Acquisition of Character Translation Rules for SNOMED CT Localizations

Jose Antonio Miñarro-Giménez[a]
Johannes Hellrich[b]
Stefan Schulz[a]

[a]Institute for Medical Informatics, Statistics and Documentation, Medical University of Graz, Austria
[b]Jena University Language and Information Engineering Lab
Jena, Germany

Introduction

- Interoperability of EHR data across languages requires the translation of medical terminologies such as SNOMED CT.
- SNOMED CT is currently available fully or partially in English, Spanish, French, Danish, Dutch, Swedish.
- Mostly only preferred terms / FSN are translated: terminology mismatch with clinicians' language
- Human translation of SNOMED CT is costly
  - English SNOMED CT:
    ~300,000 concepts
    ~700,000 terms

http://ihtsdo.org/snomed-ct

Introduction

- Machine translation techniques in combination with manual curation could reduce the cost of producing term translations or localised entry terms.
- Statistical machine translation (SMT) systems are based on the existence of parallel texts to generate a translation model
- Rule-based translation systems are based on a set of translation rules.
- Medical terminologies contains many terms derived from Greek or Latin origins which are shared across languages.
  - Appendicitis → Appendizitis
  - Appedectomy → Appendektomie
  - Osteosarcoma → Osteosarkom
  - Cholesterol → Cholesterin

Character translation rules

- Appendicitis ⟶ Appendizitis
- Appedectomy ⟶ Appendektomie
- Osteosarcoma ⟶ Osteosarkom
- Cholesterol ⟶ Cholesterin

"icit" ⟶ "izit"

"ectomy<end>" ⟶ "ektomie<end>"

"sarc" ⟶ "sark"

"ol<end>" ⟶ "in<end>"

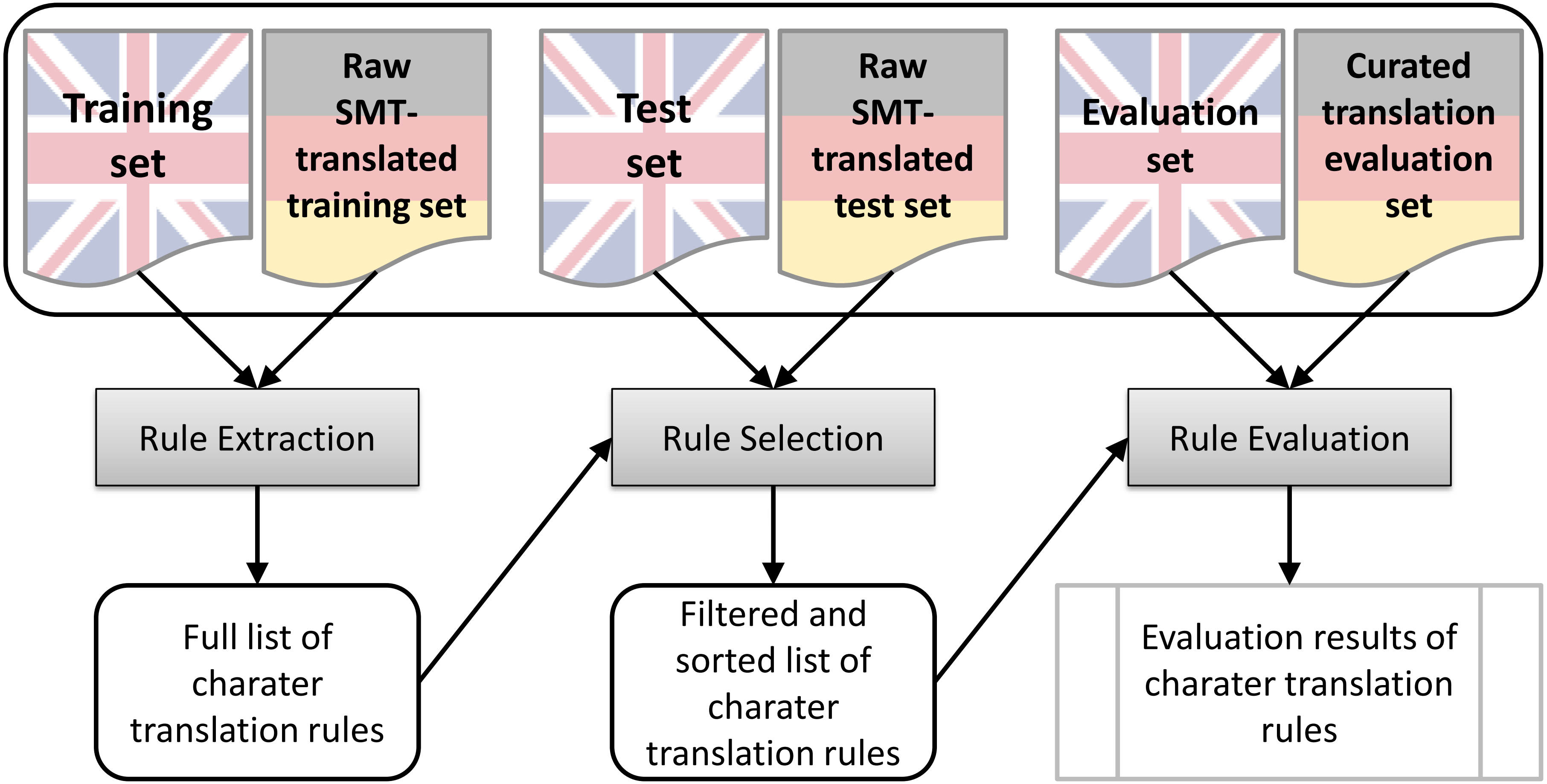
- Can such character translation ("transliteration") be learned from uncorrected machine-translated term lists?
- Can they be used for creating new localised terms in a terminology translation process?

