Improving Readability of Swedish Electronic Health Records through Lexical Simplification: First Results

Gintarė Grigonytė[a], **Maria Kvist**[bc], **Sumithra Velupillai**[b], **Mats Wirén**[a]
[a]Department of Linguistics, Stockholm University, Sweden
[b]Department of Computer and Systems Sciences, Stockholm University, Sweden
[c]Department of Learning, Informatics, Management and Ethics, Karolinska Institutet, Sweden
gintare@ling.su.se, maria.kvist@karolinska.se,
sumithra@dsv.su.se, mats.wiren@ling.su.se

Abstract

This paper describes part of an ongoing effort to improve the readability of Swedish electronic health records (EHRs). An EHR contains systematic documentation of a single patient's medical history across time, entered by healthcare professionals with the purpose of enabling safe and informed care. Linguistically, medical records exemplify a highly specialised domain, which can be superficially characterised as having telegraphic sentences involving displaced or missing words, abundant abbreviations, spelling variations including misspellings, and terminology. We report results on lexical simplification of Swedish EHRs, by which we mean detecting the unknown, out-of-dictionary words and trying to resolve them either as compounded known words, abbreviations or misspellings.

1 Introduction

An electronic health record (EHR; Swedish: *patientjournal*) contains systematic documentation of a single patient's medical history across time, entered by healthcare professionals with the purpose of enabling safe and informed care. The value of EHRs is further increased by the fact that they provide a source of information for statistics and research, and a documentation for the patient through the Swedish Patient Data Act. EHRs collect information from a range of sources, such as administration of drugs and therapies, test results, preoperative notes, operative notes, progress notes, discharge notes, etc.

EHRs contain both structured parts (such as details about the patient, lab results, diagnostic codes, etc.) and unstructured parts (in the form of free text). The free-text part of EHRs is referred to as clinical text, as opposed to the kind of general medical text found in medical journals, books or web pages containing information about health care. Clinical texts have many subdomains depending on the medical speciality of the writer and the intended reader. There are more formal kinds of EHRs, such as discharge summaries and radiology reports, directed to other physicians, and more informal kinds such as daily notes, produced by nurses and physicians (as memory notes for themselves or for the team). In spite of the Patient Data Act, the patient is seldom seen as a receiver or reader of the document.

Linguistically, health records exemplify a highly specialised domain, which can be superficially characterised as having telegraphic sentences involving displaced or missing words, abundant abbreviations, undisputed misspellings, spelling variation which may or may not amount to misspellings depending on the degree of prescriptivism, and terminology. While this specialised style has evolved as an efficient means of communication between healthcare professionals, it presents formidable challenges for laymen trying to decode it.

In spite of this, there has been no previous work on the problem of automatically improving the readability of Swedish EHRs. As an initial attempt in this direction, we provide an automatic approach to the problem of lexical simplification, by which we mean detecting the unknown, out of dictionary words and trying to resolve them either as compounds generated from known words, as abbreviations or as misspellings. As an additional result, we obtain a distribution of how prevalent these problems are in the clinical domain.

2 Lexical challenges to readability of EHRs

A major reason for the obstacles to readability of EHRs for laymen stems from the fact that they

Proceedings of the 3rd Workshop on Predicting and Improving Text Readability for Target Reader Populations (PITR) @ EACL 2014, pages 74–83,
Gothenburg, Sweden, April 26-30 2014.

are written under time pressure by professionals, for professionals (Kvist et al., 2011). This results in a telegraphic style, with omissions, abbreviations and misspellings, as reported for several languages including Swedish, Finnish, English, French, Hungarian and German (Laippala et al., 2009; Friedman et al., 2002; Hagège et al., 2011; Surján and Héja, 2003; Bretschneider et al., 2013). The omitted words are often subjects, verbs, prepositions and articles (Friedman et al., 2002; Bretschneider et al., 2013).

Unsurprisingly, medical terminology abounds in EHRs. What makes this problem an even greater obstacle to readability is that many medical terms (and their inflections) originate from Latin or Greek. Different languages have adapted these terms differently (Bretschneider et al., 2013). The Swedish medical terminology went through a change during the 1990s due to a *swedification* of diagnostic expressions performed in the 1987 update of the Swedish version of ICD, the International Classification of Diseases[1]. For this version, the Swedish National Board of Health and Welfare decided to partly change the terminology of traditional Latin- and Greek-rooted words to a spelling compatible to Swedish spelling rules, as well as abandoning the original rules for inflection (Smedby, 1991). In this spelling reform, *c* and *ch* pronounced as *k* was changed to *k*, *ph* was changed to *f*, *th* to *t*, and *oe* was changed to *e*. For example, the technical term for cholecsystitis (inflammation of the gall bladder) is spelled *kolecystit* in contemporary Swedish, thus following the convention of changing *ch* to *k* and removing the Latin ending of *-is*. The results[2] of exact matching to *kolecystit* (English: cholecystitis) and some presumed spelling variants clearly demonstrate the slow progress (Table 1).

