An ontology-based approach for SNOMED CT translation

Mário J. Silva, Tiago Chaves and Bárbara Simões
Instituto Superior Técnico, Universidade de Lisboa and INESC-ID, Portugal

ABSTRACT

SNOMED CT is a comprehensive multilingual class hierarchy of medical terms used in clinical records. Few translations are available, but, as new concepts and revisions are continuously being added, the manual translation and revision of the terms will remain a major endeavour. We propose a new approach for translating SNOMED CT terms (or named entities) using ontology mapping methods and various existing multilingual resources with translated concepts. Our purpose is generating initial candidate translations, already close to those proposed by medical experts, to be later used in a curated translation process. Our method for automatically translating SNOMED CT is being developed for Portuguese, using DBPedia, ICD-9 and Google Translate as sources of candidate translations of the clinical terms, which could later be verified. Initial results, using a manually translated Portuguese catalog of allergies and adverse reactions (CPARA) to SNOMED CT as ground truth, show that it has high potential.

1 INTRODUCTION

SNOMED Clinical Terms[1], or SNOMED CT, is a comprehensive multilingual class hierarchy of terms used in clinical records, with extensive overlapping and synonymous descriptions. The primary purpose of SNOMED CT is to encode the meanings of the terminology used in health information, supporting the effective clinical recording of data with the aim of improving patient care. SNOMED CT provides the core general terminology for electronic health records. With about 300,000 active terms, SNOMED CT spans clinical findings, symptoms, diagnoses, procedures, body structures, organisms and other etiologies, substances, pharmaceuticals, devices and specimen.

The need to interchange medical records across states is demanding the development of faster methods to obtain approved, standards-based, translations of medical records, in particular SNOMED CT. The standardisation of clinical terms and their translations to other languages is very important for the unification of the electronic health records worldwide. However, the manual translation and revision of the terms, synonyms and definitions to a new language is a major endeavour. SNOMED CT is presently available in US and UK English, Spanish, Danish and Swedish. It is also being translated to several other languages, but there is no translation of SNOMED CT to Portuguese or an official initiative to develop and maintain that translation. Hence, a tool to automatically translate SNOMED CT to Portuguese would assist in the production of a release to be validated and improved in a subsequent step at a much lower cost than conducting the process manually.

As new trannslations, concepts and revisions are continuously being added, the manual translation and revision of the terms will remain a major endeavour. This paper describes our work on the development of an automatic translator of SNOMED CT to Portuguese as an assisting tool that could be used for the production of a future standard translation of SNOMED CT. We take the approach of using available classifications and automatic translation services as ontologies that can be aligned and later navigated to provide the translations of such technical terms. In our method, we start by identifying existing alignments between SNOMED CT and other selected ontologies, including the releases of SNOMED CT in different languages. For the Portuguese translation, given its proximity to Spanish, many terms in the Spanish release of SNOMED CT have almost identical spelling. There are other medical terminologies for which multiple translations have been published, such as ICD (International Classification of Diseases) [2]. Another major resource is DBPedia, an ontology derived from Wikipedia, which is very rich in medical terms (Lehmann *et al.*, 2015). After the collection of these multilingual ontologies and published mappings between their terms, we derive additional alignments using the ontology mapping algorithms implemented in AgreementMakerLight, a scalable automated ontology matching system developed primarily for the life sciences domain Faria *et al.* (2013). To obtain correspondences between terms in two distinct languages, we can also explore online translation services, such as the Google Translate Service[3] or Microsoft translator[4] to generate additional mappings.

To show the feasibility of the above outlined approach for automatically generating translations of SNOMED CT terms to Portuguese, we evaluated the translations obtained with the alignments against the translations of a set of SNOMED CT terms that have been mapped by medical experts to terms of the Portuguese catalog of allergies and adverse reactions (SPMS, 2015). The latest release of CPARA includes curated translations of SNOMED CT terms. The evaluation shows promising results. The ontology-mapping translation method achieved an accuracy of 89% and coverage of 37% for the set of 191 terms on the translation of the CPARA vocabulary terms previously hand-mapped to SNOMED CT (using case-insensitive string comparison).

