METHODS FOR PROCESSING NOISY TEXTS AND THEIR APPLICATION TO HUNGARIAN CLINICAL NOTES

Theses of the PhD Dissertation

Borbála Siklósi

Roska Tamás Doctoral School of Sciences and Technology

Pármány Péter Catholic University, Faculty of Information Technology and Bionics

Academic advisor:
Dr Gábor Prószéky

2015

Methods for processing noisy texts and their application to Hungarian clinical notes

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 research I was trying to transform raw clinical records written in the Hungarian medical language into a normalized set of documents to provide proper input to such higher-level processing methods.

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 experts. However, locally relevant characteristics of applied medical protocols or information relevant to locally prevailing phenomena 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. Clinical records are documents created at clinical settings with the purpose of documenting every-day clinical cases or treatments. The quality of this type of text stays far behind that of biomedical texts, which are also the object of several studies. Biomedical texts, mainly written in English, are the ones that are published in scientific journals, books, proceedings, etc. These are written in the standard language, in accordance with orthographic rules (Sager et al. (1994); Meystre et al. (2008)). On the contrary, clinical records are created as unstructured texts without using any proofing tools, resulting in texts full of spelling errors and nonstandard use of word forms in a language that is usually a mixture of the local language (Hungarian in our case) and Latin (Siklósi et al., 2012, 2013). These texts are also characterized by a high ratio of abbreviated forms, most of them used in an arbitrary manner. Moreover, in many cases, full statements are written in a special notational language (Barrows et al., 2000) that is often used in clinical settings, consisting only, or mostly of abbreviated forms.

Another characteristic of clinical records is that the target readers are usually the doctors themselves, thus using their own unique language and notational habits is not perceived to cause any loss in the information to be stored and retrieved. However, beyond the primary aim of recording patient history, these documents contain much more information which, if extracted, could be useful for other fields of medicine as well. In order to access this implicit knowledge, an efficient representation of the facts and statements recorded in the texts should be created.

This noisy and domain-specific character of clinical texts makes it much more challenging to process than general and even biomedical texts. **The goal of my Thesis research was to create preprocessing methods designed explicitly to Hungarian clinical records, preparing them to reach a normalized representation suitable for deeper processing and information extraction.** The performance of any text processing algorithm depends on the quality of the input text created by humans (doctors). Relevant and correct information is only extractable if it is present in the input. My goal was to reconstruct the

intended information in these clinical documents from their noisy and malformed state to provide them to higher level processing units.

There are two main approaches in processing clinical documents (Meystre et al., 2008). Methods falling into the first category apply rule-based algorithms. These are usually small, domain-specific applications that are expensive and time-consuming to build. The second group includes statistical algorithms. Though such methods are more and more popular, the main drawback is the need of large datasets, which are usually hard to obtain due to ethical issues. Moreover, supervised methods also require high quality annotations needed to be created manually by domain experts (in our case having both linguistic and medical expertise). Moreover, applications used for processing domain-specific texts are usually supported by some hand-made lexical resources, such as ontologies or vocabularies. In the case of less-resourced languages, there are very few such datasets and their construction needs quite an amount of human work.

In order to be able to support the adaptation of existing tools, and the building of structured resources, I examined a corpus of Hungarian ophthalmology notes. In this research, statistical methods are applied to the corpus in order to capture as much information as possible based on the raw data. Even though the results of each module are not robust representations of the underlying information, these groups of semi-structured data can be used in the real construction process.

2

New scientific results

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 will be available for researchers and medical 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. In the case of less-resourced languages, such as Hungarian, the lack of structured resources, like UMLS, Snomed, etc. makes it very hard to produce results comparable to those achieved by solutions for major languages. One way to overcome this problem could be the translation of these resources, however, doing it manually would require a huge amount of work, and automated methods that could support the translation effort are also of low quality for these languages.

Moreover, the quality of the documents created in the clinical settings is much worse, than that of general texts. Thus, the goal of this research was to transform raw clinical documents to a normalized representation that is appropriate for further processing. The methods applied are based on statistical algorithms, exploiting the information found within the corpus itself even at such preprocessing steps.