Resources

- All single word types occurring in all descriptions of the international version of SNOMED CT (7/2013)
- Machine translated using Google Translate into German
- Reasons for non-translated words
  - Rare words (not enough training material)
  - English or Latin words in German source
  - Co-incidence of German and English words
- Gold standard: manually curated list of 29,790 translated Google translated words. 1:1 translation (most likely target word and inflectional form)

Methods

List of word translation pairs



Training set

Raw SMT-translated training set

Test set

Raw SMT-translated test set

Evaluation set

Curated translation evaluation set

Rule Extraction

Rule Selection

Rule Evaluation

Full list of charater translation rules

Filtered and sorted list of charater translation rules

Evaluation results of charater translation rules

Rule extraction

- Test all combinations of characters substitution between source and target word.
- Limit the total number of combinations by defining:
    - max and min allowed length of the substitution strings in the source and target word.
    - max length difference between source and target substitution strings.
- A rule is extracted when its improves the translation
- A rule improves a translation if Levenstheins’ distance between the rule-translated word and the SMT translated word is lower than between the source word and the SMT translated word.

Rule Selection

- The set of extracted rules is tested to obtain the best rule list with highest improvement in the test dataset.
- The rules are grouped by the overlapping source substitution strings. For each group the one is selected that produces the highest number of translation improvements.
- The rank of selected rules is computed on:
  1. Highest number of improved translations
  2. Lowest number of deteriorated translations
  3. Highest number of words translated correctly

Overlapping group

- “ct” → “kt”
- “vect” → “vekt”
- “ecto” → “ekto”
- “ectomy” → “ektomie”

Examples of charater translation rules EN→DE

Rank	Rule	Example
1	“ine_” → “in_”	“_Adenine_” → “_Adenin_”
2	“ate_” → “at_”	“_Fibrate_” → “_Fibrat_”
3	“ia_” → “ie_”	“_Anemia_” → “_Anemie_”
4	“ide_” → “id_”	“_Choride_” → “_Chlorid_”
5	“sis_” → “se_”	“_Analysis_” → “_Analyse_”
6	“one_” → “on_”	“_Deoxycortone_” → “_Deoxycorton_”
7	“sm_” → “smus_”	“_Albinism_” → “_Albinismus_”
8	“ole_” → “ol_”	“_Phenole_” → “_Phenol_”
9	“hy_” → “hie_”	“_Hypertrophy_” → “_Hypertrophie_”
10	“my_” → “mie_”	"_Gastrotomy_” → “_Gastrotomie_”

Rule Evaluation

- Gold standard contains manually curated list of 29,790 translated single words.
- The selected and sorted list of rules is evaluated using the gold standard.
  1. Rule-translated words are obtained.
  2. Statistical machine translated (SMT) words are obtained using Google Translate.
  3. Levenstheins' edit distance is calculated between the rule-translated words and the gold standard and also between the SMT words and the gold stardard.
  4. The calculated edit distances are compared.

Rule Evaluation

- A list of 286 rules was created
- Google translate produced 87% of correct translations
- Rule translations obtained 60% of correct translations
- The rule approach improved 55% of **not** correctly translated words by Google Translate
- The rule approach correctly translated 27% of **not** correctly translated words by Google translate.

Conclusions

- Character translation rules can be automatically obtained from SMT translated word lists
- Inflection and variability of words in target language (German) complicates the exact translation based on rules.
- The rule based approach cannot process words that do not share common roots.
- SMT produces better results than rule-based translations. However, character translation rules are promising add-ons to translate words that are not translated by SMT
  - "Long tail" of low frequency medical terms
- Interesting approach for closely related languages?