Automatic structuring and correction suggestion system for Hungarian clinical records

Borbála Siklósi, György Orosz, Attila Novák, Gábor Prószéky
Pázmány Péter Catholic University Faculty of Information Technology
H-1083 Budapest, Práter street 50/a
E-mail: siklosi.borbala@itk.ppke.hu, oroszgy@itk.ppke.hu, novak.attila@itk.ppke.hu, proszeky@itk.ppke.hu

Abstract

The first steps of processing clinical documents are structuring and normalization. In this paper we demonstrate how we compensate the lack of any structure in the raw data by transforming simple formatting features automatically to structural units. Then we developed an algorithm to separate running text from tabular and numerical data. Finally we generated correcting suggestions for word forms recognized to be incorrect. Some evaluation results are also provided for using the system as automatically correcting input texts by choosing the best possible suggestion from the generated list. Our method is based on the statistical characteristics of our Hungarian clinical data set and on the HUMor Hungarian morphological analyzer. The conclusions claim that our algorithm is not able to correct all mistakes by itself, but is a very powerful tool to help manually correcting Hungarian medical texts in order to produce a correct text corpus of such a domain.

Keywords: spelling correction, clinical text mining, language models, morphology, agglutinative, biomedical corpora

1. Introduction

In most hospitals medical records are only used for archiving and documenting a patient's medical history. Though it has been quite a long time since hospitals started using digital ways for written text document creation instead of handwriting and they have produced a huge amount of domain specific data, they later use them only to lookup the medical history of individual patients. Digitized records of patients' medical history could be used for a much wider range of purposes. It would be a reasonable expectation to be able to search and find trustworthy information, reveal extended knowledge and deeper relations. Language technology, ontologies and statistical algorithms make a deeper analysis of text possible, which may open the prospect of exploration of hidden information inherent in the texts, such as relations between drugs and other treatments and their effects. However, the way clinical records are currently stored in Hungarian hospitals does not even make free text search possible, the look-up of records is only available referring to certain fields, such as the name of the patient.

Aiming at such a goal, i.e. implementing an intelligent medical system requires a robust representation of data. This includes well determined relations between and within the records and filling these structures with valid textual data. In this paper we describe how the structure of the medical records is established and the method of automatic transformation. Basic links between individual records are also recognized, such as medical prehistory of a patient. Then, after the elimination of non-textual data, we demonstrate a basic method for correcting spelling errors in the textual parts with an algorithm that is able to handle both the language and domain specific phenomena.

2. Representation of medical texts

We were provided anonymized clinical records from various departments, we chose one of them, i.e. ophthalmology to build the system that can be extended later to other departments as well. The first phase of processing raw documents is to compensate the lack of structural information. Due to the lack of a sophisticated clinical documentation system, the structure of raw medical documents can only be inspected in the formatting or by understanding the actual content. Besides basic separations - that are not even unified through documents - there were no other aspects of determining structural units. Moreover a significant portion of the records were redundant: medical history of a patient is sometimes copied to later documents at least partially, making subsequent documents longer without additional information regarding the content itself. However these repetitions will provide the base of linking each segment of a long lasting medical process.

2.1 XML structure

Wide-spread practice for representing structure of texts is to use XML to describe each part of the document. In our case it is not only for storing data in a standard format, but also representing the identified internal structure of the texts which are recognized by basic text mining procedures, such as transforming formatting elements to structural identifiers or applying recognition algorithms for certain surface patterns. After tagging the available metadata and performing these transformations the structural units of the medical records are the followings:

- the *whole copy* in its original form of the document is stored to be used at later stages.
- *content*: parts of the records that are in free text form are further divided into sections such as header, diagnoses, applied treatments, status,

operation, symptoms, etc.