2.1 Representational schema

Wide-spread practice for representing structure of texts is to use XML to describe each part of the document. In my 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.

The resulting structure defines the separable parts of each record; however there are still several types of data within these structural units. 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. I filtered out these expressions to get a set of records containing only natural text. To solve this issue, unsupervised clustering algorithms were applied.

Digging deeper into the textual contents of the documents, a more detailed representation of these text fragments was necessary. That is why I stored each word in each sentence in an

individual data tag, augmented with several information. Such information are the *original form* of the word, the *corrected form*, its *lemma* and *part-of-speech* tag, and some phrase level information such as different types of *named entities*.

At this point, the textual content segments, each intended to appear under various subheadings, still remained as a mixture under a *content* tag. The original sections under these subheadings (*header*, *diagnoses*, *applied treatments*, *status*, *operation*, *symptoms*, etc.) contain different types of statements requiring different methods of higher-level processing. Moreover, the underlying information had to be handled in different ways, unique to each subheading. Thus, I implemented a method for categorizing lines of statements into their intended subheading. This was performed in two steps. First, formatting clues were recognized and labelled. These labelled lines were used as the training set for the second step, in which unlabelled lines were categorized by finding the most similar tag collection based on the tf-idf weighted cosine similarity measure.

THESIS 1:

I defined a flexible representational schema for Hungarian clinical records and developed an algorithm that is able to transform raw documents to the defined structure.

Related publications: 1, 4, 10, 16, 17

2.2 Automatic spelling correction

In Hungarian hospitals, 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 was a prerequisite of any further linguistic processing.

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.

I prepared a method for considering textual context when recognizing and correcting spelling errors. My system applies methods of Statistical Machine Translation (SMT), based on a word-based system for generating correction candidates. First a context-unaware word-based approach was created for generating correction suggestions, then I integrated this into an SMT framework. My system is able to correct certain errors with high accuracy, and, due to its parametrization, it can be tuned to the actual task. Thus, the presented method is able

to correct single errors in words automatically, making a firm base for creating a normalized version of the clinical records corpus in order to apply higher-level processing.

2.2.1 The word-based correction suggestion system

First, a word-based system was implemented that generates correction candidates for single words based on several simple word lists, some frequency lists and a linear scoring system.

At the beginning of the correction process, word forms that are contained in a list of stopwords and abbreviations are identified. For these words, no suggestions are generated. For the rest of the words, the correction suggestion algorithm is applied. For each word, a list of suggestion candidates are generated that contains word forms within one edit distance from the original form. The possible suggestions generated by a wide-coverage Hungarian morphological analyzer (Prószéky and Kis, 1999; Novák, 2003) are also added to this list.

In the second phase, these candidates are ranked using a scoring method based on (1) the weighted linear combination of scores assigned by several different frequency lists, (2) the weight coming from a confusion matrix of single-edit-distance corrections, (3) the features of the original word form, and (4) the judgement of the morphological analyzer. The system is parametrized to assign much weight to frequency data coming from the domain-specific corpus, which ensures not coercing medical terminology into word forms frequent in general out-of-domain text. Thus a ranked list of correction candidates is generated to all words in the text (except for the abbreviations and stopwords). However, only those are considered to be relevant, where the score of the first ranked suggestion is higher than that of the original word. This system was able to recognize most spelling errors and the list of the 5 highest ranked automatically generated corrections contained the actually correct one in 99.12% of the corrections in the test set.

2.2.2 Application of statistical machine translation to error corrections

Since my goal was to create fully automatic correction, rather than offering the user a set of corrections, the system should be able to automatically find the most appropriate correction. In order to achieve this goal, the ranking of the word-based system based on morphology and word frequency data proved to be insufficient. To improve the accuracy of the system, lexical context also needed to be considered.