2 RESOURCES AND RELATED WORK

In our work, we used the the 01/2014 International (English) distribution of SNOMED CT and the Spanish version dated from April 2014, both provided by the NLM (National Library of Medicine) institutional site [5]. The distribution also includes a mapping between ICD-9, a WHO classification of diseases, and SNOMED CT. This mapping can be used to link SNOMED CT

[1] http://www.ihtsdo.org/snomed-ct

[2] http://www.who.int/classifications/icd/en/

[3] https://translate.google.com/

[4] http://www.microsoft.com/translator/translator-api.aspx

[5] http://www.nlm.nih.gov/research/umls/Snomed/snomed_main.html

codes to the ICD-9 Portuguese terms in a translation provided by the Portuguese Ministry of Health [6].

There are no comprehensive medical terminologies for European Portuguese. In addition to ICD-9 in European Portuguese, ICD-10 has been manually translated to Brazilian Portuguese [7]. There is also an English to Brazilian Portuguese dictionary of medical terms(Stedman, 2003). However, there are a number of terminological differences between these two variants of the language. Other terminologies, such as ICPC [8], have been translated [9], but they have a much narrower scope than ICD.

In computing, the translation of a terminology, such as the set of SNOMED CT terms, is an instance of a common task in Natural Language Processing (NLP) , designated as Named Entity Translation Ling *et al.* (2011). The task is formulated as the problem of, given a set of labels (named entities) in a source language, obtaining the translations of these entities in a target language. Langlais *et al.* (2008) researched the translation of medical terms using a bilingual lexicon. Recently,Abdoune *et al.* (2013) performed an automatic translation of the CORE subset of SNOMED CT to French by mapping this subset to four French-translated terminologies integrated in the UMLS Metathesaurus: SNOMED international, ICD10, MedDRA and MeSH. They were able to map 89% of the preferred terms of the CORE Subset of SNOMED CT with at least one preferred term in one of the four terminologies.

Other approaches for generating translations have been attempted. Algorithms based on linguistic rules are particularly useful for languages which are poor in language resources, like a recently proposed Basque semi-automatic translation of SNOMED CT (Perez-de Viñaspre and Oronoz, 2014). The algorithm takes an incremental approach: first a lexical translation is attempted; then if a translation is not found, generation/transcription-rules for terms, or chunk-level generation to translate a term token by token are used; finally, a rule-based automatic translation system is used to find a translation.

In this work, we explore DBPedia, an ontology derived from Wikipedia, as an alternative source of term translations (Lehmann *et al.*, 2015). We apply ontology matching methods to align DBPedia and SNOMED CT, along with other web-based services, like Google Translate. The DBPedia is a potentially rich resource for medical terms mappings, given that the English and Portuguese Wikipedias are among the largest. To map these ontologies we used AgreementMakerLight, an ontology matching system developed to tackle large ontology matching problems, and focused in particular on the biomedical domain (Faria *et al.*, 2013). This system can handle the mapping of very large ontologies, as it is the case with SNOMED CT and DBPedia. AgreementMakerLight is derived from the AgreementMaker ontology matching systems (Cruz *et al.*, 2009). The alignments produced by AgreementMaker combine multiple matching algorithms, in three layers: the first layer uses string matching methods to identify similar labels, the second matches ontology structures, and the third layer combines the results from the matchers in the first two layers. The initial experiments reported in this paper only used the first layer algorithms to perform the alignments.

Medical terms, like named entities in general, can matched using similarity metrics like the Jaro distance, initially proposed for record linkage systems (Porter and Winkler, 1997). The Jaro distance has been used for the evaluation of automatic translations of named entitites. It accounts the number of transpositions between two input strings and also the number of different characters, resulting in a numeric distance in the $[0, 1]$ range.

