MEDINFO 2013
C.U. Lehmann et al. (Eds.)

doi:10.3233/978-1-61499-289-9-581

Machine vs. Human Translation of SNOMED CT Terms

Stefan Schulz[a,c], Johannes Bernhardt-Melischnig[a], Markus Kreuzthaler[a], Philipp Daumke[b], Martin Boeker[c]

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

Abstract

Objective: In the context of past and current SNOMED CT translation projects we compare three kinds of SNOMED CT translations from English to German by: (t_1) professional medical translators; (t_2) a free Web-based machine translation service; (t_3) medical students.
Methods: 500 SNOMED CT fully specified names from the (English) International release were randomly selected. Based on this, German translations t_1, t_2, and t_3 were generated. A German and an Austrian physician rated the translations for linguistic correctness and content fidelity.
Results: Kappa for inter-rater reliability was 0.4 for linguistic correctness and 0.23 for content fidelity. Average ratings of linguistic correctness did not differ significantly between human translation scenarios. Content fidelity was rated slightly better for student translators compared to professional translators. Comparing machine to human translation, the linguistic correctness differed about 0.5 scale units in favour of the human translation and about 0.25 regarding content fidelity, equally in favour of the human translation.
Conclusion: The results demonstrate that low-cost translation solutions of medical terms may produce surprisingly good results. Although we would not recommend low-cost translation for producing standardized preferred terms, this approach can be useful for creating additional language-specific entry terms. This may serve several important use cases. We also recommend testing this method to bootstrap a crowdsourcing process, by which term translations are gathered, improved, maintained, and rated by the user community.

Keywords:
SNOMED CT, Translations, Natural Language Processing.

Introduction

Background

The use of terminological standards throughout Europe is increasingly being requested for both routine documentation and secondary use scenarios. SNOMED CT has a great potential to take that role as the most comprehensive medical terminology developed to date [1,2]. Translated versions are important for bringing value-added applications into clinical routine in non-English speaking countries.

Survey of current translation projects

Besides the US and UK versions, a complete (Latin American) Spanish translation of SNOMED CT is being maintained by the International Health Standards Development Organisation (IHTSDO) [3]. Since the creation of this organisation in 2007, the need for translation has become apparent to non-English speaking member countries. Guidelines were elaborated by the Translation Special Interest Group of IHTSDO [4], and three major translation projects were initiated, of which two are complete. One of the first countries to set up a translation project was Denmark [5]. It was completed in 2009. The Swedish translation [6] was completed in 2010. Both countries used the same workflow. The translation of SNOMED CT into Canadian French [7] is ongoing. All SNOMED CT translations vary in quality and extent. For instance, the Danish and Swedish versions are restricted to the translation of preferred terms thus lacking synonymous descriptions.
Other European countries have decided to translate minor parts of SNOMED CT (Belgium, Lithuania, Estonia, Spain), focusing on subsets. The Netherlands have decided to educate clinicians in SNOMED CT before they begin to translate into Dutch, also starting with smaller reference sets needed in the health care sector.
There also exists a German translation, finished in 2004. In 2002, the College of American Pathologists (CAP) had commissioned a Dutch translation company to translate SNOMED CT into German, a process that consumed about 11.5 person years, carried out by nine medical translators for translation and review, enhanced by eight medical doctors in the editorial board. This version has never been released, as for none of the countries in which German is an official language has a roadmap for introducing SNOMED CT. The hesitation is partly due to the opinion that there is not enough evidence for benefit of SNOMED CT in large scale applications [8]. Another reason is the non-uniform attitudes toward SNOMED CT in the German healthcare industry. The German Association of the Healthcare IT Industry (VHitG) confirmed that some healthcare providers, whose products rely on custom terminology solutions, fear that their market position will be weakened with the introduction of SNOMED CT. The partly exhausting experiences with the introduction of the German electronic health card [9,10] currently hamper the discussions about novel technologies.

As the abovementioned translation projects attested, translating a huge terminology like SNOMED CT with more than 300,000 preferred terms is costly both in terms of human and time resources. One could argue that for applications that require less accurate translations, machine translation technology could be employed. The field of machine translation is rapidly developing, mostly due to the availability of huge training data on the web [11]. Free Machine translation interfaces have been offered to WWW users by the main search engine providers, e.g. Google Translate [12], and they have become increasingly popular. For the German language, a group of professional translators recently evaluated the quality of this service on a mixed corpus. The professionals rated the linguistic correctness with the grading of 1.5 (out of 6), corresponding to "good – very good", whereas the correctness of the content was rated 4.5 (out of 6), which is corresponds to "sufficient to poor". The conclusion was that machine translation cannot seriously compete with human translation [13].