As medical literature is predominantly written in English nowadays, physicians increasingly get exposed to the English spelling of Latin and Greek words rather than the Swedish one. This has resulted in a multitude of alternate spellings of several medical terms. For example, *tachycardia* (rapid heart) is correctly spelled *takykardi*, but is also frequently found as *tachycardi*, *tachykardi*, and *takycardi* (Kvist et al., 2011). A similar French study found this kind of spelling variation to be abundant as well (Ruch et al., 2003).

Term	KORP	DAY	X-RAY
kolecystit	**51**	**48**	**84**
colecystit	0	1	8
cholecystit	4	88	1613

Table 1: Alternate spellings of the Swedish medical term *kolecystit* (eng. cholecystitis) in the Swedish corpus collection Korp, daily notes (DAY) and radiology reports (X-RAY), respectively. Correct spelling in bold.

EHRs also contain neologisms. These are often verbs, typically describing events relating to the patient in active form, such as ”the patient is infarcting” (Swedish: *patienten infarcerar*) instead of the unintentional ”the patient is having a myocardial infarction”. Similar phenomena are described by Josefsson (1999).

Abbreviations and acronyms in EHRs can follow standardised writing rules or be *ad hoc* (Liu et al., 2001). They are often domain-specific and may be found in medical dictionaries such as MeSH[3] and Snomed CT[4]. For instance, 18 of the 100 most common words in Swedish radiology reports were abbreviations, and 10 of them were domain-specific (Kvist and Velupillai, 2013). Because many medical terms are multiword expressions that are repeated frequently in a patient's EHR, the use of acronyms is very common. Skeppstedt et al. (2012) showed that 14% of diagnostic expressions were abbreviated in Swedish clinical text.

Abbreviations are often ambiguous. As an example, 33% of the short abbreviations in the UMLS terminology are ambiguous (Liu et al., 2001). Pakhomov et al. (2005) found that the abbreviation RA had more than 20 expansions in the UMLS terminology alone. Furthermore, a certain word or expression can be shortened in several different ways. For instance, in a Swedish intensive care unit, the drug Noradrenalin was creatively written in 60 different ways by the nurses (Allvin et al., 2011).

It should be noted that speech recognition, although common in many hospitals around the

[1] http://www.who.int/classifications/icd/en/

[2] Based on a subset of the Stockholm Electronic Patient Record Corpus (Dalianis et al., 2012) of 100,000 daily notes (DAY) written by physicians of varying disciplines (4 mill. tokens) and 435,000 radiology reports (X-RAY) written by radiologists (20 mill. tokens). KORP: http://spraakbanken.gu.se/korp/

[3] www.ncbi.nlm.nih.gov

[4] http://www.ihtsdo.org/

world, has not been introduced in Sweden, and many physicians and all nurses type the notes themselves. This is one explanation to the variation with respect to abbreviations.

User studies have shown that the greatest barriers for patients lie mainly in the frequent use of abbreviations, jargon and technical terminology (Pyper et al., 2004; Keselman et al., 2007; Adnan et al., 2010). The most common comprehension errors made by laymen concern clinical concepts, medical terminology and medication names. Furthermore, there are great challenges for higher-level processing like syntax and semantics (Meystre et al., 2008; Wu et al., 2013). The research presented in this paper focuses on lexical simplification of clinical text.

3 Related research

We are aware of several efforts to construct automated text simplification tools for clinical text in English (Kandula et al., 2010; Patrick et al., 2010). For Swedish, there are few studies on medical language from a readability perspective. Borin et al. (2009) present a thorough investigation on Swedish (and English) medical language, but EHR texts are explicitly not included. This section summarizes research on Swedish (clinical) text with respect to lexical simplification by handling of abbreviations, terminology and spelling correction.

3.1 Abbreviation detection

Abbreviation identification in English biomedical and clinical texts has been studied extensively (e.g. Xu et al. (2007), Liu et al. (2001)). For detection of Swedish medical abbreviations, there are fewer studies. Dannélls (2006) reports detection of acronyms in medical journal text with 98% recall and 94% precision by using part of speech information and heuristic rules. Clinical Swedish presents greater problems than medical texts, because of *ad hoc* abbreviations and noisier text. By using lexicons and a few heuristic rules, Isenius et al. (2012) report the best *F-score* of 79% for abbreviation detection in clinical Swedish.

