Informatics for Health: Connected Citizen-Led Wellness and Population Health
R. Randell et al. (Eds.)

doi:10.3233/978-1-61499-753-5-241

Improving Terminology Mapping in Clinical Text with Context-Sensitive Spelling Correction

Juliusz DZIADEK[1] , Aron HENRIKSSON and Martin DUNELD
Department of Computer and Systems Sciences, Stockholm University

Abstract. The mapping of unstructured clinical text to an ontology facilitates meaningful secondary use of health records but is non-trivial due to lexical variation and the abundance of misspellings in hurriedly produced notes. Here, we apply several spelling correction methods to Swedish medical text and evaluate their impact on SNOMED CT mapping; first in a controlled evaluation using medical literature text with induced errors, followed by a partial evaluation on clinical notes. It is shown that the best-performing method is context-sensitive, taking into account trigram frequencies and utilizing a corpus-based dictionary.

Keywords. spelling correction, terminology mapping, clinical text

1. Introduction

The increasing adoption of electronic health records (EHRs) provides access to vast amounts of digitized healthcare data, which is potentially very valuable. There are, however, challenges in analyzing EHR data, in particular when it comes in the form of unstructured text data, which is known to be noisy and contain a high degree of shorthand and misspellings [1,2]. To facilitate the secondary use of EHR data, clinical text needs to be mapped to medical ontologies like SNOMED CT [3], which exists in multiple languages and has become the de facto standard for the representation of clinical concepts. Mapping clinical text to ontologies allows us to tap into medical knowledge and to transform unstructured data into a form that can more readily be analyzed by computers.

Mapping clinical text to ontologies and standardized terminologies is, however, nontrivial, not least due to the abundance of misspellings. Systems that perform mapping in English clinical text exist, such as cTAKES [4]. However, existing methods tend to rely largely on dictionary look-up methods, which struggle with misspellings. The performance can conceivably be improved by detecting and correcting misspellings prior to the mapping process. While spelling correction of clinical text has received some attention for English [5,6], less has been done for other languages. Here, we evaluate the use of spelling correction methods to Swedish medical and clinical text, and evaluate their impact on SNOMED mapping.

[1] Corresponding author: jmail@op.pl

2. Methods & Materials

This paper explores various algorithms for spelling correction on two Swedish corpora: (1) a literature corpus, comprising edited journal articles, and (2) a clinical corpus, comprising notes from EHRs. While the motivating use case is to improve SNOMED CT mapping in clinical text, the medical corpus allows for the creation of a synthetic reference standard under the assumption that these edited texts do not contain spelling errors. The algorithms are thoroughly evaluated on the literature corpus for their ability to (1) detect and (2) correct misspellings, as well as to what degree the SNOMED CT mapping improves with the different spelling correction strategies. The algorithms are also evaluated on the clinical corpus by quantifying the number of additional SNOMED CT mappings that were made – through exact string matching – post spelling correction, in order to see how they fare on noisier input. Finally, a small-scale manual evaluation of a context-sensitive algorithm is carried out by a domain expert.

2.1. Algorithms

Correcting misspellings can be divided into two sub-tasks: (1) misspelling detection and (2) spelling correction. For misspelling detection, the Swedish spell checker Stava [7] and medical dictionaries are used. The list of candidate misspellings produced by Stava is filtered using general and domain-specific dictionaries: tokens that do not match any dictionary entry are treated as misspellings. For spelling correction, the baseline method is context-*insensitive* and based solely on Levenshtein distance. A number of context-*sensitive* methods, inspired by [5] and [6], are then evaluated and compared to the baseline method. Here, a method is defined as context-insensitive if it is deterministic w.r.t. to a token type, i.e., yield the same result irrespective of how often and where in the corpus it occurs. In contrast, the context-sensitive methods take into account not only the token type itself but also the context in which a particular token occurs and how frequent the token types are. Below is a description of the spelling correction algorithms. Two evaluations are carried out for each algorithm, employing a Levenshtein threshold of either 1 or 3 in the retrieval of replacement candidates.

Levenshtein Distance Candidates are retrieved from a dictionary and ranked according to Levenshtein distance to the misspelling, selecting the closest one. Candidates with the same Levenshtein distance are handled according to a source dictionary prioritization, whereby domain-specific dictionaries are preferred over general dictionaries. **Trigram Frequencies** Given the context of a misspelling in the form of a word trigram, where the misspelled word constitutes the middle word (or first/last if at beginning/end of sentence), the misspelled word within the trigram is replaced by any candidate with a Levenshtein distance $\leq$ t; the candidate with the highest trigram frequency in the corpus is selected. **Trigram Frequencies + Frequent Misspellings Filtering**: Like the previous algorithm, but with the difference that it employs a corpus-based dictionary which is used to filter out frequent candidates in the misspelling detection stage. **Trigram Frequency + Corpus-Based Dictionary**: Like Trigram Frequencies but employs a corpus-based dictionary which is used to filter out frequent candidates in the misspelling detection stage and, in contrast to the previous algorithm, also in the candidates retrieval stage. **Part-of-Speech Tagging + Frequent Misspellings Filtering**: When there are multiple candidates with the same Levenshtein

distance, those with the same part-of-speech are preferred. It also uses filtering of frequent tokens in the misspelling detection stage.