Purpose of This Study

The objective of this study is to compare translations of SNOMED CT fully specified names which are provided by professional medical translators, the Web-based general scope translation software Google Translate, and by lay translators (medical students).

Materials and Methods

Preparation of the Translations

The study is based on the international SNOMED CT release from July 2012 and the unofficial German translation from 2004, provided by the IHTSDO for evaluation purposes. In order to ensure comparability with the German version, only =concepts active in either version were considered. Through the assignment of random numbers to each concept we drew a random sample with n=1000 concepts. Half of this sample was used to train translators and raters. Out of the other half, the following term sets were created:

- (t_0) English fully specified names (UK English), extracted from the July 2012 International Release;
- (t_1) Corresponding German fully specified names, extracted from the 2004 unreleased German translation;
- (t_2) Machine translated version using the WWW interface of Google Translate;
- (t_3) Human translated version produced by two German medical students.

For the machine translation (t_2), we submitted the complete list of English terms (one per line, without hierarchy tags) to the Web interface and harvested the translated output. The two medical students, who produced t_3, had received only short instructions. They were told to translate the terms as literally as possible, but equally targeting correct German spelling and grammar. Non-translated words were allowed as long as the English (or Latin) original words were also commonly used in German medical texts. The translators were not given the IHTSDO translation guidelines, because they had not yet officially released in 2004 when t_1 was produced.

Both medical students worked independently. Each of them translated 300 terms, so that 100 terms were translated twice in order to acquire data for computing the inter-translator agreement. Translation disagreements were not settled. Of the terms translated twice the selection was done by random.

Experimental Setting

After harvesting all translations, a table was created in which the three translations (t_1, t_2, t_3) were juxtaposed to the English source in random order. Thus blinded, each translation was submitted to expert rating and rated on a three-point ("green" (3), "yellow" (2), "red" (1)) scale. The rating was done by the 2nd and the 5th author, both physicians. "Green" meant that the translation was fully acceptable, "yellow" meant acceptable with modifications, "red" unacceptable. This explains why we considered a three-point scale as sufficient. Compliance with the IHTSDO translation guidelines was not tested, in order to not introduce a bias that would penalize the 2004 German translation.

Each term was ranked by two criteria, *viz.*:

a. The *linguistic correctness* of the translation. Here, without consideration of the source, the correctness of grammar rules and the vocabulary choice was assessed. Terms with unusual non-translated words, grammar or spelling errors, would diminish the quality judgment, even if the translation is intelligible.
b. The *fidelity* of the translated *content*. Here, the only criterion was the closeness in meaning of the translated term to the meaning of the original. This included non-translated terms or terms belonging to a different language register (e.g. patient language), as long as the term could be expected to be intuitively understandable by an average medical practitioner.

As an outcome variable, the average rating of each translation type was computed for both *linguistic correctness* and *content fidelity*. For the estimation of inter-rater reliability, 150 terms were rated by both raters.

Besides the human comparison, a semantic proximity measurement was performed, using a morphosemantic abstraction tool, which extracts meaning-bearing atomic fragments from medical terms and maps them to a concept-like interlingua covering several European languages [14]. As an output of this process, for each medical term a representation in the form of interlingua set tokens $L = \{l_1, l_2, \ldots l_k, \ldots l_n\}$ was generated.

As a similarity metric we used a modified Jaccard distance metric, in which the sequence of elements is ignored:

$$J = (|\text{Union}(L_a, L_b)| - |\text{Intersection}(L_a, L_b)|) / |\text{Union}(L_a, L_b)|$$

Analysis

All statistical analyses were computed with the statistical package R version 2.15.1 [15]. Exact Fleiss' kappa was calculated for three-categorical inter-rater reliability both for linguistic correctness and content fidelity. For comparison of linguistic correctness and content fidelity between translation groups (t_1, t_2, t_3) ANOVA was used. To estimate the influence of translation group and/or semantic proximity on the rating of linguistic correctness and content fidelity uni- and bivariate regression was calculated.

Results

Translation Time Per Term

The two students needed, on average, 90 seconds per term. Based on a person year of 1432 hours, according to OECD data for Germany, the translation of 360,000 fully specified names (as in the unreleased German version) would then amount to 6.3 person years.

Quality of Translations and Ratings

Prior to the experiment, the concordance of the student translations was semi-quantitatively estimated. The comparison of the translation for the 100 SNOMED CT concepts that were translated twice yielded the following results (see Table 1):

Table 1 – Comparison of terms translated twice (n=100)

Term concordance	Count
No synonymy	11
Close synonymy	20
Complete synonymy	25
Minor differences in spelling and punctuation	17
Verbatim agreement	27

In the scope of the rating experiment, the inter-rater reliability was fair with 0.4 for linguistic correctness and 0.24 for content fidelity (Exact Fleiss' kappa). A possible explanation for the lower kappa on content fidelity might be that the inter-individual variability on the interpretation of near-to-correct or near-to-wrong expressions is higher, due to different intra-individual mechanisms in constructing "meaning" than the more rigid judgment on spelling and grammar for which clear rules exist.

