Acquisition of Character Translation Rules for Supporting SNOMED CT Localizations

Jose Antonio Miñarro-Giménez, Johannes Hellrich, and Stefan Schulz

Institute of Medical Informatics, Statistics, and Documentation, Medical University of Graz, Austria

Jena University Language & Information Engineering (JULIE) Lab, Friedrich-Schiller-Universität Jena, Jena, Germany

Abstract. Translating huge medical terminologies like SNOMED CT is costly and time consuming. We present a methodology that acquires substring substitution rules for single words, based on the known similarity between medical words and their translations, due to their common Latin / Greek origin. Character translation rules are automatically acquired from pairs of English words and their automated translations to German. Using a training set with single words extracted from SNOMED CT as input we obtained a list of 268 translation rules. The evaluation of these rules improved the translation of 60% of words compared to Google Translate and 55% of translated words that exactly match the right translations. On a subset of words where machine translation had failed, our method improves translation in 56% of cases, with 27% exactly matching the gold standard.

Keywords. Translations, Terminology as Topic, Natural Language Processing

Introduction

Interoperability of health data across languages requires multilingual terminologies, which demands efforts to produce term translations. Machine translation combined with manual curation could speed up the translation process and reduce its costs. Although statistical machine translation has mostly supplanted rule-based approaches [1], the latter ones are still method of choice for transliterating words between different writing systems [2], targeting proper names or technical terms, which share their morphological building blocks across languages. An advantage of rule-based approaches is that rules can be manually checked. This is especially relevant for the medical domain, where erroneous translations raise patient safety issues, causing a general mistrust towards machine translation [3]. Transliteration means, in a narrow sense, the application of rules replacing characters from one writing system into another one, e.g. the geographic name “Venezia” from the Latin writing system into Cyrillic (“Венеција”) or Greek (“Βενετία”) one. In a broader sense, transliteration is understood as including rules that adapt terms or names to different languages that use the same writing system, in our example producing translations such as “Venice” (English), or “Wenecja” (Polish). Apart from the translation of geographical names, such an approach seems promising for the creation of biomedical terms, as their morphological inventory is largely borrowed from Latin or Greek. These internationalisms were used as features for other term translation systems, yet based on

1 Corresponding Author.
manually constructed morpheme mappings [4]. Translating complex words, like compounds, in a piecemeal fashion is a general strategy for increasing coverage and accommodating newly formed words [5, 6]. In the following we will present a rule-based approach for translating content of the clinical terminology SNOMED CT.

1. Methods

Three data sets, named Training, Testing and Evaluation (gold standard), were created on a list of 89,469 German – English word pairs. This resource had been created by collecting all English terms (fully specified names, preferred terms, synonyms) from SNOMED CT (Intl. release, July 2013). The terms were tokenized and the tokens sorted by decreasing frequency. The tokenization of SNOMED CT terms divided them into word or tokens that contains any set of consecutive word characters [a-zA-Z0-9] delimited by a non-word character, e.g. whitespace character, or punctuation mark. All tokens were automatically translated using Google Translate 2, which was chosen because of its availability. The translations of all tokens that occurred more than once were curated by an expert. For our study, these tokens and translations were randomly distributed across the three data sets in which Training and Testing contain the machine translations (excluding non-translated words), whereas the Evaluation data set contains the manually validated translations.

Our methodology consists of three steps: (1) gathering rule candidates from Training; (2) extract the most relevant translation rules; and (3) testing, filtering and ranking rules based on their profit. Table 1 describes the first step. The algorithm extracts all substring (i.e. strings that are part of other strings) combinations from the source and target words of a translation in the Training data set in order to identify which ones improve the translation, measured using the Levenshtein edit distance [7] and compared against Testing dataset. For example, the English / German word pair “pelvectomy” and “pelvektomie” produces, among others, the rules (“ect” → “ekt”), (“my_” → “mie_”), (“ectomy_” → “ektomie_”). Rule candidates such as (“_pel” → “_pel”) or (“_pel” → “_pe”), do not improve the translation and, therefore, are discarded. Word boundaries are marked by the “_” character.

Table 1. Pseudo code of the algorithm for substring translation rules. The function length retrieves the length of the string. The function distance corresponds to the Levenstheins’ distance. It is the minimum number of character insertions, deletions or modifications needed to transform one word to another. The values min and max represents the allowed minimum and maximum substring sizes to tame combinatorial explosion.

