlossene traumatische Dislokation eines Fingergelenks	4
Traumatische unbelichtete Luxation des Fingergelenks	4
Geschlossene traumatische Verschiebung des Fingergelenks	3
Geschlossene traumatische Verschiebung eines Fingergelenks	2
Luxation des geschlossenen Fingers	2
Traumatische geschlossene Luxation des Fingergelenks	2
Traumatische Luxation des unbelichteten Fingergelenks	2
Traumatische unbelichtete Fingergelenkverschiebung	2
unexponierte traumatische Störung der Fingergelenke	2

70650003: *Urinary bladder stone*	[34]
Harnblasenstein	17
Harnlithiasis	14
Blasenstein	11
Berechnung der Harnblase	9
Blasensteine	8
Blasenberechnung	2
Blasensteine	2
Harnblase	2
Harnblasensteine	2

75543006: *Cerebral embolism*	[18]
Zerebrale Embolie	28
Gehirnembolie	20
Hirnembolie	18
Schlaganfall	10
Brain Embolia	3

4. Discussion

Based on the combination of free, web-based machine MT tools, highly characteristic TC distributions were observed for different configurations of translation engine, input languages, and intermediate languages. The graph in Fig. 1 shows rather distinct profiles of the different scenarios. The mostly parallel decrease of DeepL and Google outputs contrasts with the accumulation of high cardinalities close to the theoretical maximum in Systran, which means that here much less TCs coincide and that clear majority votes cannot be expected. MTP, in contrast peaks at an intermediate level and mostly produces very uneven distributions with (so far qualitatively observed) good translations at least to the two highest ranked TCs, As exemplified in the *Cholelithiasis* and *Urinary bladder stone* examples. The more we move down the list the more translations are either wrong (e.g., "Berechnung", a possible reading for calculus, but not applicable in this context) or less specific (e.g., "Cholelithiasis" without specification). How the other two examples show, there are also more even distributions, which occur, on the one hand, where long terms, naturally allow more variation like in *Closed traumatic dislocation of joint of finger*, or where several, equally good translations exist, like in *Cerebral embolism*. The latter example also shows how the method is able to support the degree of lexical variability that FSN-only translation strategies miss out.

As observed on this limited sample, for most concepts acceptable TC could be derived. For insufficient or erroneous TCs the emergence of discernible patters can be described. Based only on a small and selected sample, the presented results have the potential for bias. So far, we cannot provide metrics on the quality of TCs and consequently a quality-based ranking of TCs. Future research needs to evaluate the potential of MTP to support human terminology translation.

5. Conclusions and Outlook

We present work in progress on using machine translation (MT) for terminology translation, by leveraging several free MT tools fed by different languages and language combinations. A first qualitative analysis was promising and supports our hypothesis that a majority voting applied to many translation candidates yields higher quality results than from one single engine and input language.

Applied to SNOMED CT localisation, shortlists of the output generated by our method, e.g., integrated into existing translation workflows and tools, could provide valuable support to human translators, whose task would be mostly limited to tag the "good" translations instead of starting from scratch. This would render the translation process more cost-effective. In addition, it also finds useful synonyms.

We are currently building a gold standard in form of SNOMED CT translations done by medical experts. This is indispensable for mutually assessing different translation strategies and measuring, for instance, the comparative value of support languages, input languages and translation engines.

References

[1]   Millar J. The Need for a Global Language - SNOMED CT Introduction, Studies in Health Technology and Informatics. 2016;225:683-5.
[2]   Schulz S, Bernhardt-Melischnig J, Kreuzthaler M, Daumke P, Boeker M. Machine vs. human translation of SNOMED CT terms, Studies in Health Technology and Informatics. 2013;192(1):581-4.
[3]   Perez-de-Viñaspre O, Oronoz M. SNOMED CT in a language isolate: an algorithm for a semiautomatic translation. BMC Medical Informatics and Decision Making. 2015;15(2):5.
[4]   Wołk K, Marasek K. Neural-based Machine Translation for Medical Text Domain. Based on European Medicines Agency Leaflet Texts, Procedia Computer Science. 2015;64:2-9.
[5]   Renato A, Castaño J, Williams M, Berinsky H, Gambarte M, Park H, Pérez-Rey D, Otero C, Luna D. A Machine Translation Approach for Medical Terms, Proceedings of the 11th International Joint Conference on Biomedical Engineering Systems and Technologies; 2018;5:369-78.
[6]   Stahlberg F, Neural Machine Translation: A review, Journal of Artificial Intelligence Research. 2020; 69:343-418.
[7]   SNOMED CT Starter Set. Available at: https://www.snomed.org/news-and-events/articles/snomed-ct-starter-set-translation-rfp. Accessed March 24, 2021.
[8]   Google Translate. Available at: https://translate.google.com. Accessed March 24, 2021.
[9]   Deepl Translator. Available at: https://www.deepl.com/translator. Accessed March 24, 2021.
[10] Systran Translator. Available at: https://translate.systran.net. Accessed March 24, 2021.
[11] For a more detailed analysis, refer to the SNOMED CT MTP github repository. Available at: https://github.com/andreaprunotto/SNOMED-CT-MTP. Accessed March 24, 2021.