Resolving Abbreviations in Clinical Texts Without Pre-existing Structured Resources

Borbála Siklósi, Attila Novák, Gábor Prószéky
Pázmány Péter Catholic University, Faculty of Information Technology and Bionics,
MTA-PPKE Hungarian Language Technology Research Group
50/a Práter street, 1083 Budapest, Hungary
{surname.firstname}@itk.ppke.hu

Abstract

One of the most important topics in clinical text processing is the identification of relevant concepts. This includes the detection and resolution of abbreviations, acronyms, or other shortened forms in the documents. Even though the task of resolving abbreviations can be treated as a word sense disambiguation problem, such methods require structured lexical knowledge bases. However, for less-resourced languages such resources are not available. In this paper, a method is proposed for the disambiguation and resolution of abbreviations found in Hungarian clinical records. In order to achieve reasonable performance, a lexicon must be created for each domain. It is shown how the set of entries to be included in such a lexicon can be induced from the corpus, thus the manual effort of creating a lexicon can be reduced significantly. The results for resolving abbreviations in Hungarian clinical documents are also shown, which are achieved by using the corpus instead of non-existing structured resources.

Keywords: clinical NLP, abbreviation resolution, less-resourced languages

1. Introduction

Processing medical texts is an emerging topic in natural language processing. There are existing solutions mainly for English to extract knowledge from medical documents, which 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 (Lindberg et al., 1993), Snomed (International Health Terminology Standards Development Organisation, 2010), 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.

Beside the availability of structured resources, categories of medical text processing can also be differentiated according to the type of text being processed (Meystre et al., 2008). Most research focuses on biomedical texts that appear in books, articles, literature, etc. However, there are a growing number of studies on processing clinical texts written by doctors in the clinical settings. This paper fits this latter line of research.

In Hungarian hospitals, 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 Hungarian and Latin (Siklósi et al., 2012; Siklósi et al., 2013). These texts are also characterized by a high ratio of abbreviated forms. The use of some of these abbreviations follows some standard rules, but most of them are used in an arbitrary manner. Moreover, in most 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. Even for non-expert humans, it is a hard task to find phrase boundaries in a long sequence of shortened forms. Processing such documents is not an easy task, and resolving abbreviations is a prerequisite of further linguistic processing.

The task of abbreviation resolution is often treated as word sense disambiguation (WSD) (Navigli, 2012). The best-performing approaches of WSD use supervised machine learning techniques. In the case of less-resourced languages, however, neither manually annotated data, nor an inventory of possible senses of abbreviations are available, which are prerequisites of supervised algorithms (Nasiruddin, 2013). On the other hand, unsupervised WSD methods are composed of two phases: word sense induction (WSI) must precede the disambiguation process. Possible senses for words or abbreviations can be induced from a corpus based on contextual features. However, such methods require large corpora to work properly, especially if the ratio of ambiguous terms and abbreviations is as high as in the case of clinical texts. Due to confidentiality issues and quality problems, this approach is not promising either.

In this study, we introduce the behaviour of abbreviations in clinical documents of low-resourced languages, demonstrated with our Hungarian corpus of medical records. Then, a corpus-based approach is described for the resolution of abbreviations with using the very few lexical resources available in Hungarian. As this method did not provide acceptable results, the construction of a domain-specific lexicon was unavoidable. Instead of trying to create huge resources covering the whole field of medical expressions, it is shown that small domain-specific lexicons are satisfactory and the abbreviations to be included can be derived from the corpus itself. Finally, an analysis of the combination of these methods is presented.

2. Related work

There are several studies, most of them applied to English texts, that address the specific task of medical language processing. One of the most challenging preprocessing steps is the detection and resolution of abbreviations found in the free-text parts of these documents. As opposed to biomedical literature, where the first mention of an abbreviated form is usually preceded by its expanded form or definition, in clinical records this is not the case. That is why simple abbreviation-definition patterns are not applicable to clinical notes as described by Xu et al. in (Xu et al., 2007). The same study compares some machine learning approaches, all achieving considerable results, but even using already existing external resources, the authors admit the need of a manually created inventory. A recent study (Wu et al., 2012) compared the performance of some biomedical text processing systems trained on biomedical literature on the task of resolving abbreviations in clinical texts. All the systems (MetaMap, MedLEE and cTAKES) achieved suboptimal results calling for more advanced abbreviation recognition modules.