3 SNOMED CT TRANSLATION

Given that SNOMED CT is mostly used to provide terminology for electronic health records, the risks of using an automatically generated translation of such large collection of terms without expert validation are unacceptable. In fact, the SNOMED publisher provides detailed guidelines for validating the translations made by medical experts for the official translations available (IHTSDO, 2012). However, we believe that, if the initial quality of the automatically generated translation is high, we could later validate such candidate translations through a crowdsourcing activity, as experimented by Schulz *et al.* (2013).

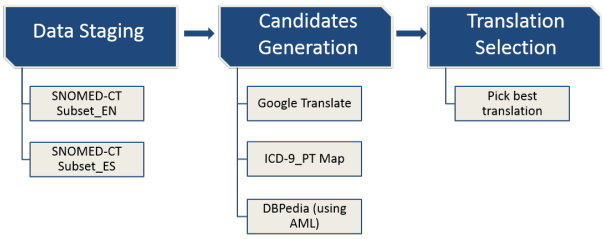

Fig. 1. The translation of SNOMED CT is preceded by a data staging phase. Once the data is prepared, translation is carried out using the implemented methods. We select the best translation candidate using an ensemble model trained that selects the best method for each class of SNOMED CT terms, based on known translations

Our approach for generating the translations of SNOMED CT terms into Portuguese is illustrated in Figure 1. We start by organising two mappings:

1. SNOMED CT to ICD-9: a correspondence between the codes of SNOMED CT and codes and descriptions of ICD-9.
2. SNOMED CT to DBPEDIA: a correspondence between SNOMED CT codes and DBPedia (English and Portuguese) page URIs, and associated page titles.

The first mapping is derived the SNOMED CT to ICD-9 mapping included in the from the UMLS distribution, which includes the correspondence between SNOMED CT and ICD-9 codes. For the second mapping, the matching algorithms implemented in AgreementMakerLight can generate an alignment between SNOMED CT terms and English DBPedia labels. Once this alignment is generated, we can map SNOMED CT codes to DBPedia URIs and then obtain the corresponding label for the Portuguese term by a simple lookup.

To obtain candidate translations for SNOMED CT terms, we implemented four translation methods:

[6] http://www.acss.min-saude.pt/Portals/0/ICD9CMOut2013.xlsx

[7] http://www.datasus.gov.br/cid10/V2008/cid10.htm

[8] http://goo.gl/IX9mqT

[9] http://icpc2.danielpinto.net/

1. Google Translate EN: the candidate translation into Portuguese of each English term in SNOMED CT is provided by the GoogleTranslate API service.
2. Google Translate ES: identical to the above, but the translation service uses the Spanish term as input.
3. ICD-9 Mapping: for a given SNOMED CT term in English, we take the corresponding code and lookup the SNOMED CT to ICD-9 mapping in the UMLS distribution to obtain the ICD-9 code and next the term description in the Portuguese version of ICD-9. This description becomes the candidate translation of the SNOMED CT term to Portuguese.
4. DBPedia Mapping: starting with a SNOMED CT term in English, we lookup the code on the SNOMED CT to DBPedia mapping and, from there, obtain the available candidate translation on the Portuguese DBPedia.

DBPedia is too big to be fully mapped in one batch with limited computing power, given the size of the ontologies involved. This would make the time required by AgreementMakerLight to align SNOMED CT with the full DBPedia prohibitive. However, it is unnecessary, given that most of DBPedia is irrelevant to the clinical domain, to use the full DBPedia. We expect that our users, domain experts in clinical specialisations, will select a batch of SNOMED CT terms of their interest at a time and create/revise the translations of the terms in that smaller set. For instance, to identify a set of allergy-related DBPedia pages to be aligned with a set of SNOMED CT terms, we used the UNIX grep tool to filter out of the DBPedia ontology every page with a label not containing any of the words of the SNOMED CT terms. This resulted in a size reduction from 2 GB to 12MB. To obtain the alignment with DBPedia, we parameterized AgreementMakerLight to consider as aligned all pairs of terms with a Jaro Distance ≥ 0.5.

