User Centred Networked Health Care
A. Moen et al. (Eds.)
IOS Press, 2011

doi:10.3233/978-1-60750-806-9-589

Truecasing Clinical Narratives

Markus KREUZTHALER[a], Stefan SCHULZ[a,b,1]
[a] *Institute for Medical Informatics, Statistics and Documentation, Medical University of Graz, Austria*
[b] *Institute of Medical Biometry and Medical Informatics, University Medical Center Freiburg, Germany*

Abstract. Truecasing, or capitalization, is the rewriting of each word of an input text with its proper case information. Many medical texts, especially those from legacy systems, are still written entirely in capitalized letters, hampering their readability. We present a pilot study that uses the World Wide Web as a corpus in order to support automatic truecasing. The texts under scrutiny were German-language pathology reports. By submitting token bigrams to the Google Web search engine we collected enough case information so that we achieved 81.3% accuracy for acronyms and 98.5% accuracy for normal words. This is all the more impressive as only half of the words used in this corpus existed in a standard medical dictionary due to the excessive use of ad-hoc single-word nominal compounds in German. Our system performed less satisfactory for spelling correction, and in three cases the proposed word substitutions altered the meaning of the input sentence. For the routine deployment of this method the dependency on a (black box) search engine must be overcome, for example by using cloud-based Web n-gram services.

Keywords. EHR, NLP, WWW

1. Background

Most significant patient-related content in electronic health records is contained in free text narratives [1, 2]. The scenarios in which these texts are produced vary across institutions, and their quality depends on the authors, the target readers, and institutional quality standards. Hastily written notes, typed by a physician or a nurse directly into the computer, tend to exhibit a lower quality in spelling, grammar, style and layout [3], compared to discharge letters, which are first dictated by a resident, then transcribed by a typist, proofread by the author, and finally validated by the staff physician before being sent out to another clinic or to the patient's GP. As well as the hurried speed in which texts are often produced, there may be technical factors responsible for the bad quality of text. Although most up-to-date text entry interfaces offer the levels of functionality users are accustomed to in modern word processors or e-mail clients, legacy systems still exist which restrict the text entry to 7bit ASCII, thus not permitting lower case characters or diacritics. As a consequence, users familiar with these systems often persist in writing in this style even after migrating to a new system.

Although it is simply a matter of time before new texts are no longer produced under these restrictions, and writers of these texts will have familiarized themselves with

[1] Corresponding Author: Stefan Schulz.

the production of correctly capitalized texts, 7bit ASCII text still continues to exist in clinical text corpora. Such legacy data is not only an important resource for retrospective research and clinical care, but also for the training of statistical natural language processing (NLP) systems [4].

The distribution of capital letters inside of a text token depends on its current context, which strongly impacts on the intelligibility of texts [5]. Practical applications of truecasing include the processing of raw input text, such as the output from speech recognition systems, as well as spelling and grammar correction systems. Just as other NLP approaches, truecasers rely on tagged corpora for the training of statistical models such as MaxEnt or SVN. Most truecasing experiments have been performed on newspaper corpora, for which the main use case was the identification of proper names characterized by initial capital letters. Languages differ in their capitalization rules, and German constitutes a special case; in contrast to most languages initial capitals are mandatory for all nouns (and nominalised adjectives and verbs) and therefore are not specific to proper names.

2. Materials and Methods

Corpus: 3,542 German-language pathology reports, containing a total of 83,818 words, were extracted from the Graz University Hospital Information System, covering a broad range of clinical disciplines. The texts had been dictated by physicians and entered by typists into a character-based user interface. The reports are entirely in upper case and do not use diacritics such as "Ä", "Ö", or "Ü".

Dictionary coverage: the coverage of these words using a German-language medical dictionary [6] was calculated.

Sampling: A random sample of 100 sentences was taken, with an average of 9.3 words (SD = 7.9; MIN = 2; MAX=38, Median = 7) per sentence. The following characters were considered sentence delimiters: [.;:!?]. Periods within abbreviations (e.g. "etc.") were not considered as delimiters.

Preparation: all remaining punctuation characters and parentheses were removed.

Gold standard: for each sentence a corrected version was created. Corrections included not only the restitution of the case, but also spelling and grammar corrections where necessary according to the 1996 German orthography reform, the medical spelling rules in accordance with German medical publishers, and [6].