Most approaches to resolving clinical abbreviations carried out in English rely on some very common medical lexical resources. Even though Xu in (Xu et al., 2007) showed that the sense inventories generated from the UMLS covered only about 35% of the abbreviations they had extracted from their corpus, these already contain definitions and possible interpretation candidates. Thus the problem can be reduced to abbreviation disambiguation, as it is carried out in (Wu et al., 2012; Xu et al., 2009; Pakhomov et al., 2005). These methods focus primarily on supervised machine learning approaches, where a part of the training corpus is labeled manually. Pakhomov in (Pakhomov, 2002) described a semi-supervised method to build training data for Maximum Entropy modeling of abbreviations automatically. In most of these studies, both training and evaluation of the systems are performed on a few manually chosen abbreviations and their disambiguation.

3. Clinical abbreviations

The use of a kind of notational text is very common in clinical documents. This dense form of documentation contains a high ratio of standard or arbitrary abbreviations and symbols, some of which may be specific to a special domain or even to a doctor or administrator. These shortened forms might refer to clinically relevant concepts or to some common phrases that are very frequent in the specific domain. For the clinicians, the meaning of most of these common phrases is as trivial as the standard shortened forms of clinical concepts due to their expertise and familiarity with the context. Some examples for abbreviations falling into these categories are shown in Table 1.

3.1. Series of abbreviations

Even though standalone abbreviated tokens are highly ambiguous, they more frequently occur as members of multiword abbreviated phrases, in which they are usually easier to interpret unambiguously. For example *o.* could stand for any word either in Hungarian or in Latin, starting with the letter *o*, even if limited to the medical domain. However,

Domain	Abbr.	Resolution	in Hungarian	in English
standard	o. d. med. gr.	oculus dexter mediocris gradus	jobb szem közepes fokú	right eye medium grade
domain-specific	o. o.	oculus os	szem csont	eye bone
domain-specific common	sü fén n	saját szemüveg fényérzés nélkül normál	saját szemüveg fényérzés nélkül normál	own glasses no sense of light normal
common words	köv lsd	következő lásd	következő lásd	next see

Table 1: Some examples for the use of simple abbreviations. Some of them are commonly known standard forms, usually of Latin origin, some others, though related to the clinical domain, might have several meanings depending on the specific sub-domain. The rest are abbreviated common words, usually of Hungarian origin, and might also refer to both clinical phrases or common words.

in our corpus of anonymized ophthalmology reports, *o.* is barely used by itself, but together with a laterality indicator, i.e. in forms such as *o. s., o. d.*, or *o. u.* meaning *oculus sinister* 'left eye', *oculus dexter* 'right eye', or *oculi utriusque* 'both eyes', respectively. In such contexts, the meaning of the abbreviated *o.* is unambiguous. It should be noted, that these are not the only representations for these abbreviated phrases, for example *oculus sinister* is also abbreviated as *o. sin., os, OS*, etc. Table 2 shows the ratio of unique abbreviation sequences of different lengths detected automatically in the corpus with the method described in Section 5.1. The number of different single-token abbreviations is roughly equal to the number of all multi-token abbreviations.

Length:	1	2	3	4	5	>5
Number:	49.53%	26.34%	15.00%	5.95%	2.16%	0.98%

Table 2: The ratio of unique abbreviation series of different lengths detected automatically in the corpus.

Thus, when performing the resolution of abbreviations, we considered series of such shortened forms instead of single tokens. A series is defined as a continuous sequence of shortened forms without any unabbreviated word breaking the sequence. These series are not necessarily coherent phrases. The individual elements of such sequences of abbreviations are by themselves highly ambiguous, and even if there were an inventory of Hungarian medical abbreviations, which does not exist, their resolution could not be solved. Moreover, the mixed use of Hungarian and Latin phrases results in abbreviated forms of words in both languages, thus the detection of the language of the abbreviation is another problem. For example, in the "sentence"

Dg : Tu. pp. inf et orbitae l. dex. ,
Cataracta incip. o. utr. , Hypertonia,

the abbreviation spans are the following:

Dg,
Tu. pp. inf,
l. dex.,
incip. o. utr..