- *metadata*: applying basic pattern recognition methods we automatically tagged such units as the type of the record, name of the institution and department where it was written, diagnoses represented in tabular forms and standard encodings of health related concepts.
- *simple named entities*: at this stage of our work we only tagged basic named entities, such as dates, doctors, operations, etc. The medical language is very sensitive to named entities, that is why handling them requires much more sophisticated algorithms, which are a matter of further research.
- *medical history*: with the help of repeated sections of medical records related to one certain patient, we have been able to build a simple network of medical processes. Since the documentation of medical history is not standardized and not consequent even for the same patient, the correspondence is determined by partial string matching and comparing algorithms. Thus we can store the identifiers of the preceding and following records.

raw medical records (ophthalmology)	6.741.435
relevant content, marked with the tag content	1.452.216
textual data in content parts	422.611

Table 1: Size of each resource
(measured in number of tokens).

2.2 Separating textual and non-textual data

The resulting structure defines the separable parts of each record; however there are still several types of data within these structural units. Thus it is not possible to apply standard text processing methods on such noisy texts. Such non-textual information inserted into free word descriptions are laboratory test results, numerical values, delimiting character series and longer chains of abbreviations and special characters. We filtered out these expressions to get a set of records containing only natural text, making it possible to use natural language processing algorithms for preprocessing. Since non-textual fragments were quite diverse especially in documents originating from different doctors, or even assistants, it was impossible to develop a robust rule-based algorithm to recognize them. To solve this issue we applied the unsupervised methods of clustering algorithms. The first assumption was to tokenize documents into sentences, however due to the domain specific behaviour and the non-well-formed written representation of the texts, there are hardly any sentence boundaries in the classical sense. Our basic units were lines (i.e. units separated with newline character) and concatenations of multiple lines where neighbouring lines were suspected to be continuation of each other. This continuation does not apply to the semantic content of the lines, rather to their behaviour regarding textual or non-textual form of information. We concatenated two lines if the end of the line was a non-sentence closing punctuation mark and the beginning of the following line was not a capital letter and not a number or if the end of the first line was an inner sentence punctuation mark. Thus such short textual fragments were kept together with more representative neighbours avoiding them to be filtered out by themselves, since their feature characteristics are very similar to those of non-textual lines. We applied k-means clustering to these concatenated lines. The goal was to split into k=2 groups, however this proved to be inefficient and could not be improved by modifying the feature set. Thus we applied k=7 clustering, where two groups were flowing texts and five were different types of non-textual fragments. The labeling of the seven groups were done offline by hand, however applying a classifier trained on these data is able to recognize and separate new sets of data automatically. Testing the efficiency of our feature set and clustering algorithm, a simple Naive Bayes classifier performed 98% accuracy on a data set of 100 lines. Portions considered to be textual information need to be normalized in terms of punctuation, spelling and the used abbreviations. A fault tolerant tokenization is applied to the running text that takes into account domain specific phenomena.

text	Zavartalan korai posztoperatív szakot követően otthonába bocsátjuk, ahol javasolt kímélő életmód mellett naponta 5x1 Tobradex (tobramycin, dexamethasone) szemcsepp alkalmazása az operált szembe. Kontroll vizsgálat megbeszélés szerint, 2010. jún. 09.-n délelőtt, Dr.Benedek Szabolcsnál klinikánk ambulanciáján, illetve panasz esetén azonnal.
non-text	V.: 0,63 +0,75 Dsph -1,00Dcyl 180° = 0,8 V: 1.0 -0.5 Dsph-al élesebb V közeli: +1.5 Dsph -al Csapody III.

Table 2: Examples of textual and non-textual fragments.

3. Spelling correction

3.1 Language and domain specific difficulties

Research in the field of clinical records processing have advanced considerably in the past decades and similar applications exist for records written in English, however, these tools are not readily applicable to other languages. In our case, the problem is not only that Hungarian is another language, but agglutination and compounding, which yield a huge number of different word forms and free word order in sentences render solutions applicable to English unfeasible. E.g. while the number of different word tokens in a 10 million word English corpus is generally below 100,000, in Hungarian it is well above 800,000. However, the 1:8 ratio does not correspond to the ratio of the number of possible word forms between the two languages: while there are about 4–5 different inflected forms for an English word, there are about a 1000 for Hungarian, which indicates that a corpus of the same size is much less representative for Hungarian than it is for English (Oravecz et al., 2002.).

