Restoring the intended structure of Hungarian ophthalmology documents

Borbála Siklósi[2] **and Attila Novák**[1,2]
[1]MTA-PPKE Hungarian Language Technology Research Group,
[2]Pázmány Péter Catholic University
Faculty of Information Technology and Bionics
50/a Práter street, 1083 Budapest, Hungary
{siklosi.borbala, novak.attila}@itk.ppke.hu

Abstract

Clinical documents have been an emerging target of natural language applications. Information stored in documents created at clinical settings can be very useful for doctors or medical experts. However, the way these documents are created and stored is often a hindrance to accessing their content. In this paper, an automatic method for restoring the intended structure of Hungarian ophthalmology documents is described. The statements in these documents in their original form appeared under various subheadings. We successfully applied our method for reassigning the correct heading for each line based on its content. The results show that the categorization was correct for 81.99% of the statements in our testset, compared to a human categorization.

1 Introduction

Documents created in clinical settings contain a large amount of practical information characteristic of the local community. Collecting and processing such documents may provide doctors and medical experts a valuable source of information (Meystre et al., 2008; Sager et al., 1994; Friedman et al., 1995).

In a broad sense, there are two sources of clinical documents regarding the nature of these textual data. First, they might be produced through an EHR (Electronic Health Records) system. In this case, practitioners or assistants type the information into a predefined template, resulting in structured documents. The granularity of this structure might depend on the actual system and the habit of its users. The second possibility is that the production of these clinical records follows the nature of traditional hand-written documents, i.e. even though they are stored in a computer, it is only used as a typewriter, resulting in raw text, having some clues of the structure only in the manual formatting. These are the two extremes, and the production of such records is usually somewhere in between, depending on institutional regulations, personal habits and the actual clinical domain.

In this paper, an automatic method is described that is able to assign labels of structural units to statements in Hungarian ophthalmology documents. In Hungarian hospitals, the usage of EHR systems is far behind expectations. Assistants or doctors are provided with some documentation templates, but most of them complain about the complexity and inflexibility of these systems. This results in keeping their own habit of documentation, filling most of the information into a single field and manually copying patient history.

Moreover, ophthalmology has been reported to be a suboptimal target of application of EHR systems in several surveys carried out in the US (Chiang et al., 2013; Redd et al., 2014; Elliott et al., 2012). The special requirements of documenting a mixture of various measurements, some of them resulting in tabular data, while others in single values or textual descriptions make the design of a usable system for storing ophthalmology reports in a structured and validated form very hard.

2 The corpus of Hungarian ophthalmology notes

We were provided with anonymized clinical records from the ophthalmology department of a Hungarian clinic. Due to the lack of a sophisticated clinical documentation system, the structure of the raw documents can only be inferred from the formatting or by understanding the actual content. Besides basic separations – that are not even uniform through documents – there were no other clues for determining structural units. Moreover, a significant portion of the records were redundant:

Proceedings of the 2015 Workshop on Biomedical Natural Language Processing (BioNLP 2015), pages 152–157,
Beijing, China, July 30, 2015.

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. 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. Figure 1 shows a document after processing, but the original format is kept in the example and the English translation is provided.

The documents of ophthalmology investigated in this research were especially characterized by nontextual information interspersed with sections containing texts. These (originally tabular) data behave as noise in such a context. Non-textual information inserted into free-word descriptions includes laboratory test results, numerical values, delimiting character series and longer chains of abbreviations and special characters. Moreover, these statements do not follow any standard patterns even by themselves and they further vary from document to document according to the style of the doctor or assistant.

Regarding the textual parts of the documents, these are also quite different from general Hungarian. Consult Siklósi et al. (2014) and Siklósi and Novák (2014) for a detailed comparison of Hungarian and the medical sublanguage.

3 Structuring and categorizing lines

First, a preprocessing chain adapted to these special characteristics was applied to the documents, which included tokenization (Orosz et al., 2013), spelling correction (Siklósi et al., 2014), and part-of-speech tagging (Orosz et al., 2014). Thus, an enriched representation of the corpus was achieved. This provided the basis for structuring and categorizing the content of each document. This was performed in two steps. First, formatting clues were recognized and labelled. Second, each line was classified into a content unit defined on statistical observations from the corpus.

3.1 Structuring

Even though the documentation system used when creating these documents did provide a basic template for labelling each section of the document to be created, these were very rarely followed by the administrative personnel. However, some of these system generated labels were printed into the final documents, which we could consider as 'clues' of the intended structure. These system generated labels followed a consistent pattern, and as such, could easily be recognized based on features such as the amount of white space at the beginning of the line, capitalization, and the recurring text of the headline. Thus, such structural units were identified and labelled with a `PART` tag.