3.2. The lexical context of abbreviation sequences

In the above example, the last section is misleading, since the token *incip.* is part of the phrase *Cataracta incip.*, i.e. it is related to its preceding neighbour, which is not included in this list as part of an abbreviation. This mixed use of a phrase is very common in the documents, with a diverse variation in using certain words in their full form or in some shortened form instead. In order to save such phrases and to keep the information relevant for the resolution of multiword abbreviations, the context of a certain length is attached to the detected series. In our experiments, the length of the context taken from both the left and right sides of the abbreviations ranged from 0 to 3 tokens. Since the average length of sentences in the corpus is 9.7 (Orosz et al., 2013), considering a larger context could span across sentences, but that would make no sense.

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. Our experimental results showed that this does not require a larger window of sampling either.

4. Resources

The corpus In our research, we used a corpus of anonymized clinical documents, all falling into the domain of ophthalmology. A portion of this corpus was set aside for testing purposes. Table 3 contains detailed information about the size of these subcorpora.

	documents	sentences	tokens	abbreviated tokens
whole corpus	2008	60660	552594	113091
test corpus	22	693	5599	765

Table 3: The size of our corpus of clinical records

External lexicon Even though there are no structured lexical resources for Hungarian, the official coding system for diseases, anatomical structures and medical procedures is available (similar to the ICD systems in English). Thus, a simple dictionary was built from the ophthalmology sections of these descriptions. The final list of phrases contained 3329 entries. However, these phrases are written in the language of the official terminology, which is different in several respects from that used in the clinical texts.

Handmade lexicon Since the official descriptions turned out not to be of much use, a domain-specific lexicon seemed to be necessary. The first step of designing such a resource is to decide what phrases to include. We assumed that the most frequent abbreviations occurring in the corpus without their expanded form ever being written out have one unambiguous resolution within a narrow domain. For example, in the domain of ophthalmology, the abbreviation *o. d.* always stands for the phrase *oculus dexter*, meaning 'right eye'. Even though it appears in various shortened forms, it is never spelled out in its full form. Thus, a frequency list of abbreviations and abbreviation series was created from the whole corpus. Then a threshold value was defined experimentally and the abbreviations with a relative frequency above this threshold were included into our set of domain-specific abbreviations. Finally, the resolution of these abbreviations were defined with the help of a medical expert.

This approach can be applied to other domains as well. Thus, our method of abbreviation resolution is applicable to new domains with a relatively small amount of manual effort. In our case of ophthalmology records, rather good coverage can be achieved even with a small lexicon of 44 entries. Adding more items to the list further improves the quality of the resolution, however, the improvement achieved by adding new items diminishes quickly.

5. Methods

The primary objective of the research described here is finding spans in a sequence of abbreviations that can be unambiguously resolved together. Since sometimes whole statements or even sentences are written using this kind of heavily abbreviated notation, it is important to find an optimal partitioning of the tokens into meaningful spans. In the previous example, the fragment *incip. o. utr.* should be divided into the spans of *incip.* and *o. utr.*, even if the abbreviation *incip.* is not relevant by itself, but still its meaning is not related to the rest of the abbreviation sequence. However *o. utr.* can be resolved with high confidence.

5.1. Detection of abbreviations

The first problem to solve when trying to handle abbreviations within running text is detecting them. Since these texts usually do not follow standard orthographic and punctuation rules, especially in the case of highly abbreviated notational text, the detection of abbreviations cannot be based on patterns formulated according to standard rules of forming abbreviations. The ending periods are usually missing, abbreviations are written with varying case (capitalization) and in varying length. For example the following forms represent the same expression, *vörös visszfény* 'red reflection': *vvf, vvfény, vörösvfény*. We applied some heuristic rules to derive relevant features as indicators of a token being an abbreviation. These features were based on the following characteristics: the presence or absence of a word-final period, the length of the token, the ratio of vowels and consonants within the token, the ratio of upper- and lowercase letters, and the judgment of a Hungarian morphological analyzer, the lexicon of which was expanded with medical terminology (Novák, 2003; Prószéky and Kis, 1999).