3.2 Compound splitting

Good compound analysis is critical especially for languages whose orthographies concatenate compound components. Swedish is among those languages, in which every such concatenation thus corresponds to a word. The most common approach to compound splitting is to base it on a lexicon providing restrictions on how different word forms can be used for generating compounds. For example, Sjöbergh and Kann (2006) used a lexicon derived from SAOL (the Swedish Academy word list), and Östling and Wirén (2013) used the SALDO lexicon of Swedish morphology (Borin and Forsberg, 2009). With this kind of approach, compound splitting is usually very reliable for genres like newspaper text, with typical accuracies for Swedish around 97%, but performs poorer in domain specific genres.

3.3 Terminology detection

The detection of English medical terminology is a widely researched area. An example of term detection in English clinical texts is Wang and Patrick (2009) work based on rule-based and machine learning methods, reporting 84% precision.

For Swedish clinical text, Kokkinakis and Thurin (2007) have employed domain terminology matching and reached 98% precision and 87% recall in detecting terms of disorders. Using similar approaches, Skeppstedt et al. (2012), reached 75% precision and 55% recall in detecting terms of disorders. With a machine learning based approach, improved results were obtained: 80% precision, 82% recall (Skeppstedt et al., 2014). Skeppstedt et al. (2012) have also demonstrated the negative influence of abbreviations and multiword expressions in their findings.

3.4 Spelling correction

A system for general spelling correction of Swedish is described by Kann et al. (1998), but we are not aware of any previous work related to spelling correction of Swedish clinical text. An example of spelling correction of clinical text for other languages is Tolentino et al. (2007), who use several algorithms for word similarity detection, including phonological homonym lookup and n-grams for contextual disambiguation. They report a precision of 64% on English medical texts. Another example is Patrick et al. (2010) and Patrick and Nguyen (2011), who combine a mixture of generation of spelling candidates based on orthographic and phonological edit distance, and a 2-word window of contextual information for ranking the spelling candidates resulting in an accuracy of 84% on English patient records. Siklóski et al. (2013) use a statistical machine translation model

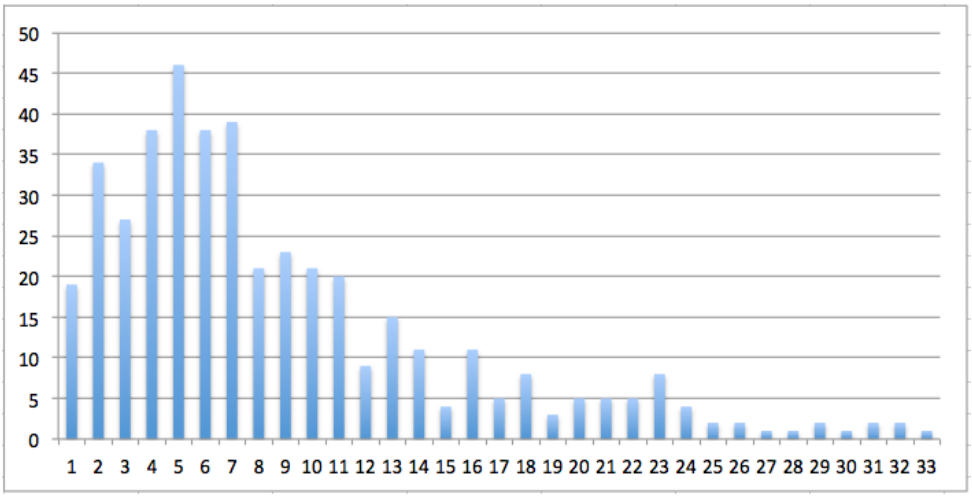

Figure 1: Distribution of 100 PR dataset sentences by length (number of sentences on the *y*-axis and number of tokens on the *x*-axis).

(with 3-grams) for spelling correction, achieving 88% accuracy on Hungarian medical texts.

4 Experimental data

This study uses clinical notes[5] from the Stockholm Electronic Patient Record corpus containing more than 600,000 patients of all ages from more than 500 health units during 2006–2013 (Dalianis et al., 2012).

A randomly selected subset of 100 daily notes from different EHRs written by physicians between 2009–2010 was used as a gold standard dataset for evaluating abbreviation detection, compound splitting and spelling corrections. This 100 daily notes dataset contains 433 sentences and 3,888 tokens, as determined by Stagger (Östling, 2013), a Swedish tokenizer and POS tagger. The majority of sentences contain between 4–11 tokens (see Figure 1.)

The text snippet in Figure 2 provides an illustrative example of the characteristics of a health record. What is immediately striking is the number of misspellings, abbreviations, compounds and words of foreign origin. But also the syntax is peculiar, alternating between telegraphic clauses with implicit arguments, and long sentences with complex embeddings.