To satisfy these two requirements, I applied Moses (Koehn et al., 2007), a widely used statistical machine translation (SMT) toolkit. During "translation", the original erroneous text is considered as the source language, while the target is its corrected, normalized version. In this case, the input of the system is the erroneous sentence: $E = e_1, e_2 \dots e_k$, and the corresponding correct sentence $C = c_1, c_2 \dots c_k$ is the expected output. Applying the noisy-channel model terminology to my spelling correction system: the original message is the correct sentence and the noisy signal received at the end of the channel data is the

corresponding sentence containing spelling errors. The output of the system trying to decode the noisy signal is the sentence $\hat{C}$, where

$$\hat{C} = argmax P(C|E) = argmax \frac{P(E|C)P(C)}{P(E)} \tag{2.1}$$

conditional probability takes its maximal value. Since $P(E)$ is constant, the denominator can be ignored, thus the product in the numerator can be derived from the statistical translation and language models.

These models in a traditional SMT task are built from a parallel corpus of the source and target languages based on the probabilities of phrases corresponding to each other. However, in my case there was no such a parallel set of documents. Thus, the creation of the translation models was substituted by three methods: (1) the word-based correction candidate generation system, (2) transformation of the distribution of various forms of abbreviations, and (3) inserting a table containing joining errors. These phrase tables are generated online, for each sentence that is to be corrected. The language model responsible for checking how well each candidate generated by the translation models fits the actual context is built using the SRILM toolkit (Stolcke et al., 2011). I have shown that the context-aware system outperformed the word-based one regarding both error detection and error correction accuracy.

THESIS 2:

I created an advanced method to automatically correct single spelling errors with high accuracy in Hungarian clinical records written in a special variant of domain-specific language containing expressions of foreign origin and a lot of abbreviations. I showed that applying a statistical machine translation framework as a spelling correction system with a language model responsible for context information is appropriate for the task and can achieve high accuracy.

Related publications: 1, 2, 6, 16, 17

2.3 Detecting and resolving abbreviations

Abbreviations occurring in clinical documents are usually ambiguous regarding not only their meaning, but the variety of the different forms they can take in the texts (for example *o.sin./o sin/o.s./os/OS*, etc.). Moreover, the ambiguity is further increased by the several resolution candidates a single abbreviated token might have (e.g. *o./f./p.*). Thus, after detecting abbreviations with the help of rules described by regular expressions, I investigated these shortened forms in the lexical context they appear in. When defining detection rules, I had to consider the non-standard usage of abbreviations, which is a very frequent phenomenon in clinical texts. The word-final period is usually missing, capitalization is used in an ad-hoc manner, compound expressions are abbreviated in several ways.

When performing the resolution of the detected abbreviations, I considered series of shortened forms (i.e. series of neighbouring tokens without any full word breaking the sequence) as single abbreviations. In such units, the number of possible resolutions of single, ambiguous tokens is reduced significantly. My goal was to find an optimal partitioning and resolution of these series in one step, i.e. having a resolved form corresponding to as much tokens as possible, while having as few partitions as possible.

Thus, in this research, a corpus-based approach was applied for the resolution of abbreviations with using the very few lexical resources available in Hungarian. Even though the first approach was based on the corpus itself, it did not provide acceptable results, thus the construction of a domain-specific lexicon was unavoidable. But, instead of trying to create huge resources covering the whole field of medical expressions, I have shown that a small, domain-specific lexicon is satisfactory, and the abbreviations to be included can be derived from the corpus itself.

Having this lexicon and the abbreviated tokens detected, the resolution was based on series of abbrevitaions. Moreover, in order to save mixed phrases (when only some parts of a multiword phrase is abbreviated) and to keep the information relevant for the resolution of multiword abbreviations, the context of a certain length was attached to the detected series. Beside completing such mixed phrases, the context also plays a role in the process of disambiguation. The meaning (i.e. the resolution) of abbreviations of the same surface form might vary in different contexts.

These abbreviation series were then matched against the corpus, looking for resolution candidates, and only unresolved fragments were completed based on searching in the lexicon. I have shown that having the corpus as the primary source is though insufficient, but provides more adequate resolutions in the actual domain, resulting in a performance of 96.5% f-measure in the case of abbreviation detection and 80.88% f-measure when resolving abbreviations of any length, while 88.05% for abbreviation series of more than one token.

THESIS 3:

I prepared an algorithm that is able to detect and resolve abbreviations in Hungarian clinical documents without relying on robust lexical resources and hand-made rules, rather applying statistical observations based on the clinical corpus.