2.2. Experimental Setup

The medical literature corpus former comprising articles from the Journal of the Swedish Medical Association (1996-2005) [8], a subset of which (~1.3M tokens and ~0.2M types) is used. This corpus is treated as a reference standard. The algorithms are applied to a corrupted version of the corpus, in which spelling errors have been artificially introduced. Following [6], the probability of a token being misspelled is set to 15%; the misspellings are randomly introduced according to one of the four types of Damerau errors: insertion, replacement, transposition or deletion, as well as a compound error (i.e., white-space deletion). By comparing the corrected versions of the corrupted corpus to the original corpus, we are able to calculate precision, recall and F1-score for both misspelling detection and spelling correction. It moreover allows SNOMED mapping to be evaluated. The clinical corpus contains notes (~4.4M tokens and ~0.1M types) extracted from a database of Swedish EHRs1 Both corpora are tokenized using the Swedish spellchecker Stava [7] and part-of-speech tagged with Stagger [9]. Dictionaries were compiled from the Swedish versions of SNOMED CT [3], MeSH [10] and ICD-10 [11]; but also from NPL [12]: a Swedish registry of pharmaceutical products; Läkemedelsboken [13]: the Swedish Medical Products Agency's guidelines for pharmaceutical treatment; and SALDO [14]: a lexical resource for modern Swedish written language. The main evaluation criterion with the clinical corpus is to what extent more SNOMED mappings are possible post spelling correction. As this evaluation method ignores the notion of mapping accuracy, one of the context-sensitive algorithms is manually evaluated by a senior physician, effectively providing an estimation of its effectiveness.

3. Results

The results obtained for the two corpora are presented separately, beginning with the medical literature corpus. Spelling detection performance is high, particularly in terms of precision, with a score of 99.02%; recall is 81.72% and F1-score is 0.895. In contrast, the spelling correction module performs considerably worse (Table 1). Spelling correction precision varies between 48% and 71%, while recall oscillates between 14% and 26%.

Table 1. Spelling correction on the medical literature corpus

	Threshold=1			Threshold=3		
Algorithm	**Precision (%)**	**Recall (%)**	**F1-score**	**Precision (%)**	**Recall (%)**	**F1-score**
Levenshtein	69.22	14.74	0.243	58.23	**26.03**	**0.360**
Trigram	69.42	14.77	0.244	57.54	25.73	0.356
Trigram + Filtering	70.89	14.70	0.244	48.34	21.18	0.295
Trigram + Dictionary	**71.09**	14.74	0.244	58.08	25.44	0.354
POS + Filtering	69.47	14.40	0.238	53.54	23.46	0.326

The impact on SNOMED mapping on the literature corpus is shown in Figure 1, from which we can see that all algorithms lead to at least 7% additional token types being

mapped. Trigram + Filtering yields the biggest increase: 18.96%. Using a higher Levenshtein threshold invariably leads to better performance on both spelling correction and SNOMED mapping. In terms of mapping precision, however, Trigram + Dictionary performs best. In this case, using a lower Levenshtein threshold invariably yields better results. When comparing the former, the context-sensitive ones all outperform the context-insensitive baseline.

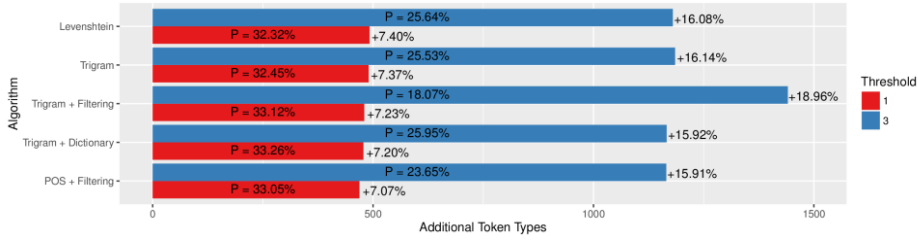

Figure 1. SNOMED mapping on medical literature.

The result of SNOMED mapping post spelling correction on the clinical corpus is shown in Figure 2, which reveals that all algorithms yield substantial mapping increases. As with the medical literature corpus, using a higher Levenshtein threshold leads to more mappings, with a context-sensitive algorithm resulting in the highest increase.

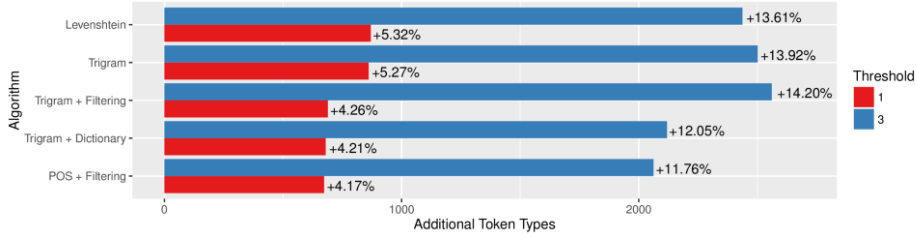

Figure 2. SNOMED mapping on clinical text.