Similarly, tables of codes were also printed by the system in a predefined format. These tables contain the BNO-codes (the Hungarian system of ICD coding) of diagnoses and the applied treatments. Such tables, though printed as raw text, could also be recognized by the spacing used in them and were labelled with an `SPART` tag.

3.2 Detecting patient history

We found it very often that findings about a patient recorded in documents of earlier visits were copied to the actual record, and in some cases minor adjustments were also introduced during the replication. Thus, although these partial recurrences contain only redundant information, they could not be recognized by simply looking for exact matches. Moreover, the short and dense statements of findings are often formatted the same way in the case of different patients or even doctors. In order to filter these copied sections, first we detect all date stamps in each document. Date stamps may occur in the headers, in the notation of some examinations, in the tables of codings or might be inserted manually at any point in the documents. The dates were labelled with a `DATE` tag. Then, the contents between these tags were ordered in increasing order and partial matches were found by comparing the `md5` coded form of each part. Those sections that had a matching under an earlier date stamp, were labelled with a `COPY` tag. Furthermore, these `DATE` tags were used to partition each document corresponding to separate visits. Thus, patient history could be retrieved by referring to the same ID and each date. All the information that was originally in a single document can thus be retrieved in order.

3.3 Categorizing statements

Even though the `PART` tags have labelled each part according to the documentation template of the system, the title of these fields is rarely in accordance with the content. For example, the status field is frequently used to include all the in-

formation, be it originally anamnesis, treatment, therapy, or any other comments. Thus, it was necessary to categorize each statement in each part of the documents. Table 1 shows the categories and their description used for classification. It should be noted, that these categories are defined directly for the ophthalmology domain. For other specialties, the tagset should be redefined.

Prior to categorization, units of statements had to be declared. The documents were exported from the original system in a way that kept the fixed width of the original input fields. Thus, linebreaks were inserted to the text at certain positions corresponding to this width. In order to restore the original units intended to be single lines, these linebreaks were deleted from the end of a line which could be continued by the next one. That is, if the second line does not start with capital letter, does not start with whitespace and if the length of the actual line plus the length of the first word of the second line is larger than the fixed width (hyphenation was not implemented in the system, thus if a word would pass the right margin, then the whole word is transmitted to a new line). Moreover, lines containing tabular data were also recognized during this processing step. The units of categorization were these concatenated lines and since these lines were either short or contained usually only one type of information, each received one tag. Longer sections of neighboring lines falling into the same category could be merged after labelling each line. The categorization was done in three steps.

First, using the preprocessed version of the texts, some patterns were identified based on part-of-speech tags and the semantic concept categories assigned to the most frequent entities. For example, due to the rare use of verbs, if a past tense verb was recognized in a sentence, it was a good indicator of being part of the anamnesis or the complaints of the patient (Siklósi, 2015).

Second, some indicator words were extracted from the documents. At the first place, these were those line initial words and short phrases that started with capital letter and were followed by a colon and some more content. These phrases were then ordered by their occurrence frequencies. Then, they were manually assigned a category label referring to the type of the statement that the phrase could be an indicator of. For example the phrase, *korábbi betegségek* 'previous illnesses' was given the label `Ana` referring to anamnesis. Table 2 shows some more examples of tags and phrases labelled by them. After having all the phrases occurring at least 10 times in the whole corpus labelled, they were matched against the lines of each document that were found in `PART` sections and were not recognized as tabular data. If the line started with a phrase or any of its variations (case variations, misspellings, punctuation marks and white spaces were allowed differences), then the line was labelled with the tag the phrase belonged to. These first two steps were able to categorize 34% of the concatenated lines in the documents.

tag	phrase	English translation
`Ana`	egyéb betegség	other illness
	panasz	complaint
	család	family
	korábbi	earlier
	hypertonia	hypertonia
	anamnézis	anamnesis
`T`	eredmény	result
	ultrahang	ultrasound
	Topo	Topo
	Schirmer	Schirmer
`RL`	réslámpa	slit lamp
	macula	macula
	fundus	fundus
	rl	sl (for slit lamp)
	lencse	lens
`Ther`	th	th (for therapy)
	szemcsepp	eyedrop
	terápia	therapy
	rendelés	prescription
	javasolt	recommended

Table 2: Examples of tags and some of the phrases labelled by the tag.