[5]Approved by the Regional Ethical Review Board in Stockholm (Etikprövningsnämnden i Stockholm), permission number 2012/2028-31/5

5 Lexical normalization of EHRs

Normalization of lexis in clinical text relies heavily on the lookup in available lexicons, corpora and domain terminologies. Although these resources usually cover the majority of words (i.e. tokens) in texts, however due to the ever evolving language and knowledge inside the domain, medical texts, when analysed with the NLP tools, also contain *unknown*[6] words. These remaining words that are not covered by any lexicon, or corpora resource, can be misspellings, abbreviations, compounds (new word formations), words in foreign languages (Latin, Greek, English), or new terms.

Our approach to dealing with *unknown* words combines a rule-based abbreviation detection and Swedish statistical language model-based compound analysis and misspelling resolution.

The following sections describe three methods that are applied in a pipeline manner. That is, first, all known abbreviations are detected and marked; second the unknown words are checked whether they are compounds; finally, for the remaining unknown words, context dependent word corrections are made.

5.1 Detecting abbreviations

This section describes the heuristics and lexicon lookup-based abbreviation detection method. The Swedish Clinical Abbreviation and Medical Terminology Matcher (SCATM) is based on

[6]By unknown words we mean words that cannot be looked up in available lexical resources or linguistically analyzed by POS tokenizer.

Original:

Cirk och **resp** stabil, **pulm ausk** något nedsatt **a-ljud bilat**, cor **RR HF** 72, **sat** 91% på 4 **l O2**. Följer *Miktionslista*. I samråd med <title> bakjour <First name> <Second name>, som bedömmer **pat** som komplicerad *sjukdomsbild*, så följer vi vitala parametrar, samt svara han ej på *smärtlindring*, så går vi vidare med **CT BÖS**.

Swedish with expanded abbreviations and corrected misspellings:

Cirkulatoriskt och respiratoriskt stabil, pulmonales auskulteras något nedsatt andningsljud bilateralt, cor regelbunden rytm hjärtfrekvens 72, saturation 91 procent på 4 liter syrgas. Följer miktionslista. I samråd med <title> bakjour <First name> <Second name> , som bedömer patienten som komplicerad sjukdomsbild, så följer vi vitala parametrar, samt svarar han ej på smärtlindring, så går vi vidare med computed tomography buköversikt.

Literal translation to English:

Circ and **resp** stable, **pulm ausc** somewhat weak **resp** sound **bilat**, cor **RR HF** 72, **sat** 91% on 4 **l O2**. Following list for micturation. Consulting <title> senior **dr** on call <First name> <Second name> , who aseses **pat** as complicated condition, so we follow vital parameters, and anwers he not to *pain-relief*, so we go on to **CT ABD**.

English translation with expanded abbreviations (extended with missing words):

Circulatory and respiratory stable, pulmonary auscultated somewhat weak respiratory sound bilaterally, heart regular rythm frequency 72, saturation 91% on 4 liter oxygen. Following list for micturation. In consultation with <title> senior doctor on call <First name> <Second name> , who assesses patient as having complex condition, we monitor vital parameters, and if he doesn't respond to pain relief, we proceed with computed tomography of abdomen.

Figure 2: Characteristics of a health record: misspellings (underline), **abbreviations** (bold), *compounds* (italic) and words of foreign origin (red).

SCAN (Isenius et al., 2012). The SCATM method uses domain-adapted Stagger (Östling, 2013) for the tokenization and POS-tagging of text. The adapted version of Stagger handles clinical-specific[7] abbreviations from three domains, i.e. radiology, emergency, and dietology. SCATM also uses several lexicons to determine whether a word is a common word (in total 122,847 in the lexicon), an abbreviation (in total 7,455 in the lexicon), a medical term (in total 17,380 in the lexicon), or a name (both first and last names, in total 404,899 in the lexicon). All words that are at most 6 characters long, or contains the characters "-" and/or "." are checked against these lexicons in a specific order in order to determine whether it is an abbreviation or not.

The SCATM method uses various lexicons[8] of Swedish medical terms, Swedish abbreviations, Swedish words and Swedish names (first and last).

5.2 Compound splitting

For compound splitting, we use a collection of lexical resources, the core of which is a full-form dictionary produced by Nordisk språkteknologi holding AS (NST), comprising 927,000 entries[9]. In addition, various resources from the medical domain have been mined for vocabulary: Swedish SNOMED[10] terminology, the Läkartidningen medical journal[11] corpus, and Swedish Web health-care guides/manuals[12].