A subset of the output (571 detected misspellings) from Trigram + Dictionary (with a threshold of 3) was then manually evaluated by a senior physician. Of these, 54.99% were categorized as spelled correctly; 38.88% were categorized as spelled incorrectly; while, for 6.13% of the tokens, the correctness could not readily be resolved. The sources of errors were spread across various token types: 68.30% were domain-specific words (including 12.78% drug names), while 17.16% were regular Swedish words. Moreover, 20.67% of the tokens were abbreviations, 21.72% inflections and 18.91% compounds. Of 222 corrected and confirmed misspellings, around 70% of the replacement candidates were marked as correct. Around 42% of the erroneous spelling correction candidates originated from the dictionary compiled from SNOMED CT.

4. Discussion

The manual evaluation revealed large differences in misspelling detection precision between the two corpora, indicating that this task is more challenging in the noisier

clinical text. It should be noted, however, that a general-purpose spell checker was employed for misspelling detection and that the only adaptation was in the form of domain-specific dictionaries. Filtering out candidates that were frequent in the corpus negatively affected performance on the medical literature corpus, probably because misspellings tend to recur. When misspellings had been correctly identified in the clinical corpus, however, replacement precision was moderate (70%). As expected, employing a lower Levenshtein threshold yields both higher spelling correction and mapping precision, while a higher threshold yields higher recall, as well as a larger number of SNOMED mappings. A context-sensitive algorithm, exploiting trigram frequency information and a corpus-based dictionary, obtained arguably the best overall results by yielding a large number of additional mappings with relatively high precision. This is encouraging, and it may be possible to obtain further improvements by taking into account additional context information: one option would be to leverage models of distributional semantics to this end. A more reliable evaluation of performance on clinical text would in the future require access to hand-annotated data.

References

[1] Helen Allvin, Elin Carlsson, Hercules Dalianis, Riitta Danielsson-Ojala, Vidas Daudaravi˘cius, Martin Hassel, Dimitrios Kokkinakis, Heljä Lundgren-Laine, Gunnar H Nilsson, Øystein Nytrø, et al. 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, 2011.

[2] Kelly Smith, Beata Megyesi, Sumithra Velupillai, and Maria Kvist. Professional language in Swedish clinical text: Linguistic characterization and comparative studies. *Nordic Journal of Linguistics*, 37(02):297–323, 2014.

[3] IHTSDO. International Health Terminology Standards Development Organisation: SNOMED CT, 2015. Accessed: July 2, 2015.

[4] Guergana K Savova, James J Masanz, Philip V Ogren, Jiaping Zheng, Sunghwan Sohn, Karin C Kipper-Schuler, and Christopher G Chute. Mayo clinical text analysis and knowledge extraction system (ctakes): architecture, component evaluation and applications. *Journal of the American Medical Informatics Association*, 17(5):507–513, 2010.

[5] Jain Zheng Patrick, Sabbagh. Spelling Correction in Clinical Notes with Emphasis on First Suggestion Accuracy. In *Proc. of the International Conference on Language Resources and Evaluation*, 2010.

[6] Patrick Ruch, Robert Bauda, and Antoine Geissbuhler. 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, 2003.

[7] Kann Hollman. En metod för svensk rättstavning baserad på bloomfilter (In Swedish), 1993. Accessed: June 19, 2015.

[8] Dimitrios Kokkinakis. The Journal of the Swedish Medical Association – a Corpus Resource for Biomedical Text Mining in Swedish. In *Proceedings of the Third Workshop on Building and Evaluating Resources for Biomedical Text Mining (BioTxtM)*, 2012.

[9] Robert Östling. Stagger: An open-source part of speech tagger for Swedish. *Northern European Journal of Language Technology* (NEJLT), 3:1–18, 2013.

[10] NLM. U.S. National Library of Medicine: MeSH (Medical Subject Headings). http://www.ncbi.nlm.nih.gov/mesh, 2015. Accessed: July 2, 2015.

[11] WHO. World Health Organization: International Classification of Diseases (ICD), 2015. Accessed: July 2, 2015.

[12] Medical Products Agency. NPL National Repository for Medicinal Products, instructions for reviewing and verifying details in the NPL, 2011. Accessed: June 19, 2015.

[13] Läkemedelsverket. Läkemedelsboken (In Swedish), 2011. Accessed: June 19, 2015.

[14] Lars Borin, Markus Forsberg, and Lennart Lnngren. The hunting of the blark - saldo, a freely available lexical database for Swedish language technology. In *Resourceful language technology. Festschrift in honor of Anna Sågvall Hein*, pages 21–32. Uppsala University, Uppsala, 2008.