Moreover, medical language contains additional difficulties. Since these records are not written by clinical experts (and there is no spell-checker in the software they use) they contain many errors of the following types:

- typing errors occurring during text input mainly by accidentally swapping letters, inserting extra letters or just missing some,
- the misuse of punctuation,
- substandard spelling with especially many errors arising from the use of special medical language with a non-standard spelling that is a haphazard mixture of what would be the standard Latin and Hungarian spelling (e.g. tension / tenzió / tenzio / tensió). Though there exists a theoretical standard for the use of such medical expressions, doctors tend to develop their own customs and it is quite difficult for even an expert to choose the right form.

Besides these errors, there are many additional difficulties that must be handled in a text mining system, which are also consequences of the special use of the language. When writing a clinical record, doctors or assistants often use short incomplete phrases instead of full sentences. The use of abbreviations does not follow any standards in the documents. Assistants do not only use standard abbreviations but abbreviate many common words as well in a rather random manner and abbreviations rarely end in a period as they should in standard orthography. Moreover, the set of abbreviations used is domain specific and also varies with the doctor or assistant typing the text. In some extreme situations it might happen that a misspelled word in one document is an intentional abbreviation or short form in the other.

For the identification of an appropriate error model of the spelling errors, a corpus of corrected clinical records is needed. There is no such corpus at all for Hungarian medical language, thus we needed to create a corrected version of our real-life medical corpus. This was necessarily a partly manual process for a subset of the corpus, but we wanted to make the correction process as efficient as possible. Our goal was to recognize misspelled word forms and automatically present possible corrections in a ranked order. Additional algorithms with manual validation could then choose the final form, which is much easier than correcting the whole corpus by hand, moreover the baseline system might be easily extended to be able to carry out the whole process trained on the already corrected corpus.

3.2 Combination of language models

Aiming at such a goal, a simple linear model was built to provide the most probable suggestions for each misspelled word. We combined several language models built on the original data set and on external resources, that are the followings (the first two used as prefilters before suggesting corrections, the rest were used for generating the suggestions):

- *stopword list*: a general stopword list for Hungarian was extended with the most common words present in our medical corpus. After creating a frequency list, these words were manually selected.
- *abbreviation list*: after automatically selecting possible abbreviations in the corpus, the generated list was manually filtered to include the possible abbreviations. Since we have not applied expert knowledge, this list should be more sophisticated for further use.
- *list of word forms licensed by morphology*: those word forms that are accepted by our Hungarian morphology (HUMor (Prószéky et al. 2005.)) were selected from the original corpus, creating a list of potentially correct word forms. To be able to handle different forms of medical expressions, the morphology was extended with lists of medicine names, substances and the content of the Hungarian medical dictionary. We built a unigram model from these accepted word forms.
- *list of word forms not licensed by morphology*: the frequency distribution of these words were taken into consideration in two ways when generating suggestions. Ones appearing just a few times in the corpus remained as unaccepted forms (transforming their frequency value to 1 - original frequency). Those ones however, whose frequency was higher than the predefined threshold were considered to be good forms, even though they were denied by the morphology. Our assumption was that it is less possible to consequently use the same erroneous word form than being that form correct and contradicting our morphology.
- *general and domain specific corpora*: we built unigram models similar to that of the above described licensed word forms from the Hungarian Szeged Korpusz and from the descriptions of the entities in the ICD coding system documentation. We assumed that both of these corpora contains only correct word forms.

After having these models created, the text to be corrected was tokenized with a language independent tokenizer that is able to handle abbreviations keeping the punctuations and letters together as one token if necessary and is robust in this aspect. The tokenizer is insensitive for punctuation errors, at the presence of any non-alphanumeric character it creates a new token. The creation of such a tool was motivated by the special language requirements and the frequent occurrence of punctuation errors. The tokenizer uses a general list of abbreviations and the aforementioned domain specific list.