THESIS 3.a:

I have shown that ambiguous abbreviations are much easier to be interpreted as members of abbreviation series, moreover, adding a one token long context to these series has also beneficial effect on the performance of the disambiguation process.

THESIS 3.b:

I have shown that the presence of a domain-specific lexicon is crucial, however it does not need to be a large, detailed knowledgebase. A small lexicon can be created by defining the resolution for the most frequent abbreviations found in a corpus of a narrow domain.

Related publications: 1, 7, 12, 13, 14

2.4 Semi-structured representation of clinical documents

Clinical documents represent a sublanguage regarding both the content and the language used to record them. However, one of the main characteristics of these texts is the high ratio of noise due to misspellings, abbreviations and incomplete syntactic structures. It has been shown that for a less-resourced language, such as Hungarian, there is a lack of lexical resources, which are used in similar studies to identify relevant concepts and relations for other languages. Thus, such lexicons should be built manually by human experts. However, an initial preprocessed transformation of the raw documents makes the task more efficient. Due to the availability of efficient implementations, statistical methods can be applied to a wide variety of text processing tasks. That is why, I have shown that corpus-based approaches (augmented with some linguistic restrictions) perform well on multiword term extraction and distributional similarity measures. Applying such methods can result in a semi-structured representation of clinical documents, appropriate for further human analyses.

2.4.1 Extracting multiword terms

In the clinical language (or in any other domain-specific or technical language), there are certain multiword terms that express a single concept. These are important to be recognized, because a disease, a treatment, a part of the body, or other relevant information can be in such a form. Moreover, these terms in the clinical reports could not be covered by a standard lexicon. This indicates the need to use some method for collocation identification. I used a modified version of the c-value algorithm (Frantzi et al., 2000). First, I defined a linguistic filter that is applied in order to ensure that the resulting list of terms contains only well-formed phrases. Phrases of the following forms were allowed:

$$(Noun|Adjective|PresentParticiple|Past(passive)Participle)^{+}Noun$$

This pattern ensures that only noun phrases are extracted and excludes fragments of frequent cooccurrences. After collecting all n-grams matching this pattern, the corresponding C-value is calculated for each of them, which is an indicator of the termhood of a phrase. The C-value is based on four components:

- the frequency of the candidate phrase;
- the frequency of the candidate phrase as a subphrase of a longer one;
- the number of these longer phrases;
- and the length of the candidate phrase.

These statistics are derived from the whole corpus of clinical notes, thus the resulting list of multiword terms are well suited in the domain and reflect their usage in the Hungarian medical language.

THESIS 4:

I have shown that corpus-based approaches augmented with some linguistic restrictions perform well on multiword term extraction from Hungarian clinical documents, resulting in a list of domain-specific terminology phrases ranked according to their termhood.

Related publications: 1, 4, 5, 11

2.4.2 Distributional behaviour of the clinical corpus

Creating groups of relevant terms in the corpus requires a similarity metric measuring the closeness of two terms. Instead of using an ontology for retrieving similarity relations between words, I applied the unsupervised method of distributional semantics. Thus, the similarity of terms is based on the way they are used in the specific corpus.

The context of a word is represented as a set of features, each feature consisting of a relation (r) and the related word (w'). In other studies these relations are usually grammatical relations, however in the case of clinical texts, the grammatical analysis performs poorly, resulting in a rather noisy model. In Carroll et al. (2012), Carroll et al. suggest using only the occurrences of surface word forms within a small window around the target word as features. In my research, a mixture of these ideas was used by applying relations based on the word forms and part-of-speech tags.

Each feature was associated with a frequency determined from the corpus. From these frequencies the amount of information contained in a tuple of (w,r,w') was computed by using maximum likelihood estimation. This is equal to the mutual information between w and w'. Then, to determine the similarity between two words (w_1 and w_2) the similarity measure described in Lin (1998) was used, i.e.:

$$SIM(w_1, w_2) = \frac{\sum_{(r,w) \in T(w_1) \bigcap T(w_2)} (I(w_1, r, w) + I(w_2, r, w))}{\sum_{(r,w) \in T(w_1)} I(w_1, r, w) + \sum_{(r,w) \in T(w_2)} I(w_2, r, w)}$$

where T(w) is the set of pairs (r,w') such that I(w,r,w') is positive.