In the third step, the rest of the lines were given a label. In order to do this, all lines labelled in the first two steps were collected for each tag (they will be referred to as tag collections). Then, for each line, the most similar tag collection was determined and the tag of this collection was assigned to the actual line. The similarity measure applied was the tf-idf weighted cosine similarity between a line (l) and a tag collection (c) defined by Formula 1.

tag	meaning	description
Tens	Tension	Measurements of the tension of the eye
V/Refr	Refraction	Refraction data
Ana	Anamnesis	Complains of the patient, other/past diseases, family history, etc.
Dg	Diagnosis	The actual diagnoses
Beav	Treatment	Applied treatments, except operations and medication
Vél	Opinion	Opinion of the doctor, except diagnoses and treatments
St	Status	Actual status of the patient
Ther	Therapy	Prescribed/applied medication
BNO	BNO (ICD)	Statements used with their BNO codes
T	Test	Tests, other than those in the Rl category
V	Visus	Visus data
Rl	Slit lamp	Tests carried out using the slit lamp (most of the tests are done with it)
Kontr	Control	Information about when the patient should return to the doctor
Mütét	Operation	Operations applied or prescribed
XXX	-	Other statements that can not be categorized

Table 1: The tags used in categorizing statements

$$sim(\vec{l},\vec{c}) = \frac{\sum\limits_{w\in l,c} tf_{w,l}tf_{w,c}(idf_w)^2}{\sqrt{\sum\limits_{l_i\in l}(tf_{l_i,l}idf_{l_i})^2}\times\sqrt{\sum\limits_{c_i\in c}(tf_{c_i,c}idf_{c_i})^2}} \quad (1)$$

, where $\vec{l}$ contained the normalized set of words in line l, and $\vec{c}$ the normalized set of words contained in the tag collection c. During normalization, stopwords and punctuation marks were removed and numbers were replaced by the character *x*, so that the actual numerical values do not mislead the representation. As a result, all lines within PART sections were labelled by a tag. Finally, tabular lines were assigned the tag Vis, since these contained the detailed information about the visual acuity of the patient.

4 Results

The labels of 1000 lines were checked manually. This testset was selected randomly only from PART sections, since the categorization was applied only to these portions of the documents. However, the label XXX was also allowed in the system when it was not able to assign any meaningful labels. The rest of the lines were assigned one of the 15 labels. Figure 1 shows the processed state of a document. In the example, the format and whitespaces of the original document is kept. Tags are shown at the beginning of the lines. Tags starting with a number of # symbols are used for the separation of structural units. Categorization is applied to lines in a Part section, here Part:Státusz. Lines ending with an @ symbol were concatenated with the next line. tags regarding structural units and classification of statements. The English translation for the meaningful parts are inserted between the lines.

In the evaluation setup, the labels were considered either as correct, non-correct or undecidable. Lines of this latter category either did not include enough information referring to the content, or it was too difficult even for the human evaluator to decide what category the line belonged to. The label XXX was accepted as correct, if the line did not belong to any category (e.g. a single date). Out of the 1000 lines in the test set, its 7.8% could not be categorized by the human expert. For the rest of the lines, 81.99% of these lines were assigned the correct label and only 18.01% the incorrect one. Regarding the errors, most of them were due to the lack of contextual information for the algorithm. For example, if the anamnesis of a patient included some surgery, then the label for surgery was assigned to it, which is correct at the level of standalone statements, but incorrect in the context of the whole document. The other main source of the errors was that some longer lines included more than one types of statements and the system was unable to choose a correct one. In these cases, the human annotation assigned the "more relevant" tag as correct. Thus, a significant part of these errors could be eliminated by a more ac-

```
###DOCTYPE:AMBULÁNS KEZELŐLAP
'T                                A M B U L Á N S    K E Z E L Ő L A P

###PART:Státusz //Status
St       Státusz

##DOCDATE##
##DATE-TIME##
XXX 'T        2010.10.19 12:28   Székelyhidi/Füst

Beav 'C        Olvasó szemüveget szeretne. Néha könnyeznek a szemei.
                    //S/he would like reading glasses, eyes are sometimes watering.
V         V:0,7+0,75Dsph=1,0
V           1,0 +0,5 Dsph élesebb

V\Refr         +2.0 Dsph mko Cs IV

St         St.o.u: halvány kh, ép cornea, csarnok kp mély tiszta, iris ép békés, pupilla@
                    // St.o.u: blanch conj, intact cornea, chamber deep clean, iris intact, calm, papil
    'C   rekciók rendben, lencse tiszta, jó vvf.
                    //reactions allright, clean lens, good rbl.
Ther         Átfecskendezés mko sikerült.
                    //Successful squishing at both side
V\Refr 'C        Olvasó szemüveg javasolt: +2.0 Dsph mko.
                    //Reading glasses are suggested: +2.0 Dsph both side
Vél 'C        Éjszakánként m)könnygél ha szükséges.
                    //Artificial tears can be used at night if necessary
Kontr         Kontroll: panasz esetén
                    //Contorl: in case of further complaints


###SPART:Diagnózis  //Diagnoses
      Diagnózis
        DIAGNÓZISOK megnevezése                          Kód    Dátum      Év   K V T
##DOCDATE##
        Látászavar, k.m.n.                               H5390  2010.10.19         3

###SPART:Beavatkozások  //Treatments
      Beavatkozások
         Kód      Megnevezés                                    Menny.   Pont
         11041    Vizsgálat                                          1    750
```