Model	size	example
stopword list	36	az
abbreviation list	1.251	alk
licensed by morphology	4.850	pupilla
not licensed by morphology	1.660	látsziuk
Szeged Korpusz	114.205	szeretnék
ICD corpus	3.209	betegségekben

Table 3: Size of each language model and resource (measured in number of tokens) with examples.

3.3 Generating possibly correct suggestions

As the next step we filtered out those word forms that are not to be corrected. These were the ones contained in the stopword and abbreviation lists. For the rest of the words the correction suggestion algorithm is applied. For each word a list of suggestion candidates is generated that contains the word forms with one unit of Levenshtein distance difference (Levenshtein, 1965) and the possible suggestions generated by the morphology. Then these candidates are ordered with a weighted linear combination of the different language models, the weight of the Levenshtein generation and the features of the original word form. Thus a weighted suggestion list is generated to all words in the text (except for the abbreviations and stopwords), but only those will be considered to be relevant, where the score of the best weighted suggestion is higher than that of the original word. At the end we considered the ten best suggestions.

4. Results

We investigated the performance of the system as a standalone automatic correcting tool, accepting the best weighted suggestion as the correction, but also as an aiding system that is only to help manual correction at this initial state. Since we did not have a correct corpus, we had to create one manually by correcting a portion of our medical corpus. Our test set contained 100 paragraphs randomly chosen from the corpus. When creating the gold standard from this set, there were disagreements even between human correctors, that is why in several cases we had to accept more than one word form as correct. The normalization of these forms is a task of further research. We used three metrics for evaluation:

- *precision*: measures how the number of properly corrected suggestions relates to the number of all corrections, considering the best weighted suggestion as correction.
- *recall*: measures the ratio of the number of properly corrected suggestions and the number of misspelled word forms in the original text.
- *f-measure*: the average of the above two

We investigated the result measures for several combinations of weighting the above described models and features:

- *Models based on justification of morphology (VOC, OOV):* since these models are the most representatives for the given corpus, these models were considered with the highest weight.
- *Models built from external resources (ICD, Szeged):* these models are bigger, but they are more general, thus word forms are not that relevant for our raw texts. Our results reflect that though these models contribute to the quality of the corrections, they should have lower weights in order to keep the scores of medical words higher.
- *Original form (ISORIG, ORIG):* the original forms of the words received two kinds of weighting. First we scored whether if the word to be corrected is licensed by the morphology or not. The second weight was given to the original word form in the suggestion list, regardless of its correctness. This was introduced so that the system would not “correct” an incorrect word form to another incorrect form, but rather keep the original one, if no real suggestions can be provided.
- *Morphological judgment on suggestions (HUMor):* each generated suggestion licensed by the morphology received a higher weight to ensure that the final suggestions are valid words.
- *Weighted Levenshtein generation (LEV):* when generating word forms that are one Levenshtein distance far from the original one, we gave special weighting for more probable phenomena, such as swapping letters placed next to each other on the keyboard of a computer (e.g.: n-m, s-d, y-z), improper use of long and short forms of Hungarian vowels (e.g.: o-ó, u-ú, ö-ő), mixing characteristic letters of Latin (e.g.: t-c, y-i).

The best combination of weights resulting in the best result for automatic correction, i.e. evaluated on the first highest scored suggestions is displayed in table 4.

Model	weights
OOV	0.05
VOC	0.25
Szeged	0.15
ICD	0.2
HUMOR	0.15
PRECISION	**70%**
RECALL	**75%**
F-MEASURE	**72%**

Table 4: Evaluation results for the best combinations of the applied models.