5.2. Resolving abbreviations

Once the abbreviation series are extracted from a document, a maximum coverage resolution suggestion process is carried out. The steps of the algorithm are the following:

1. For each possible partitioning of the tokens into non-overlapping spans, a regular expression pattern is generated. The patterns are created by general abbreviation rules, such that each letter in the abbreviated form represents the starting letter of each word in the expanded phrase (assuming that it is an acronym). Or,

in the case of multiword abbreviations, each member represents the beginning of each word (not just the first letter) in the interpretation. Some pattern generation rules are presented in Table 4.

2. The regular expressions for each span are matched against the corpus resulting in full or partial resolutions.

3. The regular expressions are refreshed based on the results retrieved from the corpus.

4. These expressions are then matched against the lexicons.

5. The results of the spans are concatenated to cover the whole series and each such merged resolution candidates are given a score appropriate for ranking.

6. The highest ranked resolution is considered as the final one.

During this process, the optimal division of the abbreviation series and its resolution are carried out in one step. This is ensured by the scoring method. The different partitionings are ranked according to three features, optimizing for the longest coverage and best resolution of the abbreviation sequence. The features used for ranking are 1) the number of all tokens in the sequence covered by a resolved form, 2) the size of the longest span covered, 3) the size of the shortest span covered. If the sequence *Exstirp. tu. et reconstr. pp. inf. l. d.*, is partitioned as | *Exstirp. tu.* | *et* | *reconstr. pp. inf.* | *l. d.* |, with having a resolution candidate for each partition except for the word *et*, then the value of the features are 7, 3 and 2, respectively. Finally, these values given for each feature are transformed into a percentage score by a weighted combination of them.

5.3. Experiments

The algorithm described above uses three resources where the resolution candidates are searched for: 1) a lexicon containing the ophthalmology section of the official ICD coding system, 2) our handmade lexicon, and 3) the corpus itself. In our experiments, we investigated how the performance of the algorithm is influenced by the availability of these resources to the program. The goal of these experiments were threefold. First, since there is a lack of structured external resources, we wanted to investigate to what extent we could rely on a raw, domain-specific corpus to resolve abbreviations. In order to do this, the size of the corpus in which the regular expressions were matched was changed incrementally. Second, we wanted to check the hypothesis that a small, manually built lexicon (containing the most frequent abbreviations) can be built and utilized and in an effective manner. Moreover, we wanted to identify a threshold for the collected entry candidates to such a lexicon for an arbitrary domain. Third, the best performing combination was evaluated.

6. Results and discussion

A test set of 22 documents was used for evaluation purposes both for the task of abbreviation detection and resolution.

abbr	regexp	matching expansion
o. s.	`o[^ ]* s[^ ]*`	oculus sinister
os	`os[^ ]*`	osteoporosis
os	`o[^ ]* s[^ ]*`	oculus sinister

Table 4: Some of the simplest patterns generated from two short abbreviated phrases. The complexity and variability of these patterns is proportional to the length of the original abbreviation sequence.

	result
Precision:	95.99%
Recall:	97.12%
F-measure:	96.55%

Table 5: Evaluation results for abbreviation detection

The abbreviations in this set were labeled manually and resolved by a medical expert. Finally, the number of tokens labeled as abbreviations was 765. The actual meaning for 56 of them could not be specified. These included initials of doctors; author-specific shortened forms; tokenization errors, etc. Thus, the remaining 709 abbreviations were considered when evaluating our methods for abbreviation resolution.

Performance in both tasks was measured in terms of precision, recall and F-measure. For abbreviation detection, precision was calculated as the number of true positives divided by the sum of true positives and false positives and recall was defined as the number of true positives divided by the sum of true positives and false negatives. On the other hand, for the resolution task, precision was defined as the number of correctly resolved tokens divided by the number of all resolved tokens, while recall is the number of correctly resolved tokens divided by the number of all abbreviations (Cohn, 2003). Thus, if having one abbreviation resolved correctly without touching anything else, a precision of 100% could be achieved, which does not reflect the real performance. That is why F_2-measure was defined as the harmonic mean between precision and recall biased towards recall.

Table 5 shows the final results for the detection of abbreviations. Most of the errors in the detection arose from misspelled forms, tokenization errors or Latin abbreviations.

In the case of abbreviation resolution, the effect of varying four parameters were investigated. The first parameter was the size of the context taken into consideration when resolving abbreviation sequences. Second, we investigated the effect of changing the size of the corpus used for pattern matching. Third, the effect of changing the size of our handmade lexicon and finding the optimal threshold value to decide what to include in it. And fourth, the performance of the system using the best combination of these parameters was evaluated.

