Journal of Universal Computer Science, vol. 14, no. 22 (2008), 3781-3795
submitted: 20/9/08, accepted: 18/11/08, appeared: 28/12/08 © J.UCS

Semantic Information in Medical Information Systems: Utilization of Text Mining Techniques to Analyze Medical Diagnoses

Andreas Holzinger
(Research Unit HCI4MED, Institute of Medical Informatics, Statistics & Documentation
Medical University Graz, Austria
andreas.holzinger@meduni-graz.at)

Regina Geierhofer
(Research Unit HCI4MED, Institute of Medical Informatics, Statistics & Documentation
Medical University Graz, Austria
regina.geierhofer@meduni-graz.at)

Felix Mödritscher
(Institute for Information Systems and New Media
Vienna University of Economics and Business Administration, Austria
felix.moedritscher@wu-wien.ac.at)

Roland Tatzl
(IT and IT Marketing, Campus 02
Graz University of Applied Sciences, Austria
roland.tatzl@campus02.at)

Abstract: Most information in Hospitals is still only available in text format and the amount of this data is immensely increasing. Consequently, text mining is an essential area of medical informatics. With the aid of statistic and linguistic procedures, text mining software attempts to dig out (mine) information from plain text. The aim is to transform data into information. However, for the efficient support of end users, facets of computer science alone are insufficient; the next step consists of making the information both usable and useful. Consequently, aspects of cognitive psychology must be taken into account in order to enable the transformation of information into knowledge of the end users. In this paper we describe the design and development of an application for analyzing expert comments on magnetic resonance images (MRI) diagnoses by applying a text mining method in order to scan them for regional correlations. Consequently, we propose a calculation of significant co-occurrences of diseases and defined regions of the human body, in order to identify possible risks for health.

Keywords: Information Retrieval, Text Mining, Performance, Medical Documentation
Categories: H.3.1, H.3.3, I.2.7, I.7, J.3

1 Introduction and Motivation

Significant progress has been made in the last years in the application of text mining techniques, in order to cope with the rapidly increasing information overload in the area of medical literature [Sullivan et al., 1999], [Hall and Walton, 2004]. However,

developments for text mining techniques in the area of clinical information systems and medical documentation are rare [Noone et al., 1998], [Holzinger et al., 2007b]. The broad application of sophisticated medical information systems amasses large amounts of medical documents, which must be reviewed, observed and analyzed by human experts [Holzinger et al., 2007a]. All essential documents of the patient records contain at least a certain portion of data which has been entered in *free-text* fields and has long been in the focus of research [Gell et al., 1976], [Gell, 1983], [Zingmond and Lenert, 1993].

Although text can be *created* simple by the end-users, the support of automatic analysis is extremely difficult [Gregory et al., 1995], [Holzinger et al., 2000], [Lovis et al., 2000]. It is likely that some interesting and relevant relationships remain completely undiscovered, due to the fact that relevant data are scattered and no investigator has linked them together manually, an example from medical literature can be found in [Smalheiser and Swanson, 1998]. With regard to the fact that textual information is definitely an extremely important part of medical documents and that the amount of textual information is rather increasing then decreasing in future, the Institute for Medical Informatics, Statistics and Documentation (IMI) of the Medical University Graz (www.meduni-graz.at/imi) has been carrying out a variety of projects to analyze and process medical documents by application of computer-based techniques and to present this information in an end-user centered manner.

2 Theoretical Background

2.1 Text Mining: Definition

Text mining, sometimes called *textual data mining* is generally defined as a knowledge-intensive process in which end users interact with a collection of textual information by using analytic tools; typical text mining tasks include categorization, clustering, concept extraction, production of taxonomies, sentiment analysis, document summarization and entity-relation modeling [Feldman and Sanger, 2007]. The major challenge of biomedical text mining over the next years is to make such systems useful to biomedical researchers. Certainly, this will require enhanced *access* to full text, better *understanding* of the feature space of biomedical text, enhanced methods for *measuring the usefulness* of systems to end users and continued *cooperation* with the biomedical research community to ensure that their needs and requirements are appropriately addressed [Cohen and Hersh, 2005].

Contrary to structured information, textual information is characterized by its inherently unstructured and fuzzy nature, being language and domain dependent as well as consisting of sentences and sub-sentences [Sistrom and Honeyman-Buck, 2005]. Text represents factual information in a complex, rich, and opaque manner – which makes it difficult to be analyzed by standard statistical data mining methods: relying on human analysis results in either huge workloads or the analysis of only a tiny fraction of the database [Nasukawa and Nagano, 2001].

Consequently, *text mining* stands for a multidisciplinary field involving *"information retrieval, text analysis, information extraction, clustering, categorization, visualization, database technology, machine learning and data mining"* [Tan, 1999].

Techniques in the field of text mining aim at the extraction of coherences within such unstructured information. Basically, text mining approaches apply statistical or pattern-based algorithms [Biemann et al., 2004], in order to extract significant key-word associations or to mine for prototypical documents (e.g. for part-of-speech tagging or term extraction) [Rajman and Besancon, 1998].

2.2 Text Mining: Knowledge Discovery from Data (KDD)

Text mining is considered a sub-specialty of Knowledge Discovery from Data (KDD) and has been headed strongly in the direction of Natural Language Processing (NLP) in the last decade [Liddy, 2000]. According to [Granitzer, 2006], the following phases of the text mining process can be identified:

(1) Pre-processing: As textual data is primarily encoded within documents of different formats (XML, HTML, Word, etc.), lexical analysis methods can be applied to prepare the texts, e.g. by lexical analysis and removal of layout information. Further, this phase also allows the improvement of the text quality by utilizing methods for feature analysis, such as stemming or part-of-speech tagging.