A refinement of the basic lexicon-driven technique described in the related research section is that our compound splitting makes use of contextual disambiguation. As the example of *hjärteko* illustrates, this compound can be hypothetically split into[13]:

hjärt+eko (en. cardiac+echo)

[7] Abbreviations that do not follow conventional orthography styles, e.g. a typical abbreviation *p.g.a.* (en. due to) can have the following variants *p g a, pga, p. G. A., p. gr. a.*

[8] the sources of lexicons are: anatomin.se, neuro.ki.se smittskyddsinstitutet.se, medicinskordbok.se, runeberg.org, g3.spraakdata.gu.se/saob, sv.wikipedia.org/wiki/Lista_ver_frkortningar, karolinska.se/Karolinska-Universitetslaboratoriet/Sidor-om-PTA/Analysindex-alla-enheter/Forkortningar/ and the list of Swedish names (Carlsson and Dalianis, 2010).

[9] Available at: www.nb.no/Tilbud/Forske/Spraakbanken/Tilgjengelege-ressursar/Leksikalske-ressursar

[10] www.socialstyrelsen.se/nationellehalsa/nationelltfacksprak/

[11] http://spraakbanken.gu.se/eng/research/infrastructure/korp

[12] www.1177.se and www.vardguiden.se

[13] Korp (http://spraakbanken.gu.se/korp) is a collection of Swedish corpora, comprising 1,784,019,272 tokens, as of January 2014.

KORP freq.: 642 + 5,669

hjärte+ko (en. beloved+cow)
KORP freq.: 8 + 8,597

For choosing the most likely composition in the given context, we use the Stockholm Language Model with Entropy (SLME) (Östling, 2012) which is a simple n-gram language model.

The max probability defines the correct word formation constituents:

hjärt+eko 2.3e-04
hjärte+ko 5.1e-07

The SMLE is described in the following section.

5.3 Misspelling detection

The unknown words that are not abbreviations or compounds can very likely be misspellings. Misspellings can be a result of typing errors or the lack of knowledge of the correct spelling.

Our approach to clinical Swedish misspellings is based on the best practices of spell checkers for Indo-European languages, namely the phonetic similarity key method combined with a method to measure proximity between the strings. In our spelling correction method, the Edit distance (Levenshtein, 1966) algorithm is used to measure the proximity of orthographically possible candidates. The Soundex algorithm (Knuth, 1973) shortlists the spelling candidates which are phonologically closest to the misspelled word. Further, the spelling correction candidates are analyzed in a context by using the SLME n-gram model.

The SLME employs the Google Web 1T 5-gram, 10 European Languages, Version 1, dataset for Swedish, which is the largest publically available Swedish data resource. The SLME is a simple n-gram language model, based on the Stupid Backoff Model (Brants et al., 2007). The n-gram language model calculates the probability of a word in a given context:

$$P(w_1^L) = \prod_{i=1}^{L} P(w_i|w_1^{i-1}) \approx \prod_{i=1}^{L} \hat{P}(w_i|w_{i-n+1}^{i-1}) \quad (1)$$

The maximum-likelihood probability estimates for the n-grams are calculated by their relative frequencies:

$$r(w_i|w_{i-n+1}^{i-1}) = \frac{f(w_{i-n+1}^{i})}{f(w_{i-n+1}^{i-1})} \quad (2)$$

The smoothing is used when the complete n-gram is not found. If $r(w_{i-n+1}^{i-1})$ is not found, then the model looks for $r(w_{i-n+2}^{i-1})$, $r(w_{i-n+3}^{i-1})$, and so on. The Stupid backoff (Brants et al., 2007) smoothing method uses relative frequencies instead of normalized probabilities and context-dependent discounting. Equation (3) shows how score S is calculated:

$$S(w_i|w_{i-k+1}^{i-1}) = \begin{cases} \frac{f(w_{i-k+1}^{i})}{f(w_{i-k+1}^{i-1})} & \text{if} f(w_{i-k+1}^{i})) > 0 \\ \alpha S(w_i|w_{i-k+2}^{i-1}) & \text{otherwise} \end{cases} \quad (3)$$

The backoff parameter α is set to 0.4, which was heuristically determined by (Brants et al., 2007). The recursion stops when the score for the last context word is calculated. N is the size of the corpus.

$$S(w_i) = \frac{f(w_i)}{N} \quad (4)$$

The SLME n-gram model calculates the probability of a word in a given context: $p(word|context)$. The following example[14] shows the case of spelling correction:

Original:
Vpl på onsdag. *UK* **tortdag**.
(en. Vpl on wednesday. UK thsday.)

torgdag (en. marketday): 4.2e-10
torsdag (en. Thursday): 1.1e-06

Corrected:
Vpl på onsdag. *UK* **torsdag**.

6 Experiments and results

Our approach to lexical normalization was tested against a gold standard, namely, the 100 EHR daily notes dataset. The dataset was annotated for abbreviations, compounds including abbreviations and misspellings by a physician.