Quantitative Comparison of Translations

Mean ratings and confidence intervals for the three translations t_1 - t_3 are given in Table 2, based on the 3-point scale used for rating. The semantic proximity measurement results obtained by the morphosemantic indexing method are given in the last row, with the values in the interval [0-1].
In terms of linguistic correctness, both human translation scenarios (translation experts vs. medical students) had the same average ratings, which was about half a rating unit higher than the rating of the machine output.

Table 2 – Ratings of content fidelity and linguistic correctness (mean and 95% confidence intervals). All mean differences between translation groups t_1 and t_2, and between t_2 and t_3 (humans translators vs. machine translation) are highly significant ($p<0.001$), except for the semantic proximity on t_2 and t_3 (n.s.). All differences in means between t_1 and t_3 (human translators) are not significant except for content fidelity ($p<0.05$)

	t_1 Professional Translators	t_2 Google Translate	t_3 Medical Students
Linguistic	2.84	2.23	2.84
Correctness	2.80 – 2.88	2.18 – 2.29	2.81 – 2.88
Content	2.78	2.54	2.86
Fidelity	2.75 – 2.84	2.45 – 2.60	2.83 – 2.9
Semantic	0.45	0.52	0.49
Proximity	0.43 – 0.48	0.49 – 0.55	0.45 – 0.51

Analysing the content fidelity, the medical students' translation was rated, on average, with 0.08 slightly higher than expert translators, which is significant at a 95% level. The difference between human and machine translation is smaller than in the linguistic assessment, with a difference of 0.24 (experts vs. machine) and 0.32 (students vs. machine). The values of the semantic proximity were highest for machine translation (t_2), and lowest with regard to expert translation (t_1).

Table 3 – Regression analysis for uni- and bivariate models (LC: linguistic correctness, CF: content fidelity). The group with the professional translators (t_1) is reference category for the categorial variable t. Insignificant models are omitted. In all cases, the semantic distance explains only a small part of the variance.

	β_0	β_{t2}	β_{t3}	β_{sd}	R^2
LC	2.84	-0.61	0.0		0.26
CF	2.8	-0.25	0.07		0.06
	2.62			0.23	0.02
LC	2.82	-0.61	0.0	0.3	0.26
CF	2.7	-0.27	0.06	0.25	0.08

Linear regression analysis was used to estimate the predictive relationship of translation group and / or the semantic distance on the human rating of linguistic correctness and content fidelity. Table 3 shows regression coefficients (β_{var}) and R^2 for uni- and bivariate models. As expected, the univariate model confirms the group analysis from Table 2. However, the bivariate model explains only slightly more of the variance than the univariate model due to the small explanatory power of the semantic distance in the model.

Discussion

In this study we compared SNOMED CT preferred term translations from different sources. Surprisingly, the student translations were rated as being equal in spelling and grammar, and better in content compared to the quality of reviewed professional translations. This advantage might be due to the laypersons' more practically oriented language shaped by daily communication needs.
The comparison of human translations to the output of a machine translation engine that had not been specifically trained in medical texts is equally astonishing. Whereas the machine-translated result was 0.5 points lower regarding language correctness, the difference in content intelligibility was only half of this. This finding contradicts the result reported in [13], where the language aspects of machine translations were rated as much more acceptable than the content aspects. This may be explained by the fact that the translation of terms is generally less error-prone than the translation of normal text.
The semantic proximity measure was lowest for the professional translation (t_1) and highest for the machine translation (t_2). The latter may support the acceptability of the machine translations, but should not be over-interpreted. The lower result for t_1 may reflect the professional translators' tendency towards more idiomatic translations.

Limitations of the Study

In its current layout this studies has several limitations, such as:

- The sample size is small for a rating experiment in which only small differences between groups are present, while the variance is high. Nevertheless, we succeeded in showing the significance of 0.08 units of difference.

- The same applies to the number of raters. Only two raters cannot represent a diversity of medical professions and background knowledge, which influences the estimation of linguistic correctness and even more of content fidelity.
- The weak definitions of quality criteria for judging content. The inter-rater reliability was much lower than for linguistic correctness. We assume to gain more homogenous results between raters providing more precise guidelines. On the other hand, the very open rating on content fidelity in this experiment might better approximate the flexibility between humans in the "understanding" of language and is thus more generalizable in reality.
- The sample size was too small to stratify the experiment along semantic tags (disorders, procedures, substances, organisms etc.) for which we observed various degrees of variation and translation quality. For instance, there was no need to translate Linnaean terms (e.g. Escherichia coli), as all of them were in Latin.