Reference corpus: The case information was extracted employing the WWW as a corpus. The Google search engine was used for harvesting correct case information.

Algorithm: Each sentence with n characters was dissected into overlapping bigrams $B_1 \ldots B_{n-1}$. All bigrams are sent to the search engine as a phrase search (quoted). The hits (as displayed in bold face in the summary) of the pages returned by the search engine are saved within a map data structure. The two maps from the same token (T_{k+1} which is the second token in B_k and the first token in B_{k+1}) are merged. A weight W is assigned, directly proportional to the number of occurrences in the map and indirectly proportional to the Levenshtein edit distance [7] of the term to be corrected. The token with the maximum calculated weight is accepted as the corrected token. The edit distance is used because the search engine can also return near matches for quoted phrase searches if, for example, there are very few exact matches for that phrase on the Web.

In the case that a token is not resolved by either bigram, a single-token (quoted) search is performed. If even this search does not yield any results, the token is decapitalized (with an upper case initial character) and diacritics are restored by applying the rule ["ae"→ "ä"; "Ae"→ "Ä"; "oe"→ "ö"; "Oe"→ "Ö";"ue"→ "ü"; "Ue"→ "Ü"], according to German character transcription rules.

The algorithm was implemented in Java, using JDOM, Tagsoup and XPath (XML Path Language). The search requests were spaced by moderate delays so that the strain on the search engine was minimal.

3. Results

Table 1 shows a typical correction result and clearly visualizes the increase in readability after truecasing.

Table 1. Original text (left), automatically corrected text (right).

CHRONISCHE HEPATITIS MIT GERING BIS MITTELGRADIGER AKTIVITAET (HEPATISCHER AKTIVITAETSINDEX 6 VON 18) UND MITTELGRADIGER BIS HOEHERGRADIGER PORTALER UND MITTELGRADIGER INKOMPLETTER UND KOMPLETTER PORTOPORTALER UND PORTOZENTRALER FIBROSE (FIBROSESCORE 4 VON 6)	*Chronische Hepatitis mit gering bis mittelgradiger Aktivität (hepatischer Aktivitätsindex 6 von 18) und mittelgradiger bis höhergradiger portaler und mittelgradiger inkompletter und kompletter portoportaler und portozentraler Fibrose (Fibrosescore 4 von 6).*

A comparison of the types in the entire corpus with the Pschyrembel clinical dictionary [6], a standard reference for German clinical terminology, showed an astonishingly low lexical coverage of 51%; of 7500 types in the text corpus only 3808 match any token in the entire dictionary corpus. This result is mainly due to the high productivity in single-word compounding (a similar result can be seen in [8]) and, to a minor extent, the use of spelling variants.

Figure 1. Typical search result. The bigrams in bold face are picked by the algorithm.

Figure 1 shows a fragment of a typical search result, from which the sequences in bold face are extracted. Table 2 exemplifies the decision algorithm.

Table 2. Decision algorithm for "CHRONISCHE".

Input	***GERINGGRADIGE CHRONISCHE GASTRITIS***	
Bigram 1	*GERINGGRADIGE CHRONISCHE*	
Frequency	*Geringgradige*	7
	chronische	15
	geringgradige	6
	geringgradige″	2
Bigram 2	*CHRONISCHE GASTRITIS*	
Frequency	*Chronische*	9
	Gastritis	14
	chronische	5
Merged		
Frequency	*Chronische*	9
	Gastritis	14
	Geringgradige	7
	chronische	20
	geringgradige	6
	geringgradige″	2
Decision	*chronische*	

After the automated correction procedure, 55 of 100 sentences resulted in being equivalent to the spelling and truecasing gold standard. If equivalent expressions and acceptable spelling variants are included this rate increases to 62 and 72, respectively. In several cases it was observed that a word with standard spelling was converted to the non-standard spelling variant, as the latter occurred sufficiently more frequently on the Web. It is well known that few health professionals are perfectly proficient in spelling standard Latin. Some rules are complicated; situations require that a "c" in a Latin word stem should be converted to "k" or "z" as soon as they are no longer in a Latin syntactic context, e.g. "*Ul**c**us ventriculi*" but "*Magenul**k**us*". A synopsis of the results is given in Table 3.