The last step in our method involves the application of an ensemble learning algorithm (Dietterich, 2000). Each SNOMED CT term has a class label, provided as "qualifier" in the term description. For instance, the SNOMED CT term with code `158965000` has the term "Medical practitioner (occupation)", from which we can separate the description "Medical practitioner" and class *Occupation*. Instead of choosing the best overall translation method, we identify the best translation method for each class, based on the validated translations. As this number will increase over time, we expect that ensemble learning will in the end improve the automatic translation process. However, given the small number of validated and translated terms in Portuguese that we have at this time, we still lack reliable data to evaluate this step.

4 EVALUATION

CPARA, *Catálogo Português de Alergias e Reacções Adversas*, is a list of terms related to Allergies and Adverse Reactions in use in the Portuguese National Health Service (SPMS, 2015). It was developed with the goal of unifying the classification for allergies and adverse reactions in Portugal. Given the high levels of patient mobility, physicians frequently need to know precisely which substances are known to affect an international patient. To address this need, CPARA terms have been mapped to SNOMED CT terms by a group of experts. These experts also created European Portuguese translations of the SNOMED CT Common Terms and Fully Specified Names (FSN) in the CPARA catalog. This mapping is critical to making the medical information exchanged about patients who travel internationally more accurate. In our evaluation, we used the Common Terms translations as gold standard to assess the accuracy of our translation approach. CPARA includes 191 codes and common terms of the US English distribution of SNOMED CT, and the corresponding CPARA codes and terms. In the Spanish SNOMED CT distribution there are 192 terms mapped from these 191 codes (one code is mapped to two terms).

Evaluation of the translated SNOMED CT terms started with candidate translations for the allergy-related SNOMED CT codes in CPARA generated by application of our method. We then evaluated the resulting set of translations against the ground truth composed by the corresponding CPARA terms as defined by the medical committee that defined the mapping. To assess the accuracy of the evaluated translation methods, we scored each term translation by the Jaro distance between the automatically translated term and the CPARA translation. The Jaro distance (JD) between two strings is 1 if the strings have the exact same number of characters and do not have any transposition-[10].

Prior to computing Jaro distances all the translation candidates and CPARA translations were normalised: we removed any qualifiers from the SNOMED CT candidates, deleted quotes from the CPARA translations, and converted all the named entities to lowercase (e.g, "Moderate (severity modifier) (qualifier value)" became "moderate" and "Contact metal agent (substance)" became "contact metal agent"). These preparatory steps are necessary to obtain meaningful similarity metrics, because these qualifiers are common to many terms and can be translated independently. In addition, the Jaro Distance considers the same letter in lowercase and uppercase forms as two distinct characters. The statistics of the translations obtained with each of the four implemented methods described in the previous section are given in Table 1. In these statistics, we considered as valid the translations with $JD = 1$.

Method	Source Language	Coverage	#Method	AVG JD	STDEV JD
GT	EN	100%	191	0.78	0.22
GT	ES	114%	218	0.58	0.15
ICD 9	EN	10%	20	0.61	0.12
DBPedia	EN	37%	70	0.89	0.03

Table 1. Global Results for all translation methods with the respective average Jaro Distance (AVG JD) and Standard Deviation Jaro Distance (STDEV JD). The implemented methods are Google Translate (GT), both from English (EN) and Spanish (ES) to Portuguese, ICD-9 Mapping (ICD 9), and DBPedia Mapping (DBPedia). All translations were attempted with two source languages, English (EN) and Spanish (ES). The number of terms translated by each method is given in the # Method column.

