Available online at www.sciencedirect.com

ScienceDirect

Computer Speech and Language 35 (2016) 219–233

COMPUTER SPEECH AND LANGUAGE

www.elsevier.com/locate/csl

Context-aware correction of spelling errors in Hungarian medical documents☆

Borbála Siklósi [a,*], Attila Novák [a,b,**], Gábor Prószéky [a,b]

[a] *Pázmány Péter Catholic University, Faculty of Information Technology and Bionics, 50/a Práter Street, 1083 Budapest, Hungary*
[b] *MTA-PPKE Hungarian Language Technology Research Group, 50/a Práter Street, 1083 Budapest, Hungary*

Received 8 November 2013; received in revised form 3 July 2014; accepted 6 September 2014
Available online 16 September 2014

Abstract

Owing to the growing need of acquiring medical data from clinical records, processing such documents is an important topic in natural language processing (NLP). However, for general NLP methods to work, a proper, normalized input is required. Otherwise the system is overwhelmed by the unusually high amount of noise generally characteristic of this kind of text. The different types of this noise originate from non-standard language use: short fragments instead of proper sentences, usage of Latin words, many acronyms and very frequent misspellings.

In this paper, a method is described for the automated correction of spelling errors in Hungarian clinical records. First, a word-based algorithm was implemented to generate a ranked list of correction candidates for word forms regarded as incorrect. Second, the problem of spelling correction was modelled as a translation task, where the source language is the erroneous text and the target language is the corrected one. A Statistical Machine Translation (SMT) decoder performed the task of error correction. Since no orthographically correct proofread text from this domain is available, we could not use such a corpus for training the system. Instead, the word-based system was used to create translation models. In addition, a 3-gram token-based language model was used to model lexical context. Due to the high number of abbreviations and acronyms in the texts, the behaviour of these abbreviated forms was further examined both in the case of the context-unaware word-based and the SMT-decoder-based implementations.

The results show that the SMT-based method outperforms the first candidate accuracy of the word-based ranking system. However, the normalization of abbreviations should be handled as a separate task.

© 2014 Elsevier Ltd. All rights reserved.

Keywords: Spelling correction; Medical text processing; Agglutinating languages

1. Introduction

Processing medical texts is an emerging topic in natural language processing. There are existing solutions, mainly for English, to extract knowledge from medical documents, which thus becomes available for researchers and medical

☆ This paper has been recommended for acceptance by R.K. Moore.

* Corresponding author. Tel.: +36 303346802.

** Corresponding author at: Pázmány Péter Catholic University, Faculty of Information Technology and Bionics, 50/a Práter Street, 1083 Budapest, Hungary.

E-mail addresses: siklosi.borbala@itk.ppke.hu (B. Siklósi), novak.attila@itk.ppke.hu (A. Novák), proszeky.gabor@itk.ppke.hu (G. Prószéky).

http://dx.doi.org/10.1016/j.csl.2014.09.001
0885-2308/© 2014 Elsevier Ltd. All rights reserved.

experts. However, locally relevant characteristics of applied medical protocols or information relevant to locally prevailing epidemic data can be extracted only from documents written in the language of the local community.

As Meystre et al. (2008) point out, it is crucial to distinguish between clinical and biomedical texts. Biomedical text processing methods, which are developed for handling published documents consider well-formed, proofread texts as their object, while this paper concentrates on clinical texts written by clinicians in clinical settings. Owing to the radical differences between these two types of textual input, methods of biomedical text processing cannot be applied to our corpus.

One of the earliest studies in processing clinical narratives, also mentioned in the report by Meystre et al. (2008), is that of Sager et al. (1994), relying on the sublanguage theory by Harris (2002). Based on this research, Friedman et al. (1995) developed MedLEE (Medical Language Extraction and Encoding System) that is used to extract information from clinical narratives to enhance automated decision-support systems. These systems are capable of creating complex representations of events found in clinical notes. Furthermore, they fulfill the expectations of extracting trustworthy information and revealing extended knowledge as well as deeper relations found in these texts. All these methods rely on proper, well-formed, and correct input documents.

In Hungarian hospitals, however, clinical records are created as unstructured texts, without any proofing control (e.g. spell checking). Moreover, the language of these documents contains a high ratio of word forms not commonly used: such as Latin medical terminology, abbreviations and drug names. Many of the authors of these texts are not aware of the standard orthography of this terminology. Thus the automatic analysis of such documents is rather challenging and automatic correction of the documents is a prerequisite of any further linguistic processing. The purpose of this paper is to present methods that can create a normalized representation of raw Hungarian clinical documents.

We investigated anonymized clinical records of a Hungarian clinic. The errors detected in the texts fall into the following categories: errors due to the frequent (and apparently intentional) use of non-standard orthography, unintentional mistyping, inconsistent word usage and ambiguous misspellings (e.g. misspelled abbreviations), some of which are very hard to interpret and correct even for a medical expert. Besides, there is a high number of real-word errors, i.e. otherwise correct word forms, which are incorrect in the actual context. Many misspelled words never or hardly ever occur in their orthographically standard form in our corpus of clinical records.

Kukich (1992) partitions the problem of spelling correction to three subcases as (a) non-word error detection; (b) isolated-word error correction; and (c) context-dependent word correction. However, most of the techniques described in the study by Kukich (1992) rely on a lexicon-based approach that is not applicable to agglutinating languages such as Hungarian. The problems of spelling correction for agglutinative languages is described by Oflazer and Güzey (1994). One way of handling an infinite vocabulary is applying finite-state automata or transducers, which are used in implementations by Park and Levy (2011), Noeman and Madkour (2010) and Pirinen and Lindén (2010). In our work, we aim at performing all three tasks in one step, that is, recognizing and correcting misspellings in context. Since the adequate use of this clinical language is never present in the documents, the goal to achieve is a quasi-standard representation (i.e. each concept represented by the same string for all occurrences) even if that spelling does not correspond to the orthographic standard.

In order to achieve this goal, a hybrid approach was chosen. Statistical or hybrid approaches are reported to outperform the previously prevailing rule-based methods. A widespread solution is to apply the noisy-channel model. The systems of Church and Gale (1991) and Brill and Moore (2000) apply variants of this model using different error models and probability scoring. Furthermore, Boswell (2004) emphasizes the beneficial use of a contextual language model in the case of spelling correction while adopting the noisy-channel model.

Although the idea of the noisy-channel model is the basis of statistical machine translation (SMT) algorithms, only very few studies use SMT implementations directly (Brockett et al., 2006; Ehsan and Faili, 2013). Still, the task of spelling correction can be modelled as a translation task, where the source language is the erroneous text and the target language is the corrected one. Moreover, a language model can be used in order to model lexical context, which is of crucial importance when choosing the appropriate item from the list of correction candidates. In a traditional SMT setup, a translation model is built from parallel training corpora and a language model from a target-language-side monolingual corpus. Since no such training data is available in our case, the translation model is replaced by a ranked list of correction candidates. These are produced by a hybrid system based on a rule-based morphological analyzer and several general and domain-specific statistical models.

Similar models are used by Turchin et al. (2007), where misspelled words are identified by comparing them to some predefined list of words. This baseline method is extended by doing prevalence analysis, i.e. determining the

Download English Version:

https://daneshyari.com/en/article/558241

Download Persian Version:

https://daneshyari.com/article/558241

Daneshyari.com