Figure 1 shows the results of three experimental setups ((a),(b) and (c)). In each of them, the size of the manually created lexicon was kept at a fixed size (0, 44, and 136 entries). The size of the corpus was increased in units of

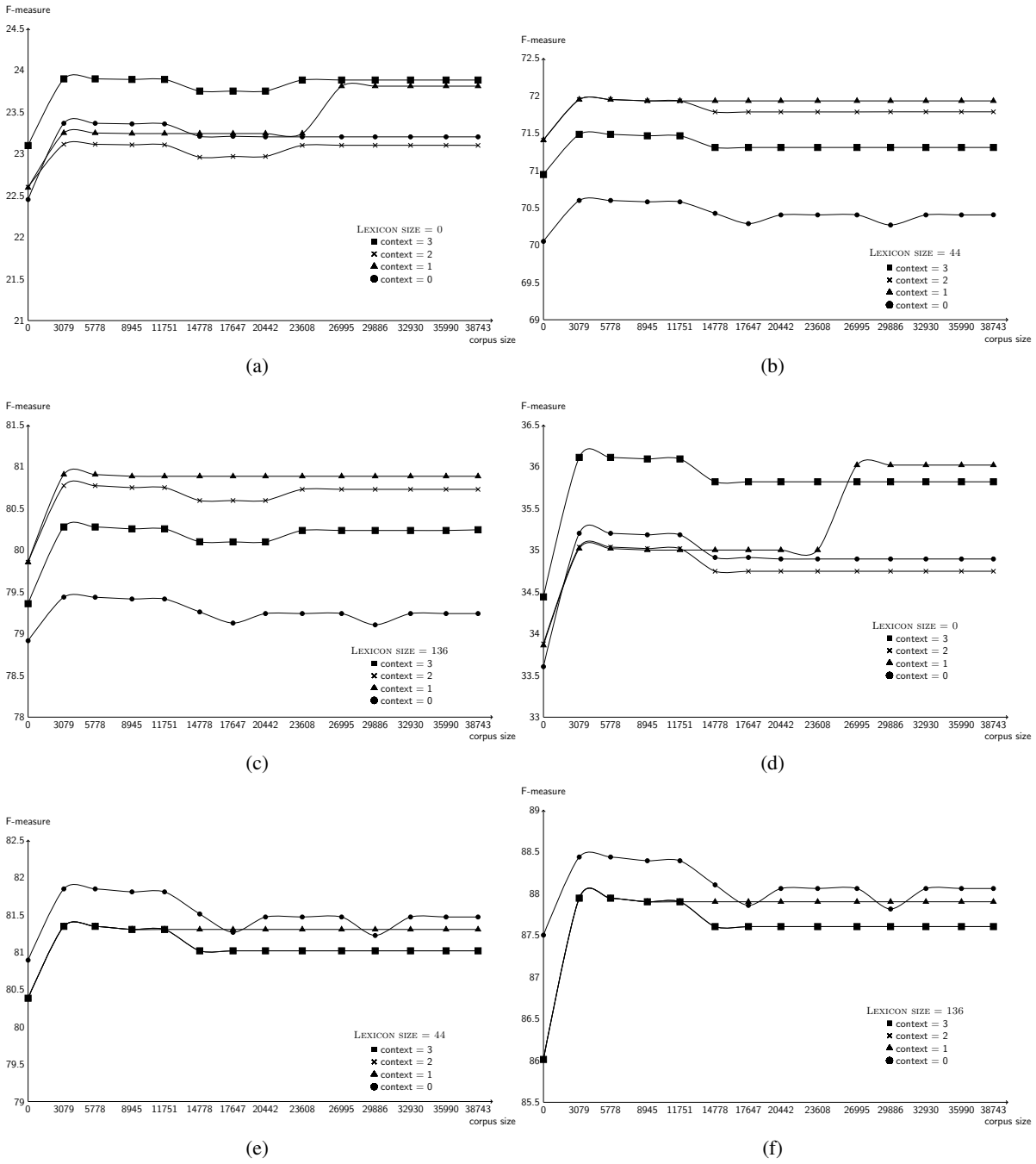

Figure 1: The performance results as a function of the corpus size for different context sizes and using a fixed portion of our handmade lexicon (0, 44 and 136 entries respectively). Graphs (a), (b) and (c) represent the results for all abbreviation series, while graphs (d), (e) and (f) represent the results for multi-token abbreviation sequences only.

