Negation detection in Norwegian medical text: Porting a Swedish NegEx to Norwegian Work in progress

Norwegian Centre for E-health Research

Stockholm University

Andrius Budrionis[1], Hercules Dalianis[1,2], Kassaye Yitbarek Yigzaw[1], Alexandra Makhlysheva[1], Taridzo Chomutare[1]

[1]Norwegian Centre for E-health Research, University Hospital of North Norway, Tromsø, Norway
[2]DSV/Stockholm University, Kista, Sweden

Abstract

This paper presents an initial effort in developing a negation detection algorithm for Norwegian clinical text. To support Norwegian text, an evaluated version of Swedish NegEx was extended by translating the negation triggers, adding more negation rules and using a pre-processed Norwegian ICD-10 codes list (2017). The Norwegian NegEx was tested on a corpus containing 170 scientific publications in Norwegian from a medical domain. NegEx found 70 negated findings/disorders. The result is not completely evaluated due to the lack of a gold standard. Need for further preprocessing of the Norwegian ICD-10 codes list for matching findings/disorders as well as challenging erroneous tokenization of Norwegian words were identified. This work pointed out the weaknesses of the current implementation and provided insights for future work.

Introduction

Negation detectors (NegEx) are available in several languages. However, to our knowledge, Norwegian NegEx has not yet been developed. A Swedish version of NegEx (originating from an American English version [1]) was presented earlier and yielded satisfactory precision and recall when tested on Swedish medical text [2]. This paper presents the development and initial tests of the Norwegian NegEx using the Swedish NegEx as a starting point.

Methods

The Swedish version of NegEx [2] was ported to Norwegian and evaluated on a Norwegian medical scientific text in the domain of gastrointestinal surgery. The Norwegian medical scientific text in this specific domain was downloaded from the *Tidsskriftet Den norske legeforening* (https://tidsskriftet.no/spesialitet/gastroenterologisk-kirurgi) and transformed from portable document format (PDF) to pure UTF-8 coded text. In total, 170 articles containing 294,745 words were chosen.

In addition to the original Swedish version of NegEx with 40 negation rules, new 27 rules were added to the Norwegian version of the NegEx. The new rules distribution was as follows: 15 Norwegian POST-negation rules (5 in Swedish version), 34 PREN-negation rules (26 in Swedish version), and 18 PSEU-DO-negation rules (9 in the Swedish version).

A Norwegian version of ICD-10 diagnosis codes and terms (https://ehelse.no/standarder-kodeverk-og-referansekatalog/helsefaglige-kodeverk/kodeverket-icd-10-og-icd-11) was used in the negation detection process. In total, 19,021 symptoms and diagnoses from the ICD-10 list and 23 significant words from the gastrointestinal surgery domain from Table V in [3] were added.

Results

When executing NegEx on the 170 articles, the system found 70 negated symptoms/diagnoses. The NegEx was also executed on a smaller sub-corpus that was manually labelled. Results are represented in Table 1.

	Words	***Findings/disorders***	***Negations***
The whole corpus	*294,745*	*1,835*	*70*
Manually labelled sub-corpus	*75,614*	*-*	*29*
Automatically labelled sub-corpus	*75,614*	*526*	*15*

Table 1. Performance and evaluation of Norwegian NegEx applied on Norwegian scientific text.

Machine-based proper negation detection

Han var ved innkomst hemodynamisk upåvirket og hadde ikke tegn til [NEGATED]peritonitt[NEGATED].

Problems with tokenization

Erroneous tokenization:
ikke -operable [NEGATED]metastaser[NEGATED]

Correct tokenization:
ikke-operable metastaser

Future work

For evaluating the Norwegian NegEx, manual labelling of the scientific publication corpus will be continued further to ensure sufficient scope and quality of the testing data. The algorithm improvement will include a more robust mechanism to capture negated symptoms/diagnoses, which currently is based on exact string matching. When the updated NegEx is evaluated on the text corpus, its performance will be tested on electronic health records in gastrointestinal surgery from the University Hospital of North Norway.

References

[1] W. W. Chapman, W. Bridewell, P. Hanbury, G. F. Cooper, and B. G. Buchanan, “A simple algorithm for identifying negated findings and diseases in discharge summaries,” J. Biomed. Inform., vol. 34, no. 5, pp. 301–310, Oct. 2001.
[2] M. Skeppstedt, “Negation detection in Swedish clinical text: An adaption of NegEx to Swedish,” J. Biomed. Semant., vol. 2 Suppl 3, p. S3, 2011.
[3] C. Soguero-Ruiz et al., “Support Vector Feature Selection for Early Detection of Anastomosis Leakage From Bag-of-Words in Electronic Health Records,” IEEE J. Biomed. Health Inform., vol. 20, no. 5, pp. 1404–1415, 2016.

For more information, contact: Andrius Budrionis, tel +47 913 60 982, e-mail: andrius.budrionis@ehealthresearch.no