Acquisition of Character Translation Rules for Supporting SNOMED CT Localizations

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

[a]Institute for Medical Informatics, Statistics and Documentatoin, Medical University of Graz, Austria

[b]Jena University Language and Information Engineering Lab Jena, Germany

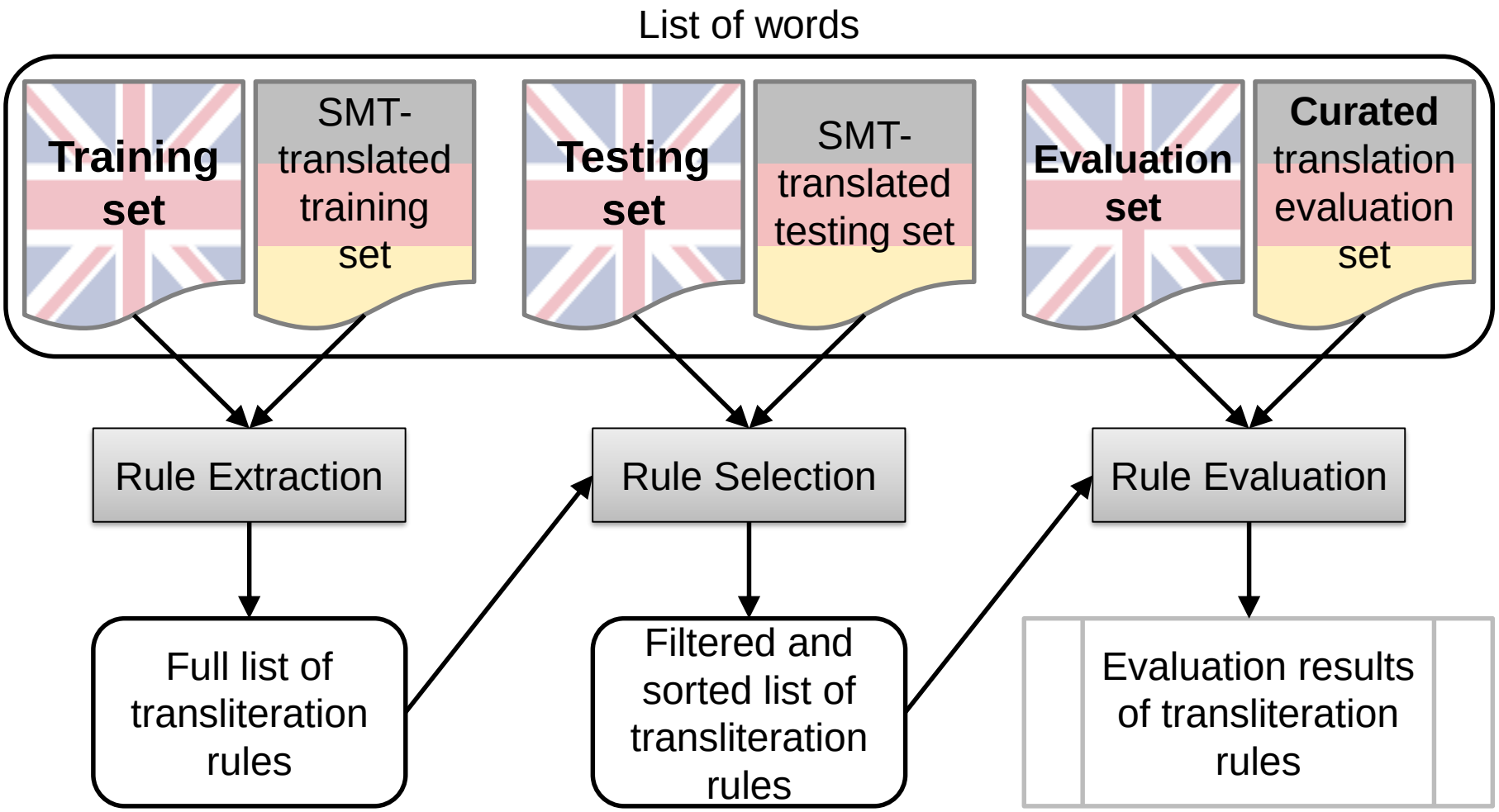
Introduction

- SNOMED CT supports the development of comprehensive high-quality clinical content in health records.
- Interoperability of EHR data across languages requires the translation of medical terminologies.
- SNOMED CT is currently available fully or partially in English, Spanish, French, Danish, Dutch, Swedish.
- Other costs related to the adoption SNOMED CT
  - Terminology license and participation in IHSTDO SIG
  - Terminology managment system and infraestructure
  - Human resources: coordination, terminologists.
  - Mapping with legacy systems.
  - ...

Introduction

- Machine translation techniques in combination with manual curation could reduce the cost of producing term translations.
- Statistical machine translation (SMT) systems are based on the existence of parallel text to generate the translation model.
- Rule-based translation systems are based on the definition of translation rules.
- Medical terminologies contains many terms derivated from Greek or Latin origins which are shared across languages.
  - Appendicitis → Apendizitis

Methods


List of words
Training set
SMT-translated training set
Testing set
SMT-translated testing set
Evaluation set
Curated translation evaluation set
Rule Extraction
Rule Selection
Rule Evaluation
Full list of transliteration rules
Filtered and sorted list of transliteration rules
Evaluation results of transliteration rules

Rule Extraction

- Testing all combinations of characters substitution between source and target word.
- Limit the total number of combinations by defining:
  - The max and min allowed length of the substitution strings in the source and target word.
  - The max number of characters between source and target substitution strings.
- A rule is extracted when the source and target substitution strings improved the translation.
- A rule Improves a translation when the Levenstheins' distance between the rule translated word and the SMT translated word is lower than the distance between the source word and the SMT translated word.