Example sentence 1 and correction:
A beteg intraorbitalis *implatatumot* is kapott ezért klinikánkon szeptember végén,október elején előzetes *telefonnegbeszélés* után kontrollvizsgálat javasolt.
A beteg intraorbitalis *implantatumot* is kapott ezért klinikánkon szeptember végén,október elején előzetes *telefonmegbeszélés* után kontrollvizsgálat javasolt.
Example sentence 2 and correction:
Meibm mirgy nyílások helyenként sárgás *kupakszeráűen* elzáródtak, ezeket megint *túvel* megnyitom
Meibom mirigy nyílások helyenként sárgás *kupakszerűen* elzáródtak, ezeket megint *tűvel* megnyitom

Table 5.a: Examples of automatically corrected sentences.

implatatumot	telefonnegbeszélés	Meibm	mirgy	kupakszeráűen	túvel
'implantatumot' : 5.60144363762e-05	'telefonmegbeszélés' : 5.87158540802e-05	'meibom' : 0.000105652387431	'mirigy' : 9.03702080337e-05	'kupakszerűen' : 5.87158540802e-05	'tűvel' : 5.88118697623e-05
'implatatumot' : 5.33130186722e-05	'telefonnegbeszélés' : 5.33130186722e-05	'meibm' : 5.06116009682e-05	'miragy' : 5.87158540802e-05	'kupakszervűen' : 5.87158541753e-05	'tevel' : 5.87158540802e-05
'ímplatatumot' : 1.875e-05	'telefónnegbeszélés' : 1.875e-05	'meíbm' : 1.875e-05	'mirgy' : 5.06116009682e-05	'kupakszeráűen' : 5.06116009682e-05	'tővel' : 5.87158540802e-05
'implatatumót' : 1.875e-05	'telefonnegbeszéléz' : 1.40625e-05	'meybm' : 1.40625e-05	'mírgy' : 1.875e-05	'kúpakszeráűen' : 1.875e-05	'túvel' : 5.06116009682e-05
'implatatúmot' : 1.875e-05	'telefonnegbessélés' : 1.40625e-05	'meilbm' : 4.6875e-06	'myrgy' : 1.40625e-05	'kupakszeráüen' : 1.875e-05	'tuvel' : 1.875e-05

Table 5.b: Detailed results of suggestions for the misspelled words in the above sentences.

The low numerical values in the table can be explained by several phenomena. The relatively small size of our test set does not reflect all types of errors. However manually creating a larger corrected text is very time and effort consuming. The system though provides great help in this task as well, so the evaluation of a generalized application will be much more accurate. Domain specific ambiguities also cause trouble at the time of evaluation. We allow the system to accept more than one correction as appropriate, but still there are several cases, where this is still a problem to decide. Thus the system might reject some correct forms while accepting other erroneous ones. The precise handling of abbreviations is still a problem, but is to be solved later on, thus it is unavoidable to fail on such fragments like “szemhéjszél idem, mérs. inj. conj, l.sin.” or “Vitr. o.s. (RM) abl. ret. miatt”. Human evaluation instead of the used metrics predicts much better results, which means that the readability of the texts has significantly improved.

Regarding the ranking of the suggestions, in 99.12% of the words of the test set, the 5 best suggestions contained the real correction. This means that using the system as an aiding tool for manual correction of medical texts is very powerful. An interactive user interface has been created to exploit the possibilities provided by such a feature, where the user can paste portions of medical texts, than the system highlights the words that it judged to be misspelled and offers the 5 best suggestions, from among which the user can choose. The scores are also displayed to give a hint to the user about the difference between each suggestions.

5. Further plans

The system at its early phase has several shortcomings regarding the generation and weighting of suggestions. Several problems are discussed above, besides which two more problems are to be solved in the near future. The first is that we are not yet taking into account the context of a word. This could solve some ambiguous cases, where no decision can be made on the word level. Considering the context as an affecting feature is also related to the task of deciding whether if a word form is an abbreviation or an incorrect word. The main difficulty for introducing this factor in the model is that a proper n-gram model is needed, which points back to the need of a correct corpus. The other important issue is that of multiple-word expressions. At its present stage the system is not able to correct such cases, when two words are written together without space between them, or vice versa. There is however a theoretical disagreement about such events, since several multiword expressions are used by doctors as one word expressions, though the standard would require them to use separately. Still these phenomena should be handled. As our test set contains examples for all these unhandled appearances, the evaluation metrics would surely be improved if these problems were solved.