We observe that the SNOMED-DBPedia alignment obtains, for a coverage of 37%, both the highest similarity (0.89) and lowest standard deviation (0.03). This shows that we have been able to accurately translate a set of SNOMED CT terms to Portuguese, using basic alignment techniques, through the SNOMED CT to DBPEDIA alignment. However, the generation of translations

[10] The computation of the Jaro distances was made with the Python Jellyfish library https://pypi.python.org/pypi/jellyfish

based on ontology alignments as proposed in this paper also has limitations. In particular, only a fraction of the translations can be obtained by this method, while Google Translate always proposed a translation. Our success with Portuguese may not be granted when aligning SNOMED CT with DBPedia in other languages with smaller Wikipedias.

Google Translate EN showed better accuracy than Google Translate ES. This result was not initially expected, because Spanish and Portuguese are close languages. This may result from the CPARA terms being originally derived from the English terminology. The number of translations obtained with Google Translate ES is higher than the number of terms in the CPARA dataset (yielding the 114% coverage). This is the result of how we have obtained the Spanish SNOMED CT candidate terms for translation. We started from the same initial SNOMED CT codes that we used for the English translation and obtained the Spanish codes matching the *concept_id* and *type_id* of the initial English terms. This generated a higher number of ES candidate terms to translate (218) then the initial EN terms (191).

To evaluate which translation method works best for each class of SNOMED CT terms, we measure which translation method performs best in each class. This method is necessary to later model an ensemble learning stage that could pick the best method for each class. To obtain the results, we divided CPARA in classes for translation purposes. These classes were extracted from the qualifiers defined for the SNOMED full specified name terms. We were interested in observing translation performance differences across classes. To measure the differences, we calculated the similarity and standard deviation as above of all the translation candidates in each class. The results are summarised in Table 2.

The SNOMED-DBPedia alignment generates better translations for all classes, except *Person* and *Qualifier Value*. The poorer performance could, however, reflect that only a small number of related identified terms in the allergy domain have been identified for both classes.

Google TranslateES has better average similarity for the *Person* class than Google Translate EN. This shows that the SNOMED CT translation from Spanish could benefit from using the Spanish language distribution for some CPARA translations.

The translations obtained with the ICD-9 mapping translator are worse than obtained by Google Translate (for both languages). This results from ICD-9 being less comprehensive than SNOMED CT. ICD codes mostly diseases, symptoms or causes of death. Therefore, many of the CPARA terms in SNOMED CT were absent in the ICD-9 to SNOMED CT mapping. The results also indicate that, as expected and observed with ICD-9, terminologies of narrower scope are not useful for translating clinical terms through ontology alignment. The ICD-9 mapping is much less successful than other resources, such as DBPedia and Google Translate, which can provide much higher coverage of candidate translations, in many cases while retaining equal or better accuracy. The ICD-9 mapping method generates a high amount of 1-to-many matchings. However, ICD-9 could still be useful in cases where it generates only one matching description, which is usually very accurate and reliable, attending that matchings between ICD and SNOMED CT and the resulting translations are validated by medical experts.

5 CONCLUSIONS AND FUTURE WORK

SNOMED CT is increasingly prevalent in the heath care sector, resulting from the increasing need to exchange medical records in mobile societies. There is also a growing general interest in accessing standardised machine-readable medical records for improving managed heath care and biomedical research.

