Department of Computer and Systems Sciences

Stockholm University

Adapting a parser to clinical text by simple pre-processing rules

1) Identify sentence types typical to the clinical domain.

Clinical text (Assessment fields)

Medical text (Läkartidningen)

Extract sentence level POS-tag sequences

POS-tagger (Granska)

Clinical text
dt nn vb nn mad x%
...

Medical text
dt nn vb nn mad y%
...

Calculate POS-tag sequence proportions

Compare proportions using Wilson score interval (95% CI)

1 736 sentence types significantly more frequent in clinical text

jj nn mad	1.27%
nn pp nn mad	0.95%
vb nn mad	0.59%
...	

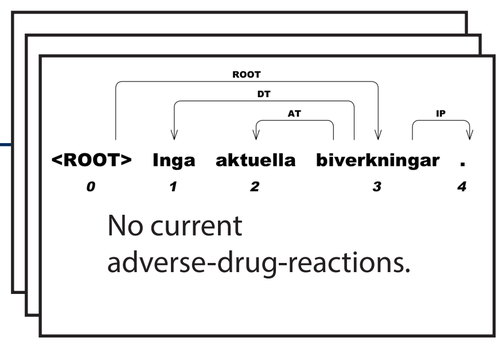
2) Identify errors made by the parser on sentence types typical to the clinical domain.

Extract sentence types for which:
min proportion in clinical text
-
max proportion in medical text

>=0.1 pp

33 sentence types (10% of the clinical sentences are of these types.)

Run Malt-parser, trained on standard Swedish, on sentences of these types.



Manually identify error types in 30x33 of the parsed sentences.

1) Abbreviated words ending with a full stop, resulting in incorrect sentence splittings.
2) Abbreviations incorrectly tagged as nouns.
3) Adjectives (often abbreviated) not recognised as such, resulting in AT relations being labelled as DT relations.
4) A general adverbial relation incorrectly assigned an adverb of place or time relation or vice versa.
5) The first word in compound expressions parsed as a determiner to the second.
6) 'nn pp nn pp nn mad' sentences with an incorrectly attributed preposition.
7) The sentence type 'nn jj' (noun adjective), for which most evaluated sentences were incorrectly parsed.
8) An omitted initial subject, resulting in the object incorrectly being parsed as the subject of the sentence.

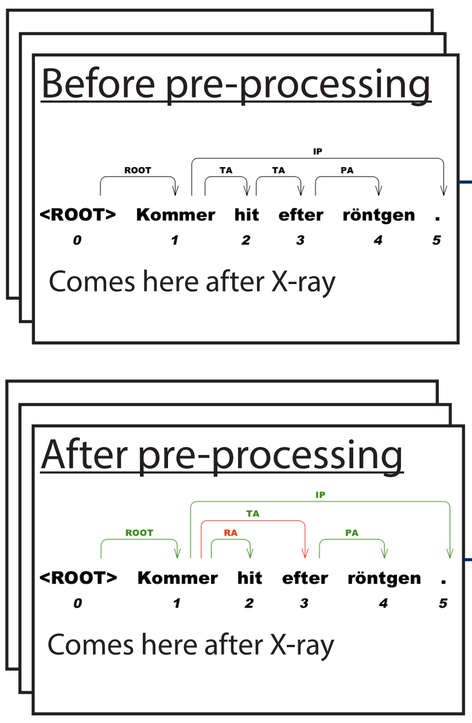
3) Exemplify how this knowledge can be used, by pre-processing rules

Apply pre-processing rules for error types 7) and 8):

* Add copula in the middle of sentences of the type 'nn jj'.

* Add the pronoun 'I' as the first word in sentences starting with a verb.

Run Maltparser on the pre-processed sentences.



Compare parsings, with and without pre-processing

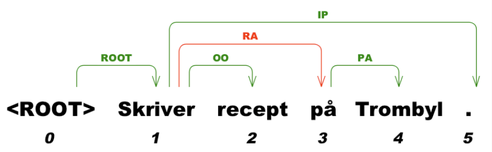
Sentences were either:

a) Given a completely correct parsing and labelling in 64% - 100% of the cases after pre-processing.

b) Slightly improved by pre-processing.

'nn jj'-sentences were correctly parsed after pre-processing. Many incorrectly labelled subject relations were changed to an object relation and some adverbs of place and time received a correct labelling.

However, sentences of the type 'vb nn pp nn mad' received an incorrect parsing after pre-processing.



4) Future work

- Apply abbreviation expansion, as error types 1), 2) and partly 3) were caused by abbreviations.
- Use medical vocabularies for identifying compound expressions (error type 5).
- Analyse shorter sequences than sentences.
- Annotate a small treebank of clinical text (for instance with the help of pre-annotation by a standard parser, improved by the methods in this study).

Maria Skeppstedt, mariask@dsv.su.se