We carried out the following experiments (see Table 2):

1. SCATM to mark abbreviations and terms;

[14] Vpl stands for *Vårdplanering* (en. planning for care), UK stands for *utskrivningsklar* (en. ready for discharge).

Method	Lexical normalization task	Gold-standard, occurences	Precision, %	Recall, %
SCATM 1	Abbreviation detection	550	**91.1**	**81.0**
SCATM 1a	Abbreviations included in compounds only	78	89.74	46.15
NoCM 1	Out-of-dictionary compound splitting	97	**83.5**	-
NoCM 1a	Out-of-dictionary compounds which include abbreviations	44	59.1	-
NoCM 2	Spelling correction	41	**54.8**	**63.12**
SCATM+NoCM	Spelling correction	41	**83.87**	**76.2**

Table 2: Results of lexical normalization.

2. NoCM (lexical normalization of compounds and misspellings as described in sections 5.2 and 5.3) to resolve compounds and misspellings;

3. The combined experiment SCATM+NoCM to resolve misspellings.

The last experimental setting was designed as a solution to deal with compounds that include abbreviations. Marking abbreviations prior to the spelling correction can help to reduce the number of false positives.

The 433 sentences contained a total of 550 abbreviations (78 of these were constituents of compound words), and 41 misspellings of which 13 were misspelled words containing abbreviations. Due to the tokenization errors, a few sentence boundaries were detected incorrectly, e.g. interrupted dates and abbreviations. Because of this some abbreviations were separated into different sentences and thus added to false negatives and false positives.

The first experiment (SCATM 1 and 1a) of detecting abbreviations achieved both high precision and recall. As a special case of demonstrating the source of errors (see SCATM 1a) is the evaluation of detecting abbreviations which are part of compounds only. The low recall is due to the design of the SCATM which does not handle words longer than 6 characters, thus resulting in compounded abbreviations like *kärlkir* or *övervak* to go undetected.

The evaluation of the second experiment (NoCM 1, 1a and 2) showed that the majority of out-of-dictionary compounds was resolved correctly (NoCM 1) and reached 83.5% precision. Errors mainly occurred due to spelling candidate ranking, e.g. even+tull instead of eventuell and compounds containing abbreviations and misspelled words. As a special case of demonstrating the source of errors of the latter (see NoCM 1a) is the evaluation of those compounds[15] only which contain abbreviations. The task of spelling correction (NoCM 2) performed poorly, reaching only 54.8% precision. This can be explained by failing to resolve misspellings in compounds where abbreviations are compounded together with a misspelled words, e.g. *aciklocvirkonc* (*aciklovir koncentrate*).

The third experiment (SCATM+NoCM) combined abbreviation detection followed by the out-of-dictionary word normalization (spelling correction and compound splitting). This setting helped to resolve the earlier source of errors, i.e. words that contain both misspelling(s) and abbreviation(s). The overall precision of spelling correction is 83.87%.

7 Conclusions

Our attempt to address the problem of lexical simplification, and, in the long run, improve readability of Swedish EHRs, by automatically detecting and resolving out of dictionary words, achieves 91.1% (abbreviations), 83.5% (compound splitting) and 83.87% (spelling correction) precision, respectively. These results are comparable to those

[15]This number of compounds is derived from the number of abbreviations included in compounds (from SCATM 1a) by selecting only those out-of -dictionary words which do not contain punctuation.

reported in similar studies on English and Hungarian patient records (Patrick et al., 2010; Siklósi et al., 2013).

Furthermore, the analysis of the gold standard data revealed that around 14% of all words in Swedish EHRs are abbreviations. More specifically, 2% of all the words are compounds including abbreviations. In contrast, and somewhat unexpectedly, only 1% are misspellings. This distribution result is an important finding for future studies in lexical simplification and readability studies of EHRs, as it might be useful for informing automatic processing approaches.

We draw two conclusions from this study. First, to advance research into the field of readability of EHRs, and thus to develop suitable readability measures it is necessary to begin by taking these findings into account and by relating abbreviations, spelling variation, misspellings, compounds and terminology to reading comprehension.

Second, as a future guideline for the overall pipeline for detecting and resolving unknown, out-of-dictionary words, we suggest handling abbreviations in a first step, and then taking care of misspellings and potential compounds. The most urgent area for future improvement of the method is to handle compound words containing both abbreviations and misspellings.

Acknowledgements

The authors wish to thank the anonymous reviewers for valuable feedback. Maria Kvist and Sumithra Velupillai were in part funded by the Vårdal Foundation, Sumithra also by the Swedish Research Council and the Swedish Fulbright commission. We thank Robert Östling who provided the POS tagger and the Stockholm Language Model with Entropy.