We introduced a new methodology for translating SNOMED CT terms, which relies primarily on aligning large ontologies, complementing language-based methods that have been proposed before. We prototyped an initial implementation of this methodology, which obtained high coverage and good accuracy, despite only using string matchers for SNOMED CT and DBPedia alignment along with the domain-independent Google Translator. A translation was considered valid when the expert mapping of an allergy-related SNOMED CT term to Portuguese is identical to the obtained using the SNOMED CT to DBPedia alignment. The accuracy under these settings was 37%. This shows that both the English and Portuguese versions of DBPedia are rich and accurately interlink with medical terms. However, the results for Portuguese may not be indicative of how this method would perform on other languages. Portuguese is one of the top-10 Wikipedia languages in terms of the total number of entries. The coverage of the obtained translations depend on how rich the Wikipedia for a target language is in covering clinical concepts and the extent to which these concepts are mapped to Wikipedia pages in languages for which a SNOMED CT translation exists. In addition, our validation experiment was confined to testing about 200 SNOMED CT Common Terms in the alergies and adverse reactions domain in European Portuguese. It is still unknown how comprehensive and accurate the English and Portuguese Wikipedias are across the full clinical domain, and how this factor affects the accuracy of the SNOMED CT translations.

Some improvements can still be added to the software implementing the presented translation method. For instance, the SNOMED CT to DBPedia alignment should explore the defined semantic relationships between classes and terms in both SNOMED CT and DBPedia. On the other hand, these relationships could be explored to generate accurate translations for untranslated terms in lexical methods to be provided. For this purpose, language resources, such as WordNet, and parallel corpora of named entities, such as previously validated SNOMED CT translations, could be used to learn how words and multi-word expressions should be properly translated.

The measured accuracy of our translation method could still be significantly increased without sacrificing the quality of translations, by relaxing the similarity threshold. The negative impacts of such relaxation are negligible, given that the generated translations will always need to be validated by experts before used in a clinical context. The expert-validation step presently relies on the review of generated translations presented on spreadsheets. A crowdsourcing platform could speed-up the process of creating and maintaining a validated translation of SNOMED CT. Moreover, active learning could also be incorporated in the crowdsourcing platform, leading to fast improvement of the proposed translations as the validated translation can also be used as input to generate good candidates (Ambati *et al.*, 2010).

A complementary assessment of the alignment approach proposed here could be obtained by applying it to the automatic translation with one of the existing released translations, e.g.

Spanish. However, given that we rely on alignments bewteen lexical resources we are not certain if the Wikipedia correspondences between clinical term pages in Spanish and English have been created based on SNOMED CT.

ACKNOWLEDGEMENTS

We thank Daniel Faria and the other members of the SOMER project for help with running AgreementMakerLight and their feedback. We also thank Dr. Anabela Santos for the help with the CPARA translation of SNOMED CT to validate our tool, and Bruno Martins for the pointers to previous works. This work was partially supported by Fundação para a Ciência e a Tecnologia (FCT), grants PTDC/EIA-EIA/119119/2010 (SOMER), UID/CEC/50021/2013 and EXCL/EEI- ESS/0257/2012 (DataStorm).

REFERENCES

Abdoune, H., Merabti, T., Darmoni, S. J., and Joubert, M. (2013). Assisting the translation of the core subset of snomed ct into french. *Studies in health technology and informatics*, **169**, 819–823. DOI:10.3233/978-1-60750-806-9-819.

Ambati, V., Vogel, S., and Carbonell, J. (2010). Active learning and crowd-sourcing for machine translation. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*.

Cruz, I. F., Antonelli, F. P., and Stroe, C. (2009). Agreementmaker: Efficient matching for large real-world schemas and ontologies. *PVLDB*, **2**(2), 1586–1589.

Dietterich, T. (2000). Ensemble methods in machine learning. In *Multiple Classifier Systems*, volume 1857 of *Lecture Notes in Computer Science*, pages 1–15. Springer Berlin Heidelberg.

Faria, D., Pesquita, C., Santos, E., Palmonari, M., Cruz, I., and Couto, F. (2013). The agreement maker light ontology matching system. In *On the Move to Meaningful Internet Systems: OTM 2013 Conferences—Confederated International Conferences*, number 8185 in Lecture Notes in Computer Science, pages 527—541. Springer.

IHTSDO (2012). *Guidelines for Management of Translation of SNOMED CT*. IHTSDO - International Health Terminology Standards Development Organisation.