6. Conclusion

The primary goal of developing our baseline algorithm was to aid the creation of a correct, reliable Hungarian medical text corpus. Having reached this goal, a more precise error model can be built to use for training a more improved system. As the results reflected, this motivation is fulfilled, since our correcting algorithm is quite efficient for such a basic aspect. The result of the system by itself could lead to several useful applications, such as at the background of a medical search engine, where both the query, and the actual result texts could be extended by other suggested forms of each word, making it possible to retrieve valuable information even if some misspellings are present on either side. The basic tagging and structuring described in the first part of this paper is also useful for storing, organizing and easier retrieving of the data. We demonstrated that the creation of an intelligent clinical system built on the knowledge lying in medical records is not trivial even in the preprocessing phase. However after some iterative application of the combination of automatic and manual work, a gradually improved corpus can be available, finally making the whole process automatic.

7. Acknowledgment

This work was partially supported by TÁMOP-4.2.2/B-10/1-2010-0014.

8. References

Brill, E., Moore, R.C. (2000). An improved error model for noisy channel spelling correction. *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*, pp. 286—293.

Contractor, D., Faruquie, T.A., Subramaniam,L.V. (2010). Unsupervised cleansing of noisy text. *Proceedings of the 23rd International Conference on Computational Linguistics*, pp. 189--196.

Farkas, R., Szarvas, Gy. (2008). Automatic Construction of rule-based ICD-9-CM coding systems. *BMC Bioinformatics*, 9

Heinze, D.T., Morsch, M.L., Holbrook, J. (2001). Mining Free-Text Medical Records. *A-Life Medical, Incorporated*, pp.254—258.

Levenshtein V. (1965). Binary codes capable of correcting spurious insertions and deletions of ones. *Problems of Information Transmission*, 1(1): pp. 8—17.

Mykowiecka, A., Marciniak, M. (2006). Domain-driven automatic spelling correction for mammography reports. *Intelligent Information Processing and Web Mining Proceedings of the International IIS: IIPWM'06. Advances in Soft Computing, Heidelberg*

Oravecz, Cs., Dienes, P. (2002). Efficient Stochastic Part-of-Speech Tagging for Hungarian. *Third International Conference on Language Resources and Evaluation*, pp. 710—717.

Patrick J., Sabbagh, M., Jain, S., Zheng, H. (2010). Spelling Correction in Clinical Notes with Emphasis on First Suggestion Accuracy. *2nd Workshop on Building and Evaluating Resources for Biomedical Text Mining*, pp. 2--8.

Pirinen, T.A., Lindén, K. (2010). Finite-State Spell-Checking with Weighted Language and Error Models – Building and Evaluating Spell-Checkers with Wikipedia as Corpus. *SaLTMiL Workshop on Creation and Use of Basic Lexical Resources for Less-Resourced Languages, LREC 2010*, pp.13—18.

Prószéky, G., Novák, A. (2005). Computational Morphologies for Small Uralic Languages. *Inquiries into Words, Constraints and Contexts*, pp 150—157.

Rebholz-Schuhmann, D., Kirsch, H., Gaudan, S., Arregui, M., Nenadic, G. (2005). Annotation and Disambiguation of Semantic Types in Biomedical Text: a Cascaded Approach to Named Entity Recognition. *Proceedings of the EACL Workshop on Multi-Dimensional Markup in NLP*.

Stevenson M., Guo, Y., Al Amri, A., Gaizauskas, R. (2009). Disambiguation of biomedical abbreviations. *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing*, pp. 71.