around 3000 sentences in each step and the performance for context sizes 0 to 3 tokens were measured. Replacing each abbreviation with its definition from the lexicon (if it was included in the lexicon) was considered as the baseline. In the first case, without the lexicon, this baseline was an F-measure of 0%, in the second case 60.52%, and in the third case, it was 75.71%. As it can be seen from the graphs, these values are quite below the performance of our final combined system. In each case, considering the tokens without any context always performs worst, however, taking a context larger than one token before and after the abbreviation has a positive effect only if the manually created lexicon is not used. In this case, the system with a context size of three tokens performed best.

Increasing the size of the corpus had a similar effect. When the domain-specific lexicon is available, then the only significant change in performance occurred when adding the first portion of the corpus. Further increasing its size did not influence the performance of these setups. This is due to the relatively small size of the corpus and the noisy nature of the texts.

In (Siklósi and Novák, 2013), it was reported that by in-

creasing the size of the corpus, the difference between the performance achieved by using lexicons of different sizes can be made up. However, in that study, abbreviations of multiple tokens were considered (i.e. no single abbreviated tokens) and the test set contained unique abbreviations. Thus we also performed the evaluation tests on abbreviation sequences of length greater than 1 (see graphs (d), (e) and (f) of Figure 1). Comparing the behaviour of the algorithm for all and for multiple-token abbreviations, there are two main differences. First, the performance values are higher for longer series of abbreviations. Second, taking a context of any size performs worse than having only the abbreviation by itself in the case of such longer series. Moreover, we found that, when using the lexicon, adding too much of the corpus will generate noise for the resolution process instead of enhancing the quality.

In order to find an optimal relative occurrence frequency threshold value for abbreviations that should be included in the handmade lexicon, the largest corpus size was used and the abbreviations were added to the lexicon incrementally. Figure 2 shows the change in the threshold and the performance as a function of the number of entries in the lexicon. The results are in accordance with our assumption that including the most frequent abbreviations in the lexicon has a more significant role than creating a large, more detailed lexicon. Adding only the first 10 most frequent abbreviations to the lexicon results in a 30% increase in the performance. Even though the performance grows further if adding more entries, an optimal threshold can be set by cutting the long tail of the graphs. Thus, in our case, this cut was done where abbreviations with relative frequency higher than 0.0025 were added to the lexicon. It resulted in a lexicon size of 44 entries.

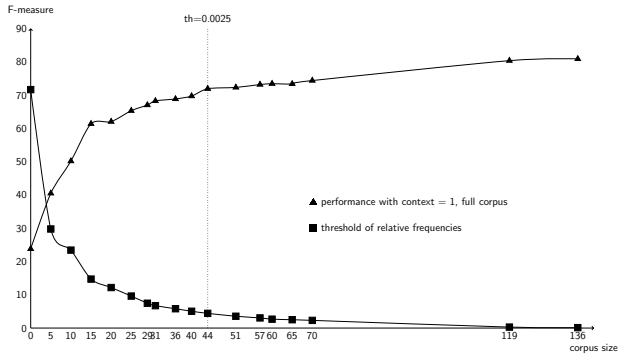

Figure 2: The change in the threshold and the performance as a function of the number of entries in the lexicon. Decreasing the threshold (measured in relative corpus frequency) below the value of 0.0025 does not produce a significant increase in the performance relative to the manual effort needed to define the meaning of these abbreviations. The F-measure values here correspond to a context size of one token and the whole corpus is used for pattern matching.

Even though the above investigations are important in order to be able to generalize the system to other domains or languages, and filling the gap caused by the lack of structured resources, our goal was also to achieve the best results in resolving abbreviations in Hungarian clinical records in the domain of ophthalmology. The best results compared to those achieved by using the above described threshold are shown in Table 6. In the case of the final setup, pattern matching was applied to the whole corpus with taking a one-token context around each abbreviation, and using an enlarged version of our lexicon to 136 entries. (We have no data about further increasing this size.) Thus, an F-measure of 80.88% was achieved for all abbreviations and 88.05% for abbreviation series consisting of multiple tokens.

length	lexicon size	precision	recall	f2
all	136	93.23%	78.29%	80.88%
	44	88.37%	68.73%	71.92%
>1	136	96.22%	86.23%	88.05%
	44	90.63%	79.46%	81.46%

Table 6: The best performance achieved for the ophthalmology corpus for all abbreviations and for abbreviation series of length greater than 1.