References

M. Adnan, J. Warren, and M. Orr. 2010. Assessing text characteristics of electronic discharge summaries and their implications for patient readability. In *Proceedings of the Fourth Australasian Workshop on Health Informatics and Knowledge Management - Volume 108*, HIKM '10, pages 77–84, Darlinghurst, Australia, Australia. Australian Computer Society, Inc.

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

L. Borin and M. Forsberg. 2009. All in the family: A comparison of SALDO and WordNet. In *Proceedings of the Nodalida 2009 Workshop on WordNets and other Lexical Semantic Resources*, pages 7–12. NEALT.

L. Borin, N. Grabar, M. Gronostaj, C. Hallett, D. Hardcastle, D. Kokkinakis, S. Williams, and A. Willis. 2009. Semantic Mining Deliverable D27.2: Empowering the patient with language technology. Technical report, Semantic Mining (NOE 507505).

T. Brants, A. C. Popat, P. Xu, F. J. Och, and J. Dean. 2007. Large language models in machine translation. In *In Proceedings of the 2007 Joint Conference EMNLP-CoNLL*, pages 858–867.

C. Bretschneider, S. Zillner, and M. Hammon. 2013. Identifying pathological findings in German radiology reports using a syntacto-semantic parsing approach. In *Proceedings of the 2013 Workshop on Biomedical Natural Language Processing (BioNLP 2013)*. ACL.

E. Carlsson and H. Dalianis. 2010. Influence of Module Order on Rule-Based De-identification of Personal Names in Electronic Patient Records Written in Swedish. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation, LREC 2010*, pages 3071–3075, Valletta, Malta, May 19–21.

H. Dalianis, M. Hassel, A. Henriksson, and M. Skeppstedt. 2012. Stockholm EPR Corpus: A Clinical Database Used to Improve Health Care. In Pierre Nugues, editor, *Proc. 4th SLTC, 2012*, pages 17–18, Lund, October 25-26.

D. Dannélls. 2006. Automatic acronym recognition. In *Proceedings of the 11th conference on European chapter of the Association for Computational Linguistics (EACL)*.

C. Friedman, P. Kra, and A. Rzhetsky. 2002. Two biomedical sublanguages: a description based on the theories of Zellig Harris. *Journal of Biomedical Informatics*, 35(4):222–235.

C. Hagège, P. Marchal, Q. Gicquel, S. Darmoni, S. Pereira, and M. Metzger. 2011. Linguistic and temporal processing for discovering hospital acquired infection from patient records. In *Proceedings of the ECAI 2010 Conference on Knowledge Representation for Health-care*, KR4HC'10, pages 70–84, Berlin, Heidelberg. Springer-Verlag.

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, Rome, Italy, September. CLEF.

G. Josefsson. 1999. Få feber eller tempa? Några tankar om agentivitet i medicinskt fackspråk.

S. Kandula, D. Curtis, and Q. Zeng-Treitler. 2010. A Semantic and Syntactic Text Simplification Tool for Health Content. In *Proc AMIA 2010*, pages 366–370.

V. Kann, R. Domeij, J. Hollman, and M. Tillenius. 1998. Implementation Aspects and Applications of a Spelling Correction Algorithm. . Technical Report TRITA-NA-9813, NADA, KTH.

A. Keselman, L. Slaughter, CA. Smith, H. Kim, G. Divita, A. Browne, and et al. 2007. Towards consumer-friendly PHRs: patients experience with reviewing their health records. In *AMIA Annu Symp Proc 2007*, pages 399–403.

D. E. Knuth, 1973. *The Art of Computer Programming: Volume 3, Sorting and Searching*, pages 391–392. Addison-Wesley.

D. Kokkinakis and A. Thurin. 2007. Identification of Entity References in Hospital Discharge Letters. In *Proceedings of the 16th Nordic Conference of Computational Linguistics (NODALIDA) 2007*, pages 329–332, Tartu, Estonia.

M. Kvist and S. Velupillai. 2013. Professional Language in Swedish Radiology Reports – Characterization for Patient-Adapted Text Simplification. In *Proceedings of the Scandinavian Conference on Health Informatics 2013*, Copenhagen, Denmark, August. Linköping University Electronic Press, Linköpings universitet.

M. Kvist, M. Skeppstedt, S. Velupillai, and H. Dalianis. 2011. Modeling human comprehension of swedish medical records for intelligent access and summarization systems, a physician's perspective. In *Proc. 9th Scandinavian Conference on Health Informatics, SHI*, Oslo, August.

V. Laippala, F. Ginter, S. Pyysalo, and T. Salakoski. 2009. Towards automated processing of clinical Finnish: Sublanguage analysis and a rule-based parser. *Int journal of medical informatics*, 78:e7–e12.

