Initial Results in the Development of SCAN

A Swedish Clinical Abbreviation Normalizer

Niklas Isenius[1], Sumithra Velupillai[1], and Maria Kvist[1, 2]

[1]Dept. of Computer & Systems Sciences, Stockholm University, Sweden
[2] Dept. of Clinical Immunology & Transfusion Medicine, Karolinska University Hospital, Sweden
[nikl-ise, sumithra]@dsv.su.se, maria.kvist@karolinska.se

Abstract. Abbreviations are common in clinical documentation, as this type of text is written under time-pressure and serves mostly for internal communication. This study attempts to apply and extend existing rule-based algorithms that have been developed for English and Swedish abbreviation detection, in order to create an abbreviation detection algorithm for Swedish clinical texts that can identify and suggest definitions for abbreviations and acronyms. This can be used as a pre-processing step for further information extraction and text mining models, as well as for readability solutions.

Through a literature review, a number of heuristics were defined for automatic abbreviation detection. These were used in the construction of the Swedish Clinical Abbreviation Normalizer (SCAN). The heuristics were: a) freely available external resources: a dictionary of general Swedish, a dictionary of medical terms and a dictionary of known Swedish medical abbreviations, b) maximum word lengths (from three to eight characters), and c) heuristics for handling common patterns such as hyphenation. For each token in the text, the algorithm checks whether it is a known word in one of the lexicons, and whether it fulfills the criteria for word length and the created heuristics. The final algorithm was evaluated on a set of 300 Swedish clinical notes from an emergency department at the Karolinska University Hospital, Stockholm. These notes were annotated for abbreviations, a total of 2,050 tokens. This set was annotated by a physician accustomed to reading and writing medical records.

The algorithm was tested in different variants, where the word lists were modified, heuristics adapted to characteristics found in the texts, and different combinations of word lengths. The best performing version of the algorithm achieved an F-Measure score of 79%, with 76% recall and 81% precision, which is a considerable improvement over the baseline where each token was only matched against the word lists (51% F-measure, 87% recall, 36% precision). Not surprisingly, precision results are higher when the maximum word length is set to the lowest (three), and recall results higher when it is set to the highest (eight).

Algorithms for rule-based systems, mainly developed for English, can be successfully adapted for abbreviation detection in Swedish medical records. System performance relies heavily on the quality of the external resources, as well as on the created heuristics. In order to improve results, part-of-speech in-

formation and/or local context is needed for disambiguation. In the case of Swedish, compounding also needs to be handled.

Keywords: Automatic Abbreviation Detection; Medical Records; Clinical Text Mining

1 Introduction

Abbreviations are common in clinical documentation, as this type of text is written under time-pressure and serves mostly for internal communication. For text mining and information extraction techniques, as well as for readability for e.g. patients and caregivers from other specialties, it would be beneficial to automatically identify and expand abbreviations to their full-form counterparts. However, most research on abbreviation detection and expansion has been performed on English texts. This study attempts to apply and extend existing rule-based algorithms that have been developed for English and Swedish[a] abbreviation detection,[1-5] in order to create an abbreviation detection algorithm for Swedish clinical texts that can identify and suggest definitions for abbreviations and acronyms. This can be used as a pre-processing step for further information extraction and text mining models, as well as for readability solutions.

2 Materials and Methods

Through a literature review, a number of heuristics were defined for automatic abbreviation[b] detection and were used for the construction of the Swedish Clinical Abbreviation Normalizer (SCAN). These heuristics were:

1. freely available external resources:
   - a dictionary of general Swedish,[c]
   - a dictionary of medical terms,[d] and
   - a dictionary of known Swedish medical abbreviations,[6]
2. maximum word lengths (from three to eight characters), and
3. heuristics for handling common patterns such as hyphenation.

For each token in the text, the algorithm checks whether it is a known word in one of the lexicons, and whether it fulfils the criteria for word length and the created heuristics. The final algorithm was evaluated on a set of 300 Swedish clinical notes from an emergency department at the Karolinska University Hospital, Stockholm.[e] These notes were annotated for abbreviations, resulting in a total of 2,050. This set was annotated by a physician accustomed to reading and writing medical records.

---

[a] This study is applied on Swedish biomedical texts, not on clinical documentation.

[b] In the study presented here, acronyms are included in the definition of abbreviations.

