Semiautomated acquisition of localised SNOMED CT content

Stefan Schulz

Rationale

- Lack of availability of SNOMED CT translations for many languages
- Translation costs and time
- Many repetitive passages in SNOMED CT
  114 concepts with "off-road motor vehicle, except snow vehicle"
- Many use cases such as NLP do not require "canonical" fully specified names but close to user terms as used in clinical documents
- "Imperfect", scalable solutions for terminology localisation, combining human with machine translation

Approach (1)

- Material: SNOMED CT "descriptions" in English version (approx. 780,000)
- Preprocessing: substitution of digits

  "urethral catheter with 900mL drainage bag" →

  "urethral catheter with °°°mL drainage bag"
- Raw machine translation of word n-grams (n = 1 – 6) to German using Google translate
- Ranking of word n-grams by frequency
- Manual curation of most frequent n-gram translations

Approach (2)

4192	On examination -	Bei der Untersuchung -	Bei der Untersuchung -
3301	malignant neoplasm of	bösartige Neubildung der	bösartige Neubildung des
2283	of lesion of	der Läsion	einer Läsion des
2051	of undetermined intent	von unbestimmt	unbestimmter Absicht
1838	Blood group antigen	Blutgruppen -Antigen	Blutgruppenantigen
1758	Blood group antibody	Blutgruppenantikörper	Blutgruppenantikörper
1750	specific IgE antibody	spezifische IgE- Antikörper	spezifische IgE-Antikörper
1739	IgE antibody measurement	IgE-Antikörper- Messung	
1660	structure (body structure)	Struktur (Anatomische Struktur),	Struktur (Anatomische Struktur),
1567	Neoplasm of uncertain	Neubildung unsicheren	Neubildung unsicheren
1532	Finding related to	Zugehörige zu	Befund zugehörig zu
1482	Adverse reaction (disorder)	Nebenwirkung (Störung)	Nebenwirkung (Erkrankung)
1411	Salmonella enterica subsp.	Salmonella enterica subsp.	Salmonella enterica subsp.
1394	enterica subsp. enterica	enterica subsp. ente	
1392	subsp. enterica ser.	subsp. ente ser.	
1390	Benign neoplasm of	Gutartige Neubildung	Gutartige Neubildung des
1346	function (observable entity)	-Funktion (beobachtbaren Person)	Funktion (beobachtbaren Einheit)
1341	related to ability	bezogen auf die Fähigkeit	
1339	to ability to	die Fähigkeit zu	
1290	using fluoroscopic guidance	mit Durchleuchtungskontrolle	mit Durchleuchtungskontrolle
1278	specific immunoglobulin E	spezifische Immunglobulin E	spezifisches Immunglobulin E

Manual curation (1)

- Single tokens:
  - Identification of components of Linnean names (Latin endings, not to be translated)
  - Learning of character translation rules from raw translations. Application of these rules to non-translated words

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

"icit" → "izit"

"ectomy<end>" → "ektomie<end>"

"sarc" → "sark"

"ol<end>" → "in<end>"

Jose Antonio Miñarro-Giménez, Johannes Hellrich, Stefan Schulz. Acquisition of Character Translation Rules for SNOMED CT Localizations. Medical Informatics Europe (MIE) 2015

Manual curation (2)

- Token translation issues:
  - preference of adjective inflection (German: 6 forms): neuter nominative singular
  - noun inflections: nominative singular
  - ambiguous words: opt for most plausible translations
  - so far: no 1:n translations
- Ngram (n > 1) translation issues
  - "of" → "des" (definite article, genitive, male, sing.)

- Additional resource generated out of source: "Most frequent patterns", generated out of (nearly) all permutations with wildcards
- Chunks translated by existing ngrams translations and manually revised

66772212868	165	311	* * of * bone	* * von * Knochen	* * des * _Knochens
38797258228	166	311	* - * * antigen *	* - * * Antigen *	* - * * Antigen *
34940214572	167	311	Poisoning by *	Vergiftung durch *	Vergiftung durch *
46226296366	168	309	Salmonella IIIb *	Salmonella IIIb *	Salmonella IIIb *
45200244212	169	308	* - Human * antigen *	* - Menschliche * Antigen *	* - Menschliches * Antigen *
239170962	170	308	HLA * * * antigen *	HLA * * * Antigen *	HLA * * * Antigen *
8706272368	171	308	O/E - * * * *	bei Untersuchung - * * * *	bei Untersuchung - * * * *
16539307143	172	306	History of * * *	Vorgeschichte eines * * *	Vorgeschichte eines * * *
17596206641	173	306	HLA - Human * antigen *	HLA - HUMAN * Antigen *	HLA - Humanes * Antigen *
50352116198	174	305	* * of * * vertebra	* * von * * Wirbel	* * des * * _Wirbels
95074268029	175	304	* * neoplasm of *	* * Neubildung der *	* * Neubildung des *
72842155237	176	304	* * * of * * of * *	* * * von * * von * *	* * * des * * des * *
16495242534	177	302	* of * gland	* von * Drüse	* der *drüse
34462108808	178	301	* * of skin	* * der Haut	* * der Haut

- Automated term translation
  - identification of "best fitting pattern"

    "malignant neoplasm of **the head** of **the pancreas**" →

    "malignant neoplasm of * * of * * "
  - translation of "fillers"
    - by lookup in curated n-gram translation lists
    - longest match from left to right
    - if no translation found use "n-1"-gram
    - token translation: all single tokens had been machine-translated and at least partly reviewed
    - no token translation: use English word.

- Translation available at SEMCARE dropbox
- Detailed error analysis to be done
- Know to do items:
  - Chunking prior to n-gram generation to avoid arbitrarily delineated term fragments
  - post processing to correct number/gender/case inflection errors
  - requires identification of noun gender and rules for noun inflection endings
- Modification of approach to generate more synonyms

Discussion

- Method still to be evaluated
- Has consumed so far approx. 3 PMs
- Is it worth while to invest more in this method?
  - for SEMCARE
  - for other use cases
  - for other language pairs