VI Levenshtein. 1966. Binary Codes Capable of Correcting Deletions, Insertions and Reversals. *Soviet Physics Doklady*, 10:707–710.

H. Liu, Y. A. Lussier, and C. Friedman. 2001. Disambiguating Ambiguous Biomedical Terms in Biomedical Narrative Text: An Unsupervised Method. *Journal of Biomedical Informatics*, 34:249–261.

S. M. Meystre, G. K. Savova, K. C. Kipper-Schuler, and John E. Hurdle. 2008. Extracting information from textual documents in the electronic health record: a review of recent research. *IMIA Yearbook of Medical Informatics 2008. 47 Suppl 1:138-154*.

R. Östling and M. Wirén, 2013. *Compounding in a Swedish Blog Corpus*, pages 45–63. Stockholm Studies in Modern Philology. New series 16. Stockholm university.

R. Östling. 2012. http://www.ling.su.se/english/nlp/tools/slme/stockholm-language-model-with-entropy-slme-1.101098 .

R. Östling. 2013. Stagger: an Open-Source Part of Speech Tagger for Swedish. *Northern European Journal of Language Technology*, 3:1–18.

S. Pakhomov, T. Pedersen, and C. G. Chute. 2005. Abbreviation and Acronym Disambiguation in Clinical Discourse. In *Proc AMIA 2005*, pages 589–593.

J. Patrick and D. Nguyen. 2011. Automated Proof Reading of Clinical Notes. In Helena Hong Gao and Minghui Dong, editors, *PACLIC*, pages 303–312. Digital Enhancement of Cognitive Development, Waseda University.

J. Patrick, M. Sabbagh, S. Jain, and H. Zheng. 2010. Spelling correction in Clinical Notes with Emphasis on First Suggestion Accuracy. In *2nd Workshop on Building and Evaluating Resources for Biomedical Text Mining*, pages 2–8.

C. Pyper, J. Amery, M. Watson, and C. Crook. 2004. Patients experiences when accessing their on-line electronic patient records in primary care. *The British Journal of General Practice*, 54:38–43.

P. Ruch, R. Baud, and A. Geissbühler. 2003. Using lexical disambiguation and named-entity recognition to improve spelling correction in the electronic patient record. *Artificial Intelligence in Medicine*, 29(1-2):169–184.

B. Siklósi, A. Novák, and G. Prószéky, 2013. *Context-Aware Correction of Spelling Errors in Hungarian Medical Documents*, pages 248–259. Number Lecture Notes in Computer Science 7978. Springer Berlin Heidelberg.

J. Sjöbergh and V. Kann. 2006. Vad kan statistik avslöja om svenska sammansättningar? *Språk och stil*, 1:199–214.

M. Skeppstedt, M. Kvist, and H Dalianis. 2012. Rule-based Entity Recognition and Coverage of SNOMED CT in Swedish Clinical Text. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation, LREC 2012*, pages 1250–1257, Istanbul, Turkey, May 23–25.

M. Skeppstedt, M. Kvist, G. H. Nilsson, and H. Dalianis. 2014. Automatic recognition of disorders, findings, pharmaceuticals and body structures from

clinical text: An annotation and machine learning study. *Journal of Biomedical Informatics*, http://dx.doi.org/10.1016/j.jbi.2014.01.012.

B. Smedby. 1991. Medicinens Språk: språket i sjukdomsklassifikationen – mer konsekvent försvenskning eftersträvas [Language of Medicine: the language of diagnose classification - more consequent Swedification sought]. *Läkartidningen*, pages 1519–1520.

G. Surján and G. Héja. 2003. About the language of Hungarian discharge reports. *Stud Health Technol Inform*, 95:869–873.

H. D. Tolentino, M. D. Matters, W. Walop, B. Law, W. Tong, F. Liu, P. A. Fontelo, K. Kohl, and D. C. Payne. 2007. A UMLS-based spell checker for natural language processing in vaccine safety. *BMC Med. Inf. & Decision Making*, 7.

Y. Wang and J. Patrick. 2009. Cascading classifiers for named entity recognition in clinical notes. In *Proceedings of the Workshop on Biomedical Information Extraction*, WBIE '09, pages 42–49, Stroudsburg, PA, USA. Association for Computational Linguistics.

D. T. Y. Wu, D. A. Hanauer, Q. Mei, P. M. Clark, L. C. An, J. Lei, J. Proulx, Q. Zeng-Treitler, and K. Zheng. 2013. Applying Multiple Methods to Assess the Readability of a Large Corpus of Medical Documents. *Stud Health Technol Inform*, 192:647–651.

H. Xu, P. D. Stetson, and C. Friedman. 2007. A Study of Abbreviations in Clinical Notes. In *Proc AMIA 2007*, pages 821–825.