Rule Selection

- The set of extracted rules is tested to obtain the best list of rule with highest improvement using the testing dataset.
- The set of extracted rules are grouped by the overlapping source substitution strings and we select for each group the one which better translatate the testing dataset more times.
- The selected rules from each group is sorted based on the overall improvement achieved with the testing dataset.
- The rank of selected rules depend on:
  1. Highest number of improved translations.
  2. Lowest number of deteriorated translations.
  3. Highest number of words correctly translated.

Overlapping group

- "ct" → "kt"
- "vect" → "vekt"
- "ecto" → "ekto"
- "ectomy" → "ektomie"

List of Transliteration 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_”	Gastronomy” → “Gastronomie”

Rule Evaluation

- Gold standard contains 29,790 manually curated list of translated 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.
  3. The Levenstheins' 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 distances are compared.

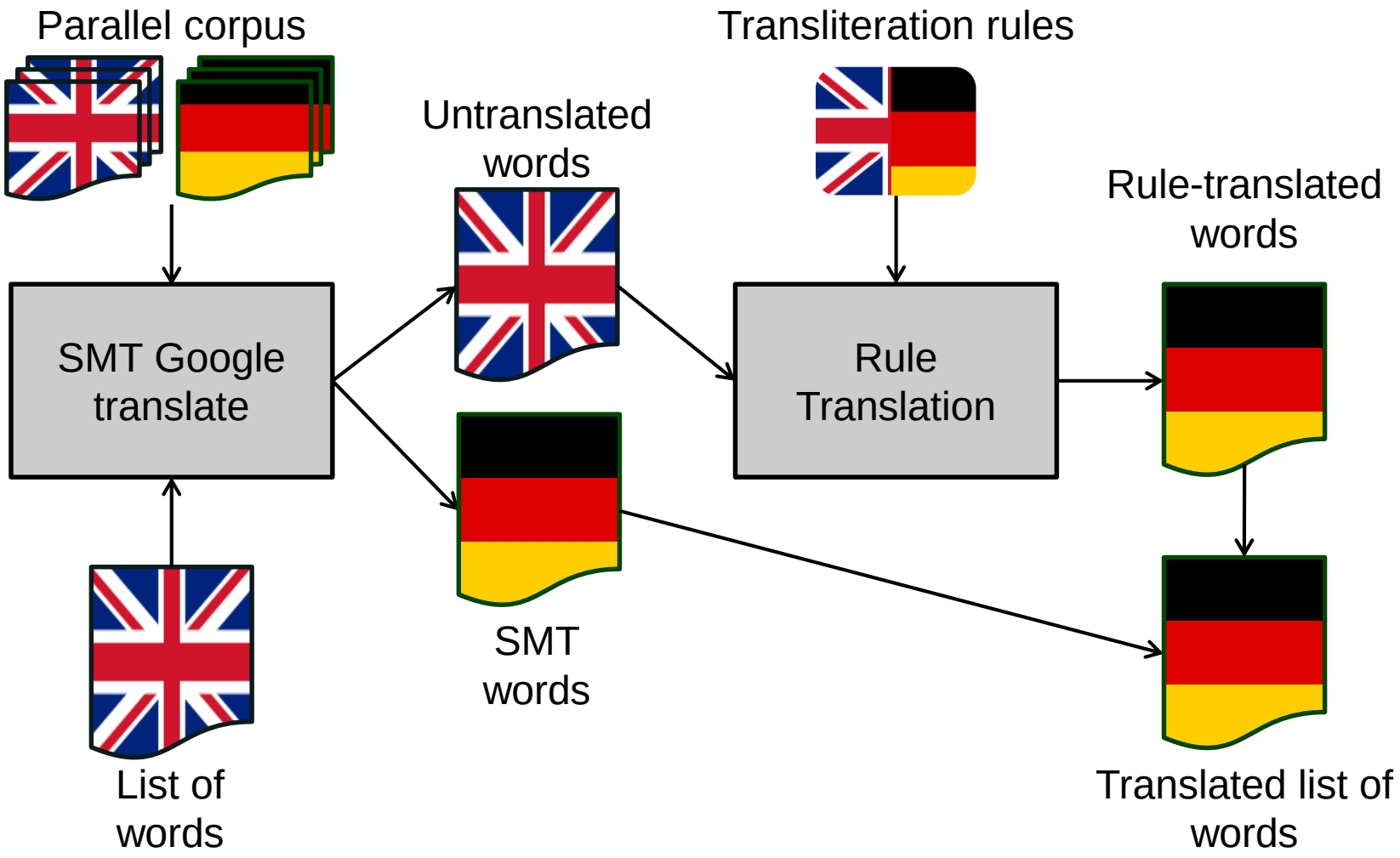
Results

- A list of 286 rules was created.
- Google translate produced 87% of correct translations
- Rule translations obtained 60% of correct translations.
- Rule approach improved 55% of **not** correctly translated words by Google translate.
- Rule approach correctly translated 27% of **not** correctly translated words by Google translate.
- The 59% of all words in the evaluation dataset have the same in English and German, e.g. “serum”, “escherichia”

Conclusions

- Translation rules can be automatically obtained from parallel corpus produced by a statistical machine translation.
- Inflection and variability of words in target language (German) complicates the exact translation based on rules.
- Rule based approach cannot deal with words that do not share common root.
- Statistical machine translation produces better results than rule-based translations. However, Rule approach could improve the translation of words that are not translated by the statistical machine translation.
  - Low frequency terms in specific domains, such as medicine.

Combined translation approach



Questions