Table 3. Results.

Correction Phenomenon		**Total**	**Units**
Right case correction of normal words	896	909	tokens
Right case correction of acronyms	13	16	tokens
Meaning of sentence affected by correction	3	100	sentences
Spelling / grammar error corrected	1	5	sentences
New grammar error after processing		1	sentence

The figures show an impressive accuracy of 98.5% of capitalized non-acronym tokens which were transformed into the correct case. The rate is not as good for acronyms (81.3%). Also, the procedure affected the meaning of three of the one hundred sentences. Only one of five known spelling errors was corrected, and one additional grammar error was introduced after processing.

The accuracy of 98.5% slightly outperforms the truecasing result reported by [4] on news articles. However, our method does not clearly separate between truecasing and spelling correction. This was justifiable under the constraints of our research, as the motivating factor for this work was the restrictive nature of the 7-bit ASCII character set which does not only preclude the use of lower-case characters but also of diacritics (in the case of German, mainly the "ä", "ö", "ü", and "ß" characters). The dependence on the Google Web search interface, and its non-predictable output in those cases

where there was no match, led to strange corrections such as, for instance, proposing "*maximaler*" as a correction for "*minimaler*". This distortion of a document's content is, of course, not acceptable, and challenges the unsupervised applicability of the truecasing system. In a future version we will therefore introduce a more conservative edit distance threshold for corrections (after applying the diacritic transcription rules).

Additionally, the dependence on Google Search as a black box system, which can not tolerate any major upscaling, is an unknown quantity upon which no routine system could realistically be based. An alternative would be to use Web n-gram services made available by Yahoo!, Google, and Microsoft Research [9].

4. Conclusions

We demonstrated that the use of the World Wide Web as a corpus can impressively improve the legibility of legacy texts in medical record systems that use 7-bit ASCII encoding. As the texts under scrutiny were German-language pathology reports, both German diacritics and its associated capitalization rules had to be taken into account. By submitting token bigrams to the Google Web search engine we collected enough case information so that we achieved an accuracy of 81.3% for acronyms and of 98.5% for normal words. This is all the more impressive as only half of the word types used in this corpus could be found in a comprehensive standard medical dictionary. Our system performed less satisfactory for spelling correction, and in three cases proposed word substitutions that altered the meaning of the input sentence. For the routine deployment of this method the dependency on a (black box) search engine must be overcome, for example by using cloud-based Web n-gram services.

References

[1] Barry J. Value of unstructured patient narratives. Current EHRs capture most information--patient demographics, medications and problem lists--as structured data, and often codify the details to support billing instead of clinical activities. *Health Management Technology*. 2010; 31 (7): 6-7.

[2] Schiff GD, Bates DW. Can electronic clinical documentation help prevent diagnostic errors. New England Journal of Medicine. 2010; 25; 362(12): 1066-1069.

[3] Peters AC, Nohama P, Pacheco E, Schulz S. Análise de erros de linguagem em sumários de alta.. *XII Congresso Brasileiro de Informática na Saúde*, Oct 18-22, 2010, Porto de Galinhas, Brazil: http://www.itarget.com.br/newclients/cbis2010.com.br

[4] Lita LV, et al. tRuEcasIng. *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL 2003)*, July 7-12, Sapporo, Japan.

[5] Batista F, et al. Language Dynamics and Capitalization using Maximum Entropy. *Proceedings of ACL-08: HLT*, Short Papers (Companion Volume), pages 1-4, Columbus, Ohio, USA, June 2008.

[6] Pschyrembel W. Pschyrembel Klinisches Wörterbuch Version 2. CD-ROM for Windows 3.x/95/98 de Gruyter, Bln. 1999; ISBN: 3110166208.

[7] Levenshtein VI. Binary codes capable of correcting deletions, insertions, and reversals. *In Soviet Physics. Doklady*, volume 10, pages 707-710, 1966.

[8] Schulz S, Hahn U. Morpheme-based, cross-lingual indexing for medical document retrieval. *International Journal of Medical Informatics* 2000 Sep; 58-59:87-99.

[9] Zhai, et al. Web N-gram Workshop. *Workshop of the 33rd International ACM SIGIR Conference* (2010) http://research.microsoft.com/en-us/events/webngram/sigir2010web_ngram_workshop_proceedings.pdf.