[c] `http://runeberg.org/words/` and `http://g3.spraakdata.gu.se/saob/`

[d] `http://www.fass.se`

[e] Ethical approval is granted by the Regional Ethical Review Board in Stockholm (Etikprövningsnämnden I Stockholm), permission number 2009/1742-31/5

3 Results

The algorithm was tested in different variants, where the word lists were modified, heuristics adapted to characteristics found in the texts, and different combinations of word lengths. The best performing version of the algorithm achieved an F-Measure score of 79%, with 76% recall and 81% precision, which is a considerable improvement over the baseline where each token was only matched against the word lists (51% F-measure, 87% recall, 36% precision), see Table 1. Not surprisingly, precision results are higher when the maximum word length is set to the lowest (three), and recall results higher when it is set to the highest (eight).

Table 1. SCAN results, precision, recall and F-measure. Baseline = each token is checked against the word lists. ImprovedListLength4 = token length 4 and modified word lists.

Version	Recall %	Precision %	F-measure %
Baseline	87	36	51
ImprovedListLength3	68	**82**	75
ImprovedListLength4	76	81	**79**
ImprovedListLength8	**83**	62	71

4 Discussion

In this study, a rule-based abbreviation system tailored for Swedish medical records is presented. The system relies on lexicons and heuristics and the overall results are encouraging. Through an error analysis, different types of common errors have been identified, such as ambiguous words (e.g. *hö* which is a valid word (hay) but also a common abbreviation for *höger* (right)) and abbreviations within compounds, e.g. *lungrtg* (x-ray of lungs, x-ray abbreviated *rtg*). Compared to similar research for English (e.g. Xu et al.[1]), results are lower. Relying on word lengths and external lexicons limits coverage. In order to improve results, part-of-speech information and/or local context is needed for disambiguation, for instance, which would probably improve precision results. In the case of Swedish, compounding also needs to be handled.

5 Conclusion

Algorithms for rule-based systems, mainly developed for English, can be successfully adapted for abbreviation detection in Swedish medical records. System performance relies heavily on the quality of the external resources, as well as on the created heuristics. However, promising results are obtained without extensive tailoring of previous algorithms developed for other languages. In the case of Swedish, language-specific properties need to be addressed. Future work involves expanding the abbreviations to their definitions, where emphasis should be put on improving precision rather than recall, as precision would be more important for this task (an erroneous expansion would create unfortunate misunderstandings). In the current system definitions are provided if they exist in the list of known medical abbreviations but no evaluation has yet been performed on this part. Detecting abbreviations automatically in medical

records has great importance for information access from this type of text. In one study, extraction of disorders were hampered by abbreviations, as 14% of disorders in clinical text were written as abbreviations and was not found when matched to SNOMED CT[7]. If these could be correctly expanded to their full-length counterpart, automatic disorder extraction would of course also improve. With rule-based algorithms, very little language specific tailoring is needed, at least in the case of English versus Swedish. However, for better performance, some issues need to be handled, e.g. compound splitting and disambiguation.

Acknowledgements

The authors wish to thank the anonymous and known reviewers for their comments and feedback.

References

1. Xu, H., Stetson, P.D., Friedman, C., 2007. A Study of Abbreviations in Clinical Notes, in AMIA Annual Symposium Proceedings 2007, pp. 821–825.
2. Yeates, S., 1999. Automatic Extraction of Acronyms from Text, in Proceedings of the Third New Zealand Computer Science Research Students' Conference, pp. 117-124.
3. Park, Y., Byrd, R.J., 2001. Hybrid text mining for finding abbreviations and their definitions, in Proceedings of Empirical Methods in Natural Language Processing 2001, pp. 126-133.
4. Larkey, L.S., Ogilvie, P., Price, M.A., Tamilio, B., 2000. Acrophile: An Automated Acronym Extractor and Server, in Proceedings of the Fifth ACM Conference on Digital Libraries, pp. 205-214.
5. Dannélls, D., 2003. Acronym Recognition. Master Thesis. Department of Linguistics, Göteborg University, Sweden
6. Cederblom, S., 2005. Medicinska förkortningar och akronymer. Lund: Studentlitteratur AB.
7. Skeppstedt, M., Kvist, M., Dalianis, H., 2012. Rule-based Entity Recognition and Coverage of SNOMED CT in Swedish Clinical Text. In Proc. LREC 2012, Istanbul, Turkey.