23rd International Conference of the European Federation for Medical Informatics
User Centred Networked Health Care - A. Moen et al. (Eds.)
MIE 2011 / CD / Posters

Using SNOMED CT for High Precision Entity Recognition in Swedish Clinical text

Maria SKEPPELSTEDT[a,1], Hercules DALIANIS[a]
[a] *Department of Computer and Systems Sciences (DSV)*
Stockholm University, Forum 100, 164 40 Kista, Sweden

Abstract An evaluation was performed of retrieval of findings in Swedish clinical text through exact string matching against SNOMED CT. The aim was to create a system for retrieving clinical findings with high precision that for example can be used as training data for machine learning. The evaluation was performed on previously manually annotated findings, and the best approach showed a precision of 93 percent.

Keywords. Clinical text, Swedish, Findings, SNOMED CT

1. Introduction

A lot of valuable unstructured information in form of free text is today entered into electronic patient records systems. An example of this type of unstructured information is clinical findings. One way to extract this information is to manually annotate a limited set and then use this as training data for a machine learning system that detects the findings in free text. However, to manually annotate a large corpus is expensive in terms of time and money, and therefore there is a need for additional data. We propose a two-step approach, in which findings are first retrieved with high precision and low recall through exact string matching against SNOMED terms. The retrieved findings can then be used as a part of the training data for a machine learning system. A similar approach is for example described by Niu et al [1]. We here evaluate a system that aims at retrieving findings with high precision, and that could be used in the first phase of this two-step approach.

2. Method

As a gold standard for our experiment we have used a subset from the Stockholm EPR Corpus [2], containing one Swedish ICU clinic[2], that was annotated for *diagnosis*, *symptoms* and *findings* [3]. No attempt was made to try to distinguish between these three classes, and they were therefore merged into one class, *clinical finding*.

Five different methods for detecting these clinical finding were used. The first method performed an exact string matching against all Swedish SNOMED terms [4]

[1] Corresponding Author.

[2] The research was carried out after approval from the Regional Ethical Review Board in Stockholm, permission number 2009/1742-31/5.

with the semantic category *disorder*. The second method performed an exact string matching against the same terms, except that all unigrams and bigrams that occurred more than five times in a 600,000 token non-clinical Swedish corpus [5] were removed, thereby possibly removing terms that also have another non-clinical meaning. Thereafter, the exact string matching was carried out on SNOMED terms that belonged to either the category *disorder* or the category *finding*. A match was performed both with the complete list of terms and with a list in which the terms that occurred more than five times in the non-clinical corpus were removed. For the fifth method, SNOMED terms longer than two words were removed.

Partial recall was compared to an exact recall, both obtained through evaluating against the gold standard on a token level, but where exact recall required that all tokens in the clinical term were retrieved. The difference thus indicates instances where only a part of a clinical term is retrieved, e.g. *diabetes* instead of *diabetes mellitus*.

3. Results and conclusions

As can be seen in Table 1, the only methods that gave a high precision were a matching against the SNOMED disorders. Also, the false positives for the sting matching against disorders could all be defined as clinical findings or modifiers to clinical findings.

Table 1. Result for the five different methods for retrieving clinical entities through exact string matching.

SNOMED terms	Exact precision	Exact recall	Partial recall	Correct retrieved terms (unique)
All disorders	**0.928** (±0.031)	0.127	0.151	227 (114)
Disorders without common words	**0.929** (±0.033)	0.109	0.127	192 (97)
All disorders and findings	0.582 (±0.035)	0.231	0.293	405 (185)
Disorders and findings without common words	0.798 (±0.044)	0.133	0.161	216 (129)
Disorders and findings without common words and a maximum of two tokens	0.859 (±0.041)	0.123	0.151	205 (120)

Fewer terms were found with the second method, without an increase in precision, which makes a match against the complete list of SNOMED disorders a better choice. A disadvantage with only using the semantic type *disorder* for retrieving findings to use as training data for a machine learning system, is that it might bias the system to detect only this type and not findings in general, Therefore the two last methods might in some cases be preferable, even though they have a lower precision.

References

[1] Niu, C., Li, W., Ding, J., Srihari, R.K. *A bootstrapping approach to named entity classification using successive learners*. Proceedings of the 41st Annual Meeting of the ACL (2003), 335–342.

[2] Dalianis, H. Hassel, M. Velupillai. S. *The Stockholm EPR Corpus - Characteristics and Some Initial Findings*. In the Proceedings of ISHIMR 2009, Kalmar, Sweden, 14-16 October, (2009), 243-249

[3] Velupillai, S. Dalianis, H. Kvist, M. *Factuality Levels of Diagnoses in Swedish Clinical Text*, (2011), to be published in MIE, Int. Conf. of the European Federation for Medical Informatics Oslo, Norway.

[4] Socialstyrelsen: SNOMED CT-licens. http://www.socialstyrelsen.se/halsoinformatik/nationelltfacksprak/ snomedct-licens (Accessed 2011-01-24)

[5] M. Gellerstam Y. Cederholm, T. Rasmark *The bank of Swedish*. In: Proceedings of LREC 2000 - The 2nd International Conference on Language Resources and Evaluation, 329–333, (2000) Athens, Greece.