Conclusions and Further Research

Machine translation and the employment of student translators are considerable alternatives to facilitate the translation of standardized medical terms. However, this approach must be very carefully pondered and should be limited to certain uses that are not prone to affect patient safety. Both methods are not acceptable for the production of fully specified names that can be considered a terminological standard for the target language. However, our "suboptimal" translations can be regarded as additional (quasi-)synonymous descriptions, and may therefore be useful for content retrieval or semantic annotation of clinical texts. For the use in data acquisition forms the corresponding English fully specified name or preferred term should remain visible. Here, their use should be restricted to user groups that understand English medical terms.
The combination of machine-translated text with subsequent post-editing by humans could be another translation strategy that reduces time and produces quality translations.
An interesting route to be further explored could be the use of crowdsourcing mechanisms for terminology maintenance. Here, the English standard term would be displayed together with a machine-generated translation. Users can then add alternative translations and rate the quality of translations. In additional iterations, bad terms could be ruled out and the best translations would be determined. By this method, a rich repository of synonymous term variants would evolve, which could become an important resource for natural language processing on medical texts. We also hypothesise that SNOMED CT concepts for which a great variation of terms with rather low rating results observed are ambiguous ones, which should be submitted to quality review.
Finally, a larger study would allow us to measure translation quality stratified by SNOMED CT subhierarchies.

Acknowledgments

We thank our student translators Gerrit Merkel and Moritz Mohr. We are grateful to Lene Vistisen for information on the SNOMED CT translation processes in Denmark and Sweden. SNOMED CT related activities of the main author were supported by the German HL7 User Group.

References

[1] Ciolko E, Lu F, Joshi A. Intelligent Clinical Decision Support Systems based on SNOMED CT. Conference Proceedings of the IEEE Engineering in Medicine & Biology Society 2010; 1: 6781-6784.

[2] Nadkarni PM. Drug safety surveillance using de-identified EMR and claims data: issues and challenges. J Am Med Inform Assoc. 2010; 17 (6): 671-674.

[3] Reynoso GA et al. Development of the Spanish version of the Systematized Nomenclature of Medicine: methodology and main issues. Proceedings of the 2000 AMIA Symposium: 694-698.

[4] Translating SNOMED CT. IHTSDO – International health Terminology Standards Development Organisation. http://www.ihtsdo.org/develop/documents/translating-snomed-ct/. Last accessed: Mar 31, 2013.

[5] Høy A. Coming to Terms with SNOMED CT® Terms: Linguistic and Terminological Issues Related to the Trans-lation into Danish. In: Budin G, Laurén C, Picht H et al., Terminology Science and Research, 2006.

[6] Klein GO, Chen R. Translation of SNOMED CT - strategies and description of a pilot project. Studies in Health Technology and Informatics. 2009; 146: 673-677.

[7] Canadian NPC: Translation of SNOMED CT® - Approaches, Challenges and Lessons Learned. IHTSDO Spring 2009 Pre-Conference Workshop.

[8] Cornet R, de Keizer N. Forty years of SNOMED: a literature review. BMC Med. Inform. and Decision Making 2008 Oct 27; 8 Suppl 1: S2.

[9] Schmidt K. Test procedure for the electronic medical card. More and more physicians refuse. MMW Fortsch. der Med. 2007 Apr 5; 149(14): 55-56.

[10] Stoschek J. Electronic health card. This will be expensive for physicians. MMW Fortsch. der Med. 2007 Jan 25; 149(4): 56.

[11] Koehn P. Statistical Machine Translation. Cambridge University Press, 2010.

[12] Google Translate. http://translate.google.com/. Last accessed: Mar 31, 2013.

[13] Bundesverband der Dolmetscher und Übersetzer e.V. Übersetzerverband testet Google Translate (press release, German). http://www.bdue-bayrn.de/fileadmin/bdue/Pressemiteilugen/09.10.2012_Uebersetzerverband_testet_Google_Translate_Pressemitteilung.pdf. Last accessed: Mar 31, 2013.

[14] Daumke P, Schulz S, Müller ML, Dzeyk W, Prinzen L, Pacheco EJ. Subword-based semantic retrieval of clinical and bibliographic documents. Methods of Information in Medicine, 49:141–147, 2010.

[15] R Development Core Team. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing.

Address for correspondence

Stefan Schulz
Institut für Medizinische Informatik, Statistik und Dokumentation
Medizinische Universität Graz
Auenbruggerplatz 2/V
8036 Graz (Austria)
+43 (0) 316 385 13201
stefan.schulz@medunigraz.at