7. Conclusion

Automatic detection and resolution of abbreviations in clinical documents are usually solved by using external resources. In this study an approach was presented to solve the same problems if such resources are not available, which is the case for less-resourced languages, such as Hungarian. It has been 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. The rest of the abbreviations can be resolved based on the corpus itself. On the other hand, the role of the context of an abbreviated token was investigated from different aspects. It has been 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.

8. Acknowledgement

The authors of this paper would like to thank Ákos Kusnyerik for providing his expertise knowledge in ophthalmology.

This research was partially supported by the project grants TÁMOP–4.2.1./B–11/2-KMR-2011-0002 and TÁMOP–4.2.2./B–10/1-2010-0014.

9. References

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.

Cohn, T. (2003). Performance metrics for word sense disambiguation. In *Proceedings of the Australasian Language Technology Workshop (ALTW)*, pages 49–56, Melbourne, Australia, December.

International Health Terminology Standards Development Organisation. (2010). Snomedct: Systematized nomenclature of medicine-clinical terms.

Lindberg, D., Humphreys, B., and McCray, A. (1993). The unified medical language system. *Methods of Information in Medicine*, 32(4):281–291.

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.

Nasiruddin, M. (2013). A state of the art of word sense induction: A way towards word sense disambiguation for under-resourced languages. *CoRR*, abs/1310.1425.

Navigli, R. (2012). A quick tour of word sense disambiguation, induction and related approaches. In *Proceedings of the 38th Conference on Current Trends in Theory and Practice of Computer Science (SOFSEM)*, pages 115–129.

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

Orosz, G., Novák, A., and Prószéky, G., (2013). *Hybrid text segmentation for Hungarian clinical records*, volume 8265 of *Lecture Notes in Artificial Intelligence*. Springer-Verlag, Heidelberg.

Pakhomov, S., Pedersen, T., and Chute, C. (2005). Abbreviation and acronym disambiguation in clinical discourse. In Friedman, C., Ash, J., and Tarczy-Hornoch, P., editors, *Proceedings of the AMIA Annual Symposium*, pages 589–593, Washington DC, USA. Bethesda, MD: AMIA Press.

Pakhomov, S. (2002). Semi-supervised maximum entropy based approach to acronym and abbreviation normalization in medical texts. In Isabelle, P., editor, *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics (ACL)*, pages 160–167, Philadelphia, USA. Rochester, NY: ACL Press.

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, Stroudsburg, PA, USA. Association for Computational Linguistics.

Siklósi, B. and Novák, A., (2013). *Detection and Expansion of Abbreviations in Hungarian Clinical Notes*, volume 8265 of *Lecture Notes in Artificial Intelligence*, pages 318–328. Springer-Verlag, 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.

Siklósi, B., Novák, A., and Prószéky, G. (2013). Context-aware correction of spelling errors in Hungarian medical documents. In Dediu, A.-H., Martin-Vide, C., Mitkov, R., and Truthe, B., editors, *Statistical Language and Speech Processing*, volume LNAI 7978. Springer Verlag.

Wu, Y., Denny, J. C., Rosenbloom, S. T., Miller, R. A., Giuse, D. A., and Xu, H. (2012). A comparative study of current clinical natural language processing systems on handling abbreviations in discharge summaries. *Proceedings of the AMIA Annual Symposium*, 2012:997–1003.

Xu, H., Stetson, P., and Friedman, C. (2007). A study of abbreviations in clinical notes. In Teich, J., Suermondt, J., and Hripcsak, G., editors, *Proceedings of the AMIA Annual Symposium*, pages 821–825, Washington DC, USA. Bethesda, MD: AMIA Press.

Xu, H., Stetson, P., and Friedman, C. (2009). Methods for building sense inventories of abbreviations in clinical notes. *Journal of American Medical Informatics Association*, 16(1):103–108.