Having this metric, the pairwise distributional similarity of any two terms can be counted. In my research, however, I only dealt with nouns that appear at least twice in the corpus and multiword expressions.

The results showed that the resulting similarity relations are valid between terms, which made the application of this measure of semantic relatedness appropriate for creating conceptual clusters as the base of an ontology for the clinical domain. In order to create such a hierarchy, I applied an agglomerative clustering algorithm on the most frequent terms and nouns, where each term was represented by a feature vector containing its similarity to all the other terms. The clustering and ordering of terms extracted from clinical documents can be used directly as an initial point of a Hungarian medical ontology containing phrases used by practitioners

in their daily cases. Moreover, since each group (and each node in the hierarchy) was given a unique identifier, these can also be replaced into the original texts. Thus, a higher-level abstract representation of the documents were created, transforming each document to a set of normalized patterns.

THESIS 5:

I have applied the methods of distributional semantics to create a similarity measure between multiword terms and nouns and used it to create a hierarchical representation of the most relevant concepts in Hungarian clinical documents.

THESIS 5.a:

I created a method for automatically constructing a system of concepts that can be used as an aid in the creation of hand-made domain-specific resources and is flexible to be parametrized regarding its granularity.

THESIS 5.b:

I have shown that the resulting system of concepts is appropriate for creating an abstract representation of raw documents by mapping various occurrences of a cluster member to an identifier, thus resulting in documents containing a normalized set of patterns.

Related publications: 1, 4, 5, 11, 14

3

List of Papers

Journal publications

1 **Borbála Siklósi**, Attila Novák, Gábor Prószéky (2016): Context-aware correction of spelling errors in Hungarian medical documents, *Computer Speech & Language*, Vol.35, pp. 219-233, ISSN 0885-2308.

2 **Borbála Siklósi**, Attila Novák, György Orosz, Gábor Prószéky (2014): Processing noisy texts in Hungarian: a showcase from the clinical domain, *Jedlik Laboratories Reports*, Vol. II, no.3, pp. 5-62, ISSN 2064-3942

3 László János Laki, Attila Novák, **Borbála Siklósi**, György Orosz (2013): Syntax-based reordering in phrase-based English-Hungarian statistical machine translation. *International Journal of Computational Linguistics and Applications*, Vol. 4 no. 2. pp. 63–78, ISSN 0976-0962.

Book chapters

4 **Borbála Siklósi** (2015): Clustering Relevant Terms and Identifying Types of Statements in Clinical Records, In: A. Gelbukh (Ed.), *Lecture Notes in Computer Science Volume 9042: Computational Linguistics and Intelligent Text Processing* Springer International Publishing, Berlin Heidelberg. Part II pp. 619–630. ISBN 978-3-319-18116-5.

5 **Borbála Siklósi**, Attila Novák (2014): Identifying and Clustering Relevant Terms in Clinical Records Using Unsupervised Methods. In: Besacier, L.; Dediu, A.-H. and Martín-Vide, C. (Eds.), *Lecture Notes in Computer Science Volume 8791: Statistical Language and Speech Processing* Springer International Publishing, Berlin Heidelberg. pp. 233-243 ISBN 978-3-319-11396-8.

6 **Borbála Siklósi**, Attila Novák, Gábor Prószéky (2013): Context-Aware Correction of Spelling Errors in Hungarian Medical Documents. In: Adrian-Horia Dediu, Carlos Martín-Vide, Ruslan Mitkov, Bianca Truthe (Eds.), *Lecture Notes in Computer Science Volume 7978: Statistical Language and Speech Processing* Springer, Berlin Heidelberg. pp. 248–259 ISBN 978-3-642-39592-5.

