Department of Computer and Systems Sciences (DSV)

Stockholm University

Entity Recognition of Pharmaceutical Drugs in Swedish Clinical Text

Sidrat ul Muntaha, Maria Skeppstedt, Maria Kvist and Hercules Dalianis

An entity recognition system for expressions of pharmaceutical drugs, based on vocabulary lists from three terminologies was constructed. The system achieved a precision of 94% and a recall of 74% when evaluated on assessment texts from Swedish emergency unit health records that had been annotated for expressions denoting pharmaceutical drugs.

76-årig kvinna som under längre tid har ätit
smärtstillande, bland annat paracetamol och Voltaren.
(drug) (drug) (drug)

76 year old woman with prolonged consumption of
pain killers, such as paracetamol och Voltaren.
(drug) (drug) (drug)

Background

Data documented in narrative form in health records is difficult to use for e.g. structured summarization, advanced search, statistical analysis and data mining.

Reasoning about a patient's medication is one example of valuable information, which can be used for mining for new knowledge on pharmaceutical drugs used in health care, e.g. adverse drug reactions. It can also be useful for presenting a summarization of reasoning about a patient's medication.

Aim

Automatic recognition of pharmaceutical drug expressions in Swedish clinical text, as a first step for extracting information on medication, both for the purpose of summarization and for text mining.

Method

Vocabulary matching methods evaluated:

1. Direct match to vocabulary lists.
2. Removing Parole words from one of the vocabulary lists.
3. Also matching to terms with a Levenshtein distance of one.
4. Also matching to terms with a Levenshtein distance of two.

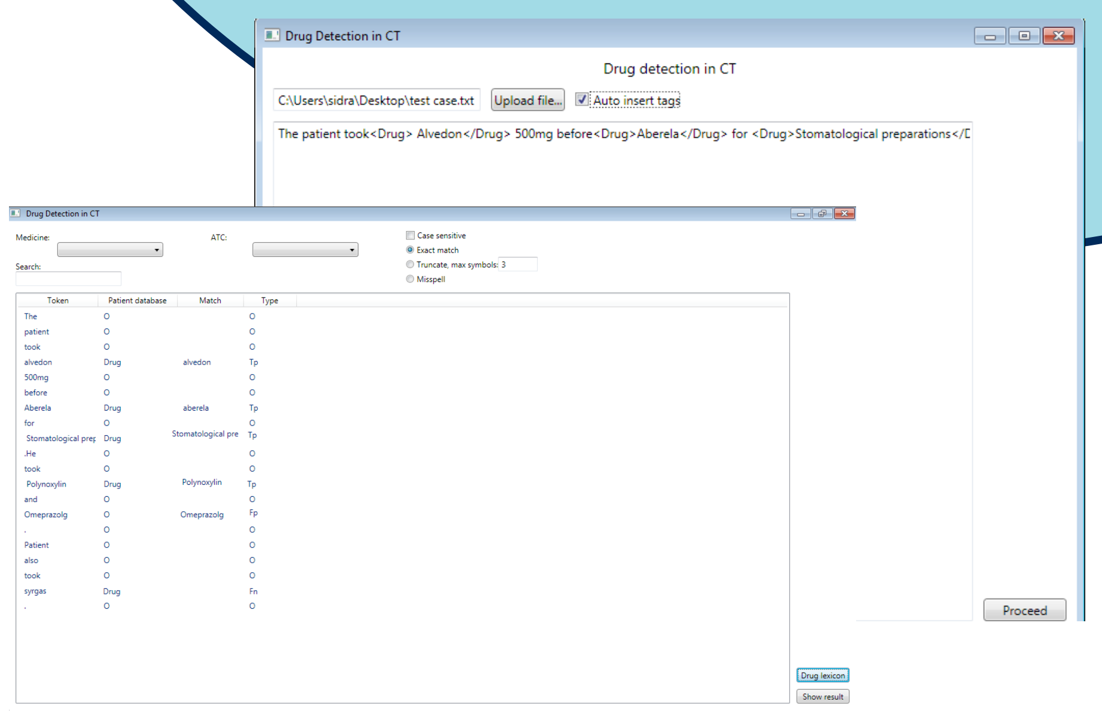
Terminologies

Total vocabulary size:
25,161 unique expressions

- FASS (Farmaceutiska Specialiteter i Sverige): Swedish product names for drugs (7,056 terms) and a list of classifications (5,062 terms)
- SNOMED CT (SNOMED Clinical Terms): terms under the main category node pharmaceutical (16,977 terms)
- MeSH (the Medical Subject Headings): Terms from the categories pharmacologic-substance (2,554 terms) and antibiotic (239 terms)

Results for expressions of pharmaceutical drugs (580 instances)

Method	Precision (CI)	Recall (CI)	F-score
Exact match	0.51 (± 0.03)	0.72 (± 0.04)	0.60
Excl. Parole	0.94 (± 0.02)	0.74 (± 0.04)	0.83
Lev dist. 1	0.91 (± 0.03)	0.74 (± 0.04)	0.82
Lev dist. 2	0.89 (± 0.03)	0.75 (± 0.04)	0.81



Drug Detection in CT

Drug detection in CT

C:\Users\sidra\Desktop\test case.txt

☑ Auto insert tags

The patient took<Drug> Alvedon</Drug> 500mg before<Drug>Aberela</Drug> for <Drug>Stomatological preparations</D

Drug Detection in CT

Medicine:

ATC:

Search:

□ Case sensitive
● Exact match
○ Truncate, max symbols: 3
○ Misspell

Token	Patient database	Match	Type
The	O		O
patient	O		O
took	O		O
alvedon	Drug	alvedon	Tp
500mg	O		O
before	O		O
Aberela	Drug	aberela	Tp
for	O		O
Stomatological prep	Drug	Stomatological pre	Tp
.He	O		O
took	O		O
Polynoxylin	Drug	Polynoxylin	Tp
and	O		O
Omeprazolg	O	Omeprazolg	Fp
.	O		O
Patient	O		O
also	O		O
took	O		O
syrgas	Drug		Fn
.	O		O

Discussion

Levenshtein distance matching method did not result in an improvement of recall, which indicates that spelling errors were not common.

Sources of error:

- Compound words.
- Expressions denoting drugs that were expressed with the effect of the drug, the disease for which it is given or the content of the medication, ('pain killer', 'heart failure medication', 'vitamin pills').
- False positives were the term 'läkemedel' ('pharmaceutical') and expressions denoting narcotics.

Future work

1. Compound splitting may be applied to improve recall.
2. Apply machine learning methods for recognizing pharmaceutical drugs. The method evaluated for this study can serve as a baseline method, and more importantly can also serve as one of the key features for such a machine learning system.

Conclusion

Expressions for pharmaceutical drugs can be automatically recognized with a precision of 94% and a recall of 74%.

Corresponding authors:
simu9349@dsv.su.se and mariask@dsv.su.se