(2) Information extraction: This stage aims at transforming the unstructured text entities into structured elements, e.g. through named entity recognition, co-reference resolution, template element filling, scenario templates and others. Therefore, techniques of machine learning or linguistic analyses are applied, whereby the quality of the outcome is highly dependent on the domain and the correctness of the phrasing.

(3) Feature generation and statistical analyses: By applying statistical methods, features of an information space can be extracted from the structured elements of a text. Hereby, methods such as frequency analyses, collocations or co-occurrences (of words) are utilized. [Zipf, 1949] examined the connection between frequency and significance of words within a text corpora, which led to *Zipf's law* as a widely used basis for analyzing word frequencies in texts. Zipf's law states that the frequency of words in a large corpus is inversely proportional to their rank [Le Quan et al., 2002]. Moreover, also word pairs (co-occurrences) and word groups (collocations) are of interest for generating features from a text.

(4) Operations on feature spaces: As a result, each text object is described with several features, which, for instance, spans an n-dimensional vector space with n equals to the number of features for a text object. This *feature space* can be utilized for further operations, e.g. comparing texts, visualizing relations in the text corpus, calculating similarities or rankings, clustering, etc.

2.3 Text Mining in Medical Documentation

According to [Hotho et al., 2005], text mining approaches can be useful *for many* different application scenarios, such as patent search, text classification for news agencies, anti-spam filtering of emails, bioinformatics, etc.

Text mining in medical documentation is an extremely important area [Buenaga et al., 2006], however, it encompasses a lot of various problems [Gell, 1983]. In this

paper, we report about our experiences on utilizing text mining techniques for medical documentation.

3 Related Work

In the last years, mining in medical digital libraries has come up with new findings and hypotheses. [Srinivasan, 2004] reports on the development of text mining methods on the basis of medical subject headings (MeSH). These algorithms generate hypotheses by identifying potentially interesting terms related to the specific input. Additionally, new protein associations have been found by clustering learning and vector classification techniques [Fu et al., 2003]. Another interesting approach utilizes a rule-based parser and co-occurrence for extracting and combining relations [Leroy et al., 2003]. Further, a new way to use thalidomide has been discovered by mapping phrases to concepts of the Unified Medical Language System [Weeber et al., 2001a]. Finally, co-occurrences are useful to build up gene networks [Jenssen et al., 2001] and to discover gene interactions [Stephens et al., 2001].

Derived from these experiences, three kinds of application areas for text mining can be outlined in the field of medical documentation:

- Firstly, such methods are applied in order to build up an infrastructure or models for biomedicine, i.e. by finding patterns or relations in texts and generating a feature space. Inspecting the projects on the basis of the well-known text mining framework GATE (http://gate.ac.uk/projects.html), MultiFlora or myGRID can be identified as examples for this approach.
- Secondly, text mining is used to observe and retrieve documents with innovative ideas in the scope of a restricted domain (cf. projects such as BioRAT or InESBi).
- Thirdly, text mining techniques are utilized to extract information or features out of a medical text corpus, which is generated as product of clinical documentation for further operations such as information retrieval. The MedDictate software comprises one solution in this scope. Although this last category of applications areas overlaps partially with the first one, there are only a few reports about the mining of medical diagnoses. In addition to these experiences, we want to report on our approach towards aiming at the detection of diseases in MRT diagnoses. In the following two sections, this project and its outcome are described in detail.

4 Methods and Materials

The success of text mining methods in medical research was the origin of our idea to apply statistical techniques in order to find hints for possible locations of diseases in MRT diagnoses. Therefore, we aimed at the topological proximity between anatomic structures and pathologic expressions and implemented a tool which calculates the significant co-occurrences of anatomic and pathologic terms within the diagnoses.

4.1 Text Corpus and Pre-Processing

Accompanying measures during the evaluation of an information extraction tool revealed the need for additional assistance in finding topological relations between

anatomic structures including regional indicators and diseases such as tumors. Given these requirements, there was a demand for domain specific databases for anatomic structures and pathologic expressions.

At the beginning of this project, we used a text corpus of about 6.000 diagnoses, which comprises comments of medical experts on magnetic resonance imaging (MRI) material. Table 1 shows the distribution of diagnoses per year.

Year	Number of diagnoses
1987	1456
1988	2143
1989	1714
1990	52
2003	331
2004	313

Table 1: Diagnoses per year

These findings derived from diverse radiologists and are completely written in capital letters and packed with medical terms. Hence, we were confronted with three serious problems: synonyms, medical dialects and abbreviations [Weeber et al., 2001b].
Further, the diagnoses are spread over a period of 17 years. As a consequence, time-dependent changes of terminology might be possible. These textual diagnoses were made anonymous and imported into a database. A distinction between non-important free text and biomedical entities as proposed by [Chen et al., 2005], p.20, does not seem to be necessary because of the pure provenience of the diagnostic corpus.

4.2 Information extraction

Considering the systemic performance and the sentence-based statistical calculation, the texts were also split up into sentences. Pre-processing of the free text is realized in a way, that the occurrences of expression pairs are counted and stored. Sentences have been identified by break-iterators of the Jakarta Group. In addition a length smaller then 15 characters has not been accepted as a complete sentence. Following this methodology, we identified 15731 sentences. Consequently, the calculation of the co-occurrences can now be executed by means of one of the three formulas which are shown in figure 1.

Performance issues demanded a full-text indexing of the diagnoses as well as a reduction of the anatomic terms. Unfortunately, two or three character words are found in many other words, consequently they must be excluded from the reference database.

Additