Detection of Spelling Errors in Swedish Clinical Text

Nizamuddin Uddin and Hercules Dalianis
Department of Computer and System Science
Stockholm University
Box 7003, 164 07 Kista, Sweden
nini5311@student.su.se, hercules@dsv.su.se

Abstract

Spelling errors are common in clinical text because such text is written under pressure and lack of time. It is mostly used for internal communication. To improve text mining and other type of text processing tools, spelling error detection and correction is needed. In this paper we will count spelling errors in Swedish clinical text. The developed algorithm uses word lists for detection such as a Swedish general dictionary, a medical dictionary and a list of abbreviations. The final algorithm has been tested on Swedish clinical corpus, we obtained 12 per cent spelling errors. After error analysis of the result, it was concluded that many errors were detected by the algorithm due to inadequate word list and faulty preprocessing such as lemmatization and compound splitting. By manually removing these correct words from the list, total spelling errors were decreased to 7.6 per cent.

Key words

Swedish clinical text, spelling detection, text mining, information extraction

1. Introduction

Today hospitals produce huge amounts of electronic patient records which can not be used by automatic text processing methods as they are written in non-standardized form. Non-standardized means that it contains many spelling errors, non-standard abbreviations and acronyms, incomplete sentences and missing subjects. The text also uses lots of jargon and informal expressions. This is due to the fact that the text is being written in a time pressure by clinicians with no support of spelling and grammar correction. Some of the clinical text has been transcribed from dictation recordings.

This clinical text is useful for clinical research to develop new tools and reveal previously unknown information, and also for care givers to improve patient health. Therefore to improve the usage of this information for research, spelling errors detection is needed.

The non-standardized form makes it difficult for processing as information extraction, text summarization, decision support and statistical analysis (Meystre et al. 2008). It would be beneficial to detect these spelling errors and obtain an overview of what types of error and how many they are. Swedish clinical text has been studied partly with respect to abbreviations (Isenius et al., 2012) and vocabulary (Dalianis et al., 2009, Allvin et al., 2011) but no one have studied the number of spelling errors in Swedish clinical text. This study has the aim to detect and count spelling errors in Swedish clinical text.

2. Previous research

Detection and correction of spelling errors have previously been studied by Kukich (1992) that presented two types of techniques for detection errors in text, N-gram analysis and dictionary lookup technique. N-gram analysis technique works by examining each n-gram input string in the predefined n-gram table for the existence of the string or its frequency, on the other hand lexical lookup technique is simply search an input string in predefined list of words or lexicons.

An algorithm for detecting and correction spelling errors in Swedish ordinary text was described in Domeij et al., (1994). Patrick and Nguyen (2006) introduced a knowledge based process for detection and correction of spelling errors in clinical text written in English. The process detected 14.3 per cent SNOMED-CT codes, and 4.7 per cent new concepts were found in the Concord hospital's summary.

Recently Isenius et al, (2012) constructed an algorithm which detects abbreviations. The algorithm was applied on Swedish clinical notes from an emergency department at Karolinska University Hospital.

The problem of spelling errors in clinical text was also addressed by Ruch et al., (2003), where 10 per cent spelling errors were detected in clinical text written in French.

3. Method

A spelling error detection algorithm was developed, which uses lexical lookup and exact matching technique for detection of spelling errors in the text. The algorithm uses Swedish general and medical dictionaries for detection of spelling errors. Finally the algorithm was tested on a Swedish Clinical Corpus[1].

[1]This research has been approved by the Regional Ethical Review Board in Stockholm (Etikprövnings-nämnden i Stockholm), permission number 2012/834-31/5

The spelling detection algorithm uses seven different word lists both Swedish medical dictionaries and ordinary Swedish dictionaries, as the Parole dictionary that contains 2.8 million Swedish words, (Parole dictionary 2013), Swedish vocabulary that contains 66,000 Swedish words, (Swedish sample vocabulary 2013), general Swedish dictionary (Svenska (TXT), 2013) that contains 46,073 words, English to Swedish dictionary that contains 24,396 words (English to Swedish, 2013), Swedish stop words list that contains 113 stop words (Swedish stop words list 2013) and medical dictionaries that contain 5,062 drug names substances, Product Names contain 7,056 drug names for example aspirin, Pharmaceutical list contain 16,977 drug names so called ATC categories from FASS (2013) and ICD-10 diagnosis code list (2013). The ICD-10 code lists contain both codes and textual descriptions of the codes.

Preprocessing

The text was tokenized by the algorithm in order to increase matching of text in the available dictionaries. Then the text was normalized by a compound splitter (Sjöbergh, 2004) and the CST lemmatizer (Jongejan & Haltrup, 2005), to increase exact matching of the text.

The evaluation of the results was carried out manually by a Swedish native speaker by reading the 1,000 most frequent errors, and 1,000 most frequent correctly spelled words to evaluate the error rate of the spelling detection algorithm.

4. Results

The algorithm was applied on a Swedish Clinical Corpus the result is shown in the following table, see Table 1.

Method	Result
Total input words	151,924
Number of spelling errors	**11,584**
Snowball dictionary (Swedish dictionary)	90,424
Pharmaceuticals (Medical dictionary)	4,585
ATC categories (Medical dictionary)	599
Product names (Medical dictionary)	2,328
Parole dictionary (Swedish dictionary)	26,093
Abbreviation + Diagnosis	6,410
Eng-Swe dictionary (Swedish dictionary)	2,183
Nordic Word Svenska (TXT) Dictionary (Swedish dictionary)	1,021

Table 1. Result of Error Detection System applied on the Swedish Clinical Corpus

The algorithm detected 12 per cent spelling errors in the Swedish Clinical Corpus which is shown in Figure 2.

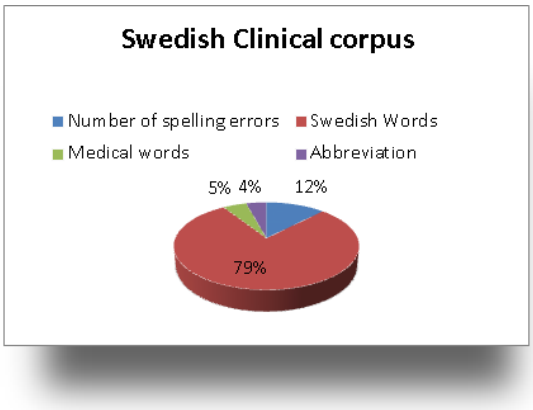

Figure 1. Spelling errors detected by the algorithm After manually analyzing the error rate (false positives) of the correctly identified words there were only 6.9 per cent spelling errors left.

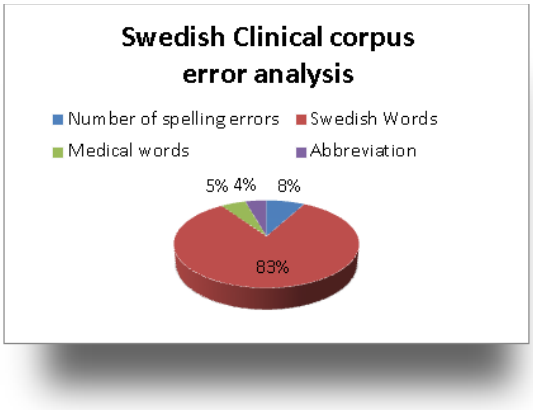

Figure 2. 1,000 most frequent spelling errors and their distribution

This gives a precision of 93 per cent of the algorithm, however the error rate for missing correctly spelled words were much higher 36.8 per cent (4,309 of 11,700 words).

The 7.6 per cent spelling errors (7,391) are after correcting for the false positives correctly spelled words that were not identified, see Table 1. 4.2 per

cent of other errors were compounds and were not de-compounded and therefore not identified and 2.9 per cent were abbreviations that were not identified.

Some common spelling error in Swedish clinical corpus are:

bätter, filtra, uppegå, någod, häel, säka, fotölj, tempa, smärtstillning, puttning, plockig, rekomendation, karnvatten, deligering, uttöka

Many of the error were due to compound splitter could not decompound medical terms as for example:

diuretika-behandling, habitual-tillstånd

(a hyphen is put to mark the compounding) or that the words were compounds of words and medical abbreviations as for example:

lungrtg, kontrollrtg, opförband

(*rtg* and *op* are abbreviations)

5. Discussion

Spelling errors are also common in other type of text as reported by Lopresti et al., (2009), Kukich, (1992) and Ehrentraut et al., (2012).

	Misspellings
Texts written in e.g.	0.2
Newspaper texts	0.05 - 0.44
Web texts	0.8
Handwritten texts	1.5 - 2.5
Typed textual	5.0 - 6.0
Patient records	10.0

Table 2. Spelling error rate in different type of text (Ehrentraut et al., 2012).

According to the above mentioned spelling error rate, error rate is much higher than other type of text. The reason is these type of text are read by a large number of people as compare to patient records which is only read by clinician.

The result support the previous research that there are around 10 per cent spelling errors detected in French clinical text as mentioned by (Ruch et al., 2003). The algorithm performance exclusively depends on the dictionaries we used. The result can be improved by adding more relevant dictionaries to the algorithm. The limitation of the developed algorithm was only to detect spelling errors not correct them.

Conclusion

A rule based algorithm was developed for Swedish clinical text in order to detect the number of spelling errors. Performance of the algorithm relies on the external resources and preprocesses and the overall result is encouraging. However the performance can be enhanced by improving external resources for example relevant dictionaries.

The lemmatization process can be improved by retraining the CST lemmatizer on Swedish medical texts instead of standard Swedish text. It was also found during the error analysis that some Swedish medical compounds were not splitted by the compound splitter and this process needs to be improved in order to minimize the error rate in the Swedish clinical text.

