Department of Computer and Systems Sciences

Stockholm University

Rule-based Entity Recognition and Coverage of SNOMED CT in Swedish Clinical Text

Maria Skeppstedt, Maria Kvist and Hercules Dalianis

mariask@dsv.su.se, maria.kvist@karolinska.se and hercules@dsv.su.se

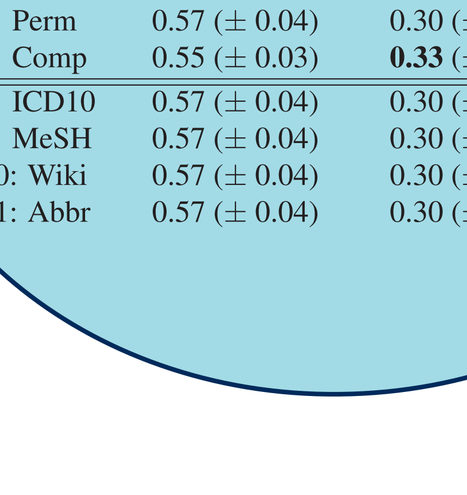
76-årig kvinna med hypertoni (disorder) och angina pectoris (disorder). Inkommer med centrala bröstsmärtor (finding).

76-year old woman with hypertension (disorder) and angina pectoris (disorder). Admitted to hospital with severe chestpain (finding).

Named entity recognition of the clinical entities *disorders*, *findings* and *body structures* is needed for information extraction from unstructured text in health records. Swedish clinical notes from an emergency unit were annotated and used for evaluating a rule- and terminology-based entity recognition system. Several preprocessing techniques were used for matching the text to terms in the medical terminology SNOMED CT belonging to the categories disorder, finding or body structure. Thereafter, four additional terminologies were added.

Results

- A large proportion of disorders (14%) and findings (12%) were written as abbreviations, which were not regognised by the system.
- Entities containing more than two tokens were not recognised by the system.
- The best results were achieved when all terminologies were used:
  Body structure: Precision: 0.74, Recall: 0.80
  Disorder: Precision: 0.75, Recall: 0.55
  Finding: Precision: 0.57, Recall: 0.30



Conclusions

Low recall for disorders and findings shows that:

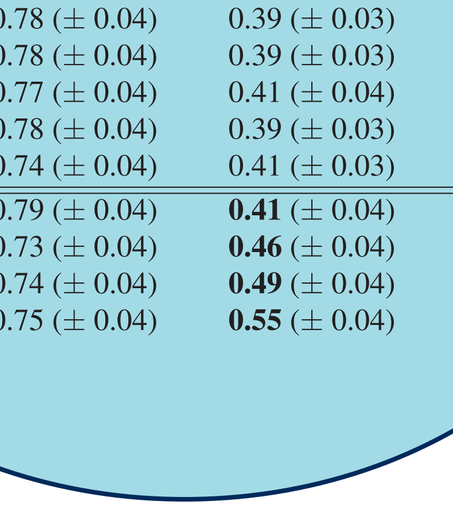
- Additional methods are needed for entity recognition.
- There are many expressions in clinical text that are not included in SNOMED CT.

Future work

- Expanding the corpus and measure inter-annotator agreement between several annotators.
- Evaluating the constructed rule-based system on clinical text from another domain.
- Methods for expansion of abbreviations.
- Applying this entity recognition system to a larger clinical corpus in order to study the prevalence of different clinical findings as well as connections between them.
- Applying machine learning methods. The rule-based entity recognition system developed for this study will be used both as a baseline and for generating features for the machine learning system.

Body structure (264 instances)

Nr.	Prec. (95% CI)	Recall (95% CI)	F-Score
1: Base	0.11 (± 0.14)	0.01 (± 0.01)	0.01
2: Lemm	0.09 (± 0.12)	0.01 (± 0.01)	0.01
3: Stop	0.41 (± 0.04)	**0.79** (± 0.05)	0.54
4: Qual	**0.73** (± 0.05)	0.77 (± 0.05)	0.75
5: Leve	0.72 (± 0.05)	0.78 (± 0.05)	0.75
6: Perm	0.73 (± 0.05)	0.77 (± 0.05)	0.75
7: Comp	0.6 (± 0.05)	0.78 (± 0.05)	0.68
9: MeSH	0.74 (± 0.05)	0.80 (± 0.05)	0.76
11: Abbr	0.74 (± 0.05)	0.80 (± 0.05)	**0.77**



Finding (1,319 instances)

Nr.	Prec. (95% CI)	Recall (95% CI)	F-Score
1: Base	0.51 (± 0.04)	0.23 (± 0.02)	0.31
2: Lemm	0.52 (± 0.04)	**0.29** (± 0.02)	0.37
3: Stop	0.53 (± 0.04)	0.29 (± 0.02)	0.37
4: Qual	**0.57** (± 0.04)	0.30 (± 0.02)	0.39
5: Leve	0.57 (± 0.04)	0.30 (± 0.02)	0.39
6: Perm	0.57 (± 0.04)	0.30 (± 0.02)	0.39
7: Comp	0.55 (± 0.03)	**0.33** (± 0.03)	**0.41**
8: ICD10	0.57 (± 0.04)	0.30 (± 0.02)	0.39
9: MeSH	0.57 (± 0.04)	0.30 (± 0.02)	0.39
10: Wiki	0.57 (± 0.04)	0.30 (± 0.02)	0.39
11: Abbr	0.57 (± 0.04)	0.30 (± 0.02)	0.39

Preprocessing/Terminologies

1. Direct match to SNOMED CT disorders, findings and body structures
2. Lemmatisation
3. Stop word filtering
4. Additional SNOMED CT categories
5. Levenshtein distance of one
6. Permutations of tokens in text
7. Compound word splitting
8. ICD-10, International classification of diseases
9. MeSH, Controlled vocabulary of Medical Subject Headings
10. Wikipedia list of diseases
11. List of medical abbreviations and acronyms

Disorder (759 instances)

Nr.	Prec. (95% CI)	Recall (95% CI)	F-Score
1: Base	0.78 (± 0.04)	0.38 (± 0.03)	0.51
2: Lemm	0.78 (± 0.04)	0.39 (± 0.03)	0.52
3: Stop	0.78 (± 0.04)	0.39 (± 0.03)	0.52
4: Qual	0.78 (± 0.04)	0.39 (± 0.03)	0.52
5: Leve	0.77 (± 0.04)	0.41 (± 0.04)	0.54
6: Perm	0.78 (± 0.04)	0.39 (± 0.03)	0.52
7: Comp	0.74 (± 0.04)	0.41 (± 0.03)	0.52
8: ICD10	0.79 (± 0.04)	**0.41** (± 0.04)	0.54
9: MeSH	0.73 (± 0.04)	**0.46** (± 0.04)	0.56
10: Wiki	0.74 (± 0.04)	**0.49** (± 0.04)	0.59
11: Abbr	0.75 (± 0.04)	**0.55** (± 0.04)	**0.63**