7 **Borbála Siklósi**, Attila Novák (2013): Detection and Expansion of Abbreviations in Hungarian Clinical Notes. In: F. Castro, A. Gelbukh, M.G. Mendoza (Eds.): *Lecture Notes in Computer Science, Vol. 8265: Advances in Artificial Intelligence and Its Applications.* Springer, Berlin Heidelberg. pp. 318–328. ISBN 978-3-642-45114-0

8 György Orosz, László János Laki, Attila Novák, **Borbála Siklósi** (2013): Improved Hungarian Morphological Disambiguation with Tagger Combination. In: Habernal, Ivan; Matousek, Vaclav (Eds.) *Lecture Notes in Computer Science, Vol. 8082: Text, Speech, and Dialogue* Pilsen, Czech Republic. Springer, Berlin Heidelberg. pp. 280–287. ISBN: 978-3-642-40584-6.

Conference proceedings

9 Novák Attila, **Siklósi Borbála** (2015): Automatic Diacritics Restoration for Hungarian. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, Lisbon, Portugal: Association for Computational Linguistics. pp. 2286–91.

10 **Siklósi Borbála**, Novák Attila (2015): Restoring the intended structure of Hungarian ophthalmology documents. BioNLP Workshop at the 53^{rd} Annual Meeting of the Association for Computational Linguistics, ACL 2015. Beijing, China, July 26-31, 2015

11 **Siklósi Borbála**, Novák Attila (2015): Nem felügyelt módszerek alkalmazása releváns kifejezések azonosítására és csoportosítására klinikai dokumentumokban. In: Tanács Attila, Varga Viktor, Vincze Veronika (szerk.) *XI. Magyar Számítógépes Nyelvészeti Konferencia.* Szegedi Tudományegyetem, Informatikai Tanszékcsoport, Szeged. pp. 237-248

12 **Borbála Siklósi**, Attila Novák, Gábor Prószéky (2014): Resolving Abbreviations in Clinical Texts Without Pre-existing Structured Resources. In: *Proceedings of the Fourth Workshop on Building and Evaluating Resources for Health and Biomedical Text Processing (BioTxtM 2014).* Reykjavík

13 **Siklósi Borbála**, Novák Attila (2014): Rec. et exp. aut. Abbr. mnyelv. KLIN. szöv-ben – rövidítések automatikus felismerése és feloldása magyar nyelvű klinikai szövegekben. In: Tanács Attila, Varga Viktor, Vincze Veronika (szerk.) *X. Magyar Számítógépes Nyelvészeti Konferencia.* Szegedi Tudományegyetem, Informatikai Tanszékcsoport, Szeged. pp. 167–176. ISBN 978-963-306-246-3

14 **Siklósi Borbála**, Novák Attila (2014): A magyar beteg. In: Tanács Attila, Varga Viktor, Vincze Veronika (szerk.) *X. Magyar Számítógépes Nyelvészeti Konferencia.* Szegedi Tudományegyetem, Informatikai Tanszékcsoport, Szeged. pp. 188–198. ISBN 978-963-306-246-3

15 **Siklósi Borbála**, Novák Attila, Prószéky Gábor (2013): Helyesírási hibák automatikus javítása orvosi szövegekben a szövegkörnyezet figyelembevételével. In: Tanács Attila; Vincze Veronika (szerk.) *A IX. Magyar Számítógépes Nyelvészeti Konferencia előadásai.* SZTE, Szeged. pp. 148–158 ISBN 978-963-306-189-3

16 **Borbála Siklósi**, György Orosz, Attila Novák, Gábor Prószéky (2012): Automatic structuring and correction suggestion system for Hungarian clinical records. In: *LREC-2012: SALTMIL-AfLaT Workshop on "Language technology for normalisation of less-resourced languages".* Istanbul, Turkey, 2012. pp. 29–34

17 **Siklósi Borbála**, Orosz György, Novák Attila (2011): Magyar nyelvű klinikai dokumentumok előfeldolgozása. In: Tanács Attila, Varga Viktor, Vincze Veronika (szerk.) *VIII. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2011).* Szegedi Tudományegyetem, pp. 143–340