<table>
<thead>
<tr>
<th>For each source: S and target: T words in training set:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Obtain all substrings Ss of S where length(Ss) &gt;= min AND length(Ss) &lt;= max</td>
</tr>
<tr>
<td>Obtain all substrings Ts of T where length(Ts) &gt;= min AND length(Ts) &lt;= max</td>
</tr>
<tr>
<td>For each pair of obtained substrings: (Ss, Ts)</td>
</tr>
<tr>
<td>Transform S with the pair (Ss, Ts) to obtain tS</td>
</tr>
<tr>
<td>If distance(S, T) &gt;= distance(S, tS) + distance(Ss, Ts) then save (Ss, Ts) rule.</td>
</tr>
<tr>
<td>Return all saved (Ss, Ts) rules.</td>
</tr>
</tbody>
</table>

The substring combinations are reduced by (1) max and min substring length, (2) max length difference between source and target substring, and (3) the limitation to those rule candidates that actually improve the translation of the word pair under

2 http://translate.google.com/
Additional heuristics could further rule out many candidates, blacklisting those that rarely occur and that are shown to produce adverse effects in other translations. Thus, we have defined statistics that determine the minimum threshold of the allowed frequency of valid translation rules. This threshold is calculated as the mean frequency minus its standard deviation (see Figure 1). As the frequency of each word in the whole terminology is known, the frequency of each rule can be weighted to reduce the bias of words in the training set.

\[
\bar{f} = \frac{1}{n} \sum_{i=1}^{n} f_i; \quad \sigma_f = \frac{1}{n-1} \sum_{i=1}^{n} (f_i - \bar{f})^2; \quad \text{threshold} = \bar{f} - \sigma_f
\]

**Figure 1.** Formula to obtain the minimum threshold for translation rules. \(\bar{f}\) is the arithmetic mean; \(f_i\) is the frequency of the rule ‘i’ in the training data set; and \(\sigma_f\) is the standard deviation of the frequencies.

The quality of the resulting set of rule candidates is quite low because of rule overlappings and negative impact of rules in other translations. Therefore, a testing step is needed to filter out such rule candidates, using the data set *Testing*. We calculate the number of exacted (ExTr), improved (ImTr) and deteriorated (DeTr) translations. Then, we use these statistics in our algorithm to remove the negative impact rules and ranking them by their profit, i.e. the best translation improvements. The filtering and ranking methods are described in Table 2. We also need to avoid that an entire source substring of a rule overlaps part of a source substring of another rule. This is achieved by grouping such rules and selecting for each group the combination of rules that do not overlap and produces the best translations. For example, the set of overlapped rules (“ct” \(\rightarrow\) “kt”), (“vect” \(\rightarrow\) “vekt”), (“ecto” \(\rightarrow\) “ekto”) and (“ectomy” \(\rightarrow\) “ektomie”) is, firstly, sorted by their profit according to the algorithm in Table 2. In this example the sorted list is: 1. (“ectomy” \(\rightarrow\) “ektomie”), 2. (“vect” \(\rightarrow\) “vekt”), 3. (“ct” \(\rightarrow\) “kt”) and 4. (“ecto” \(\rightarrow\) “ekto”). According to the resulting rule order, we select the rules (“ectomy” \(\rightarrow\) “ektomie”) and (“vect” \(\rightarrow\) “vekt”) because they have the highest profit and do not overlap. Finally, we rank all the selected rules of each group in a list by the decreasing rate of exact translation occurrences (ExTr) with the testing data set.

**Table 2.** Pseudo code of the algorithm for filtering and ranking translation rule candidates. ExTr is the number of occurrences of exact translations; ImTr is the number of improved translations (reduced distance); and DeTr is the number of deteriorated translations (increased distance). The rules are removed when the number of deteriorations is higher or equal than the number of improvements. The rules are grouped by their overlapping source substring and sorted by lower DeTr and higher ImTr. Finally, it selects non overlapping rules from all sorted groups is sorted by their ExTr value.

1. For each translation rule:
2. If DeTr >= ImTr then continue
3. Group rules by overlapping source substring
4. For each rule group:
5. Sort the words by first profit rate: the lowest DeTr and then by the highest ImTr.
6. Select the first sorted rules in the group that do not overlaps.
7. Sort all selected rules from each group by the highest ExTr.
8. Return the sorted list of rules.

The evaluation done uses the resulting list of the acquired and filtered translation rules with manually curated translation pairs form *Evaluation*. The rules were applied using the rule order provided in the resulting list. The translated words were compared with the manual translations to evaluate their impact on the translation.
2. Results

The system is published as a Google code project “translation-rule-extraction”\(^3\), where the latest version of the source code and the collected and resulting files are hosted. The file with SNOMED CT translations is divided into the following columns: (1) the original (English) word from SNOMED CT; (2) its Google translation (mostly single word); (3) its manually curated translation; and (4) the frequency of the original word in all English SNOMED CT terms. As a result, we obtained a list of 268 translation rules (cf. Table 3). Each rule consists of a source string and a target string that represent the substring replacement in a word translation.