Langlais, P., Yvon, F., and Zweigenbaum, P. (2008). Analogical translation of medical words in different languages. In B. Nordström and A. Ranta, editors, *Advances in Natural Language Processing*, volume 5221 of *Lecture Notes in Computer Science*, pages 284–295. Springer Berlin Heidelberg.

Lehmann, J., Isele, R., Jakob, M., Jentzsch, A., Kontokostas, D., Mendes, P. N., Hellmann, S., Morsey, M., van Kleef, P., Auer, S., and Bizer, C. (2015). DBpedia - a large-scale, multilingual knowledge base extracted from wikipedia. *Semantic Web Journal*, **6**(2), 167–195.

Ling, W., Calado, P., Martins, B., Trancoso, I., Black, A., and Coheur, L. (2011). Named entity translation using anchor texts. In *The International Workshop on Spoken Language Translation (IWSLT)*.

Perez-de Viñaspre, O. and Oronoz, M. (2014). Translating snomed ct terminology into a minor language. In *Proceedings of the 5th International Workshop on Health Text Mining and Information Analysis (Louhi)*, pages 38–45, Gothenburg, Sweden. Association for Computational Linguistics.

Porter, E. H. and Winkler, W. E. (1997). Approximate string comparison and its effect on an advanced record linkage system. In *Advanced Record Linkage System. U.S. Bureau of the Census, Research Report*, pages 190–199.

Schulz, S., Bernhardt-Melischnig, J., Kreuzthaler, M., Daumkea, P., and Boeker, M. (2013). Machine vs. human translation of SNOMED CT terms. In *MEDINFO 2013*.

SPMS (2015). CPARA – catálogo português de alergias e outras reações adversas / portuguese catalogue of allergies and other adverse reactions. Technical Report V3.0, 09-03-2015, SPMS – Serviços Partilhados do Ministério da Saúde. `http://tinyurl.com/me5jhq7`,`http://tinyurl.com/lehlhaa`.

Stedman, T. L. (2003). *Stedman's English to Portuguese and Portuguese to English Medical Dictionary*. French & European Pubns. ISBN 13: 9780785975281.

Translation Technique	Source Lang.	Class	AVG JD	STDEV JD
Google Translate	EN	Substance	0.82	0.19
		Observable Entity	0.74	NA
		Product	0.96	0.01
		Disorder	0.78	0.25
		Occupation	1.00	0.00
		Person	0.55	0.18
		Qualifier Value	0.79	0.17
		Finding	0.74	0.30
		Event	1.00	NA
		Situation	0.71	NA
		Organism	0.63	0.37
		Severity Modifier	0.79	0.30
		Contextual Qualifier	0.83	0.15
		No Qualifier	0.63	0.28
	ES	Disorder	0.66	0.11
		Substance	0.59	0.14
		Qualifier Value	0.58	0.11
		Contextual Qualifier	0.56	0.08
		Organism	0.62	0.10
		Person	0.57	0.09
		Occupation	0.67	0.04
		Finding	0.67	0.12
		Situation	0.60	0.00
		Observable Entity	0.64	0.00
		Product	0.57	0.03
		Severity Modifier	0.53	0.13
		Event	0.35	NA
		No Qualifier	0.50	0.28
ICD-9	EN	Disorder	0.60	0.12
		Finding	0.68	0.15
DBPedia Matching	EN	Disorder	0.92	0.04
		Substance	0.90	0.03
		Qualifier Value	0.72	0.05
		Event	1.00	NA
		Finding	0.99	0.00
		Organism	0.82	0.09
		Person	0.48	NA
		No Qualifier	0.83	0.09

Table 2. Scores for the different classes of SNOMED CT terms. AVG and STDEV JD column represent the averade and standard deviation of the Jaro Distance; NA indicates that STDEV cannot be obtained because there is only one translation for the class.