Figure 1: The processed state of a document.

curate segmentation for separating each statement and by the incorporation of contextual features to the categorization process, which are among our future plans.

5 Conclusion

A method for structuring Hungarian ophthalmology notes has been described. The original form of these records created at clinical settings contains a large amount of noise and lacks almost any structure. Thus, in order to be able to use these documents either as the input of information retrieval algorithms or as a searchable database for medical experts, their intended structure had to be restored by assigning medical headings to each statement. This categorization was achieved by our method in three steps, relying on (1) the formatting clues of the original documents, (2) domain-specific keywords derived from the ophthalmology notes and (3) a statistical classification approach. Compared to a manually created gold standard, the results showed relatively high accuracy.

References

Michael F. Chiang, Sarah Read-Brown, Daniel C. Tu, Dongseok Choi, David S. Sanders, Thomas S. Hwang, Steven Bailey, Daniel J. Karr, Elizabeth Cottle, John C. Morrison, David J. Wilson, and Thomas R. Yackel. 2013. Evaluation of electronic health record implementation in ophthalmology at an academic medical center (an american ophthalmological society thesis). *Trans Am Ophthalmol Soc*, 111:70–92, Sep.

Amanda Elliott, Arthur Davidson, Flora Lum, Michael Chiang, Jinan B. Saaddine, Xinzhi Zhang, John E. Crews, and Chiu-Fang Chou. 2012. Use of electronic health records and administrative data for public health surveillance of eye health and vision-related conditions. *Am J Ophthalmol*, 154(6 0):S63–S70, Dec.

C. Friedman, S.B. Johnson, B Forman, and J Starren. 1995. Architectural requirements for a multipurpose natural language processor in the clinical environment. *Proc Annu Symp Comput Appl Med Care*, pages 347–51.

S.M. Meystre, G.K. Savova, K.C. Kipper-Schuler, and JF Hurdle. 2008. Extracting information from textual documents in the electronic health record: a review of recent research. *Yearb Med Inform*, 35:128–44.

György Orosz, Attila Novák, and Gábor Prószéky, 2013. *Hybrid text segmentation for Hungarian clinical records*, volume 8265 of *Lecture Notes in Artificial Intelligence*. Springer-Verlag, Heidelberg.

György Orosz, Attila Novák, and Gábor Prószéky. 2014. Lessons learned from tagging clinical Hungarian. *International Journal of Computational Linguistics and Applications*, 5(1):159–176.

Travis K. Redd, Sarah Read-Brown, Dongseok Choi, Thomas R. Yackel, Daniel C. Tu, and Michael F. Chiang. 2014. Electronic health record impact on productivity and efficiency in an academic pediatric ophthalmology practice. *Journal of AAPOS*, 18(6):584–589.

Naomi Sager, Margaret Lyman, Christine Bucknall, Ngo Nhan, and Leo J. Tick. 1994. Natural language processing and the representation of clinical data. *Journal of the American Medical Informatics Association*, 1(2), Mar/Apr.

Borbála Siklósi and Attila Novák. 2014. A magyar beteg. *X. Magyar Számítógépes Nyelvészeti Konferencia*, pages 188–198.

Borbála Siklósi, Attila Novák, and Gábor Prószéky. 2014. Context-aware correction of spelling errors in Hungarian medical documents. *Computer Speech and Language*, in press(0):–.

Borbála Siklósi. 2015. Clustering relevant terms and identifying types of statements in clinical records. In Alexander Gelbukh, editor, *Computational Linguistics and Intelligent Text Processing*, volume 9042 of *Lecture Notes in Computer Science*, pages 619–630. Springer International Publishing.