<table>
<thead>
<tr>
<th></th>
<th>Source String</th>
<th>Target String</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>“ine_” → “in_”</td>
<td>(“Adenine” → “Adenin”)</td>
</tr>
<tr>
<td>2</td>
<td>“ate_” → “at_”</td>
<td>(“Fibrate” → “Fibrat”)</td>
</tr>
<tr>
<td>3</td>
<td>“ia_” → “ie_”</td>
<td>(“Anemia” → “Anämie”)</td>
</tr>
<tr>
<td>4</td>
<td>“ide_” → “id_”</td>
<td>(“Choride” → “Chlorid”)</td>
</tr>
<tr>
<td>5</td>
<td>“uis_” → “se_”</td>
<td>(“Analysis” → “Analyse”)</td>
</tr>
<tr>
<td>6</td>
<td>“one_” → “on_”</td>
<td>(“Deoxycortone” → “Deoxycorton”)</td>
</tr>
<tr>
<td>7</td>
<td>“sm_” → “smus_”</td>
<td>(“Albinism” → “Albinismus”)</td>
</tr>
<tr>
<td>8</td>
<td>“ole_” → “ol_”</td>
<td>(“Phenole” → “Phenol”)</td>
</tr>
<tr>
<td>9</td>
<td>“hy_” → “hie_”</td>
<td>(“Hypertrophy” → “Hypertrophie”)</td>
</tr>
<tr>
<td>10</td>
<td>“my_” → “mie_”</td>
<td>(“Gastrostomy” → “Gastrostomie”)</td>
</tr>
</tbody>
</table>

The evaluation of the resulting list of translation rules shows a performance of 60% of improved word translations out of a data set that contains 29,790 manually curated translation pairs. When we only take into account the translations that produce an exact translation the percentage is 55%. Examples of exact translations are (“thrombocythaemia” → “thrombozythämie”) or (“macroscopecally” → “makroskopisch”). However, sometimes a translation is not exact but improves the distance between the source and target words, i.e. the English word “hypoglycaemia” should be translated into the German word “hypoglykämie” but we obtained the word “hypoglycämie” in which the character ‘c’ should have changed into ‘k’. A subsequent evaluation step focused only on those words (n=13,656) for which Google Translate has not produced a translation and which do not occur in the SNOMED CT organism hierarchy (knowing that the Linnaean organism terms in Latin are used in both English and German without modification). Here, in 56% we got improved translations, and 27% of translations corresponded to the gold standard. We further analyzed a random sample of 100 translation pairs, where the translation had not matched the gold standard. We rated 16 of them as correct translations, as they constituted valid synonyms or variants. We then matched the remaining 84 word pairs against the entire German and English Wikipedia corpus. By requiring a two orders of magnitude\(^4\) higher corpus frequency of the source word compared to the target word, 34 translations could be identified as wrong. Surprisingly, 23 source words did not match any token in the English Wikipedia, mostly single-word compounds like "blepharorhytidectomy", "onychocutaneous", or "scaphycephaly".

\(^3\) https://code.google.com/p/translation-rule-extraction/
\(^4\) We set 0.3 for the words with no match in the target corpus
3. Discussion and Conclusion

The proposed approach benefits from machine translations available for frequent words in order to extract substring translation rules that are characteristic for medical terminology and can thus be applied to those words for which machine translation fails. We obtained 268 rules, which showed a clear focus on affixes ("_macr" → "_makr"; "cally_" → "sch_"). The latter also shows one weakness of our approach, viz. that the application of rules for adjective produced rather poor results: while English adjectives are invariant (not considering comparison), German adjectives have five different inflexion suffixes, whereas the (uninflected) lemma, as present in the gold standard, is relatively rare within medical terms. This occurred with about 15% of terms in our evaluation set. Further, the scope of the translation rules is limited to words that are similar because they derive from a common root, such as Greek and Latin. This is not the case with many other word pairs (e.g., “examination”; “Prüfung”), which do not have a common root and for which our rule-based approach is inappropriate.

The result of the evaluation of the resulting list of translation rules shows a performance of 60% of improved word translations, whereas the overall translation correctness of Google Translate is 87%. However, in the evaluation data set, 59% of all words were the same in English and German, such as “antigen”, “serum”, “escherichia”. Hence, if we ignore these cases, the translation correctness drops to 37% whereas the translation correctness with Google Translate is still 74%. If we look at the words for which the machine translation system failed, there are still 55% improvements and 27% correct translations. Wrong translations can be identified by lookup in large reference corpora. However we found out that a considerable number of source words are so rare that they could not be found in a large corpus like the English Wikipedia. It is for these cases the proposed translation approach seems promising, because low frequency words are missed by statistical machine translation systems. Main limitations of this approach are the morphological richness in the target language (as we could demonstrate with the phenomenon of adjective inflection) and the restriction to those words that use the morphological inventory of Greek and Latin, as a common but not universal phenomenon in medical terminology.

References