18 Laki László, Novák Attila, **Siklósi Borbála** (2013): Hunglish mondattan – átrendezésalapú angol-magyar statisztikai gépifordító-rendszer. In: Tanács Attila; Vincze Veronika (szerk.) *A IX. Magyar Számítógépes Nyelvészeti Konferencia előadásai.* SZTE, Szeged. pp. 71–82 ISBN 978-963-306-189-3

19 György Orosz, László János Laki, Attila Novák, **Borbála Siklósi** (2013): Combining Language-Independent Part-of-Speech Tagging Tools. In: J. P. Leal, R. Rocha, and A. Simoes (Eds.), *2nd Symposium on Languages, Applications and Technologies.* Porto: Schloss Dagstuhl–Leibniz-Zentrum für Informatik. pp. 249–257 ISBN 978-3-939897-52-1

20 László János Laki, Attila Novák, **Borbála Siklósi** (2013): English-to-Hungarian Morpheme-based Statistical Machine Translation System with Reordering Rules. In: Marta R. Costa-jussa, Reinhard Rapp, Patrik Lambert, Kurt Eberle, Rafael E. Banchs, Bogdan Babych (Eds.) *Proceedings of the Second Workshop on Hybrid Approaches to Machine Translation (HyTra).* Association for Computational Linguistics. pp. 42–50

Bibliography

Barrows, J. R., Busuioc, M., and Friedman, C. (2000). Limited parsing of notational text visit notes: ad-hoc vs. NLP approaches. *Proceedings of the AMIA Annual Symposium*, pages 51–55.

Carroll, J., Koeling, R., and Puri, S. (2012). Lexical acquisition for clinical text mining using distributional similarity. In Gelbukh, A., editor, *Computational Linguistics and Intelligent Text Processing*, volume 7182 of *Lecture Notes in Computer Science*, pages 232–246. Springer Berlin Heidelberg, Berlin, Heidelberg.

Frantzi, K., Ananiadou, S., and Mima, H. (2000). Automatic recognition of multi-word terms:. the C-value/NC-value method. *International Journal on Digital Libraries*, 3(2):115–130.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., and Herbst, E. (2007). Moses: Open Source Toolkit for Statistical Machine Translation. In *Proceedings of the ACL 2007 Demo and Poster Sessions*, pages 177–180, Prague. Association for Computational Linguistics.

Lin, D. (1998). Automatic retrieval and clustering of similar words. In *Proceedings of the 17th international conference on Computational linguistics - Volume 2*, COLING '98, pages 768–774, Stroudsburg, PA, USA. Association for Computational Linguistics.

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

Novák, A. (2003). Milyen a jó Humor? [What is good Humor like?]. In *I. Magyar Számítógépes Nyelvészeti Konferencia*, pages 138–144, Szeged. SZTE.

Prószéky, G. and Kis, B. (1999). A unification-based approach to morpho-syntactic parsing of agglutinative and other (highly) inflectional languages. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, ACL '99, pages 261–268, College Park, Maryland. Association for Computational Linguistics.

Sager, N., Lyman, M., Bucknall, C., Nhan, N., and Tick, L. J. (1994). Natural Language Processing and the Representation of Clinical Data. *Journal of the American Medical Informatics Association*, 1(2).

Siklósi, B., Novák, A., and Prószéky, G. (2013). Context-Aware Correction of Spelling Errors in Hungarian Medical Documents. In Dediu, A.-H., Martín-Vide, C., Mitkov, R., and Truthe, B., editors, *Statistical Language and Speech Processing*, volume 7978 of *Lecture Notes in Computer Science*, pages 248–259. Springer Berlin Heidelberg.

Siklósi, B., Orosz, G., Novák, A., and Prószéky, G. (2012). Automatic structuring and correction suggestion system for Hungarian clinical records. In De Pauw, G., De Schryver, G.-M., Forcada, M., M. Tyers, F., and Waiganjo Wagacha, P., editors, *8th SaLTMiL Workshop on Creation and use of basic lexical resources for less-resourced languages*, pages 29–34, Istanbul, Turkey.

Stolcke, A., Zheng, J., Wang, W., and Abrash, V. (2011). SRILM at Sixteen: Update and Outlook. In *Proc. IEEE Automatic Speech Recognition and Understanding Workshop*, Waikoloa, Hawaii.