Future work involves the use of web instead of dictionaries for detection and also to correct the detected spelling errors by using edit distance algorithm.

Acknowledgments

We would like to thank all the members of the Clinical Text mining group at DSV/Stockholm University for assistance for lexical resources.

References

B. Jongejan and D. Haltrup. 2005. The CST Lemmatiser. Center for Sprogteknologi, University of Copenhagen version, 2, http://cst.dk/download/cstlemma/current/doc/cstlemma.pdf, (Accessed on 15th January 2014).

C. Ehrentraut, H. Tanushi, H. Dalianis and J. Tiedemann. 2012. Detection of Hospital Acquired Infections in sparse and noisy Swedish patient records. A machine learning approach using Naïve Bayes, Support Vector Machines and C4.5. In the *Proceedings of the Sixth Workshop on Analytics for Noisy Unstructured Text Data, AND, December 9, 2012 held in conjunction with Coling 2012, Bombay*

English To Swedish, 2013. http://user.meduni-graz.at/stefan.schulz/data/SemanticMiningWP20.zip (Accessed on 20th October 2014).

FASS. 2013. Swedish environmental classification of pharmaceuticals, http://www.fass.se/LIF/produktfakta/sok_lakemedel.jsp (Accessed on 6th November 2013).

H. Allvin, E. Carlsson, H. Dalianis, R. Danielsson-Ojala, V. Daudaravicius, M. Hassel D. Kokkinakis, H. Lundgren-Laine, G. H. Nilsson, Ø. Nytrø, S. Salanterä M. Skeppstedt, H. Suominen and S. Velupillai. 2011. Characteristics of Finnish and Swedish intensive care nursing narratives: a comparative analysis to support the development of clinical language technologies. *Journal of Biomedical Semantics*

2011, 2(Suppl 3):S1, doi: 10.1186/2041-1480-2-S3-S1.

H. Dalianis, M. Hassel, and S. Velupillai. 2009. The Stockholm EPR Corpus- Characteristics and Some initial findings. In *Proceedings of the 14th International Symposium for Health Information Management Research.*

ICD-10 Diagnosis Codes List. 2013. http://people.dsv.su.se/~mariask/resources/icd-10-codes.txt (Accessed on 6th November 2013).

J Patrick and D. Nguyen. 2011. Automated proof reading of clinical notes. In *Proceedings 25th Pacific Asia Conference on Language, Information and Computation*, Singapore (pp. 303-312).

J. Sjöbergh and V. Kann. 2004. Finding the Correct Interpretation of Swedish Compounds, a Statistical Approach. *The Fourth International conference on Language Resources and Evaluation, LREC*. 2004, Lisboa, Portugal.

K. Kukich. 1992. Techniques for automatically correcting words in text. *ACM Computing Surveys (CSUR)*, 24(4), 377-439

N. Isenius. S. Velupillai and M. Kvist. 2012. Initial Results in the Development of SCAN: a Swedish Clinical Abbreviation Normalizer. In *Proceedings of the CLEF 2012 Workshop on Cross-Language Evaluation of Methods, Applications, and Resources for eHealth Document Analysis - CLEFeHealth2012*, CLEF, Rome, Italy.

P. Ruch, R. Baud and A. Geissbuhler. 2003. ´Using lexical disambiguation and named-entity recognition to improve spelling correction in the electronic patient record, *Journal of Artificial Intelligence in Medicine* 29, 169-184.

R. Domeij, J. Hollman and V. Kann. 1994. Detection of spelling errors in Swedish not using a word list en clair *J. of Quantitative Linguistics 1:195-201, 1994. QUALICO-94, 71-76, 1994.*

S. Meystre, G. K. Savova, J. Kipper-Schuler and J. E. Hurdle. 2008. Extracting information from textual documents in the electronic health record: a review of recent research. *Year Med Inform*, 35, 128-44.

Swedish Parole dictionary. 2013. https://svn.spraakdata.gu.se/sb-arkiv/pub/korpusar/corpora.html/parole/parole.xml.bz (Accessed on 20th November 2013).

Swedish sample vocabulary. 2013. http://snowball.tartarus.org/algorithms/swedish/voc.txt (Accessed on 20th October 2014).

Swedish stop words list. 2013. http://snowball.tartarus.org/algorithms/swedish/stop.txt (Accessed on 6th November 2013).

Svenska (TXT) 2013 http://runeberg.org/words/fr-svenska.txt.

V. Liu and J. R. Curran. Web text corpus for natural language processing. 2006. In *Proceedings of the 11th Meeting of the European Chapter of the Association for Computational Linguistics (EACL),* 233–240, 2006.

S. Velupillai, H. Dalianis, M. Hassel and G. H. Nilsson. 2009. Developing a standard for de-identifying electronic patient records written in Swedish: precision, recall and F-measure in a manual and computerized annotation trial. International Journal of Medical Informatics (2009), doi:10.1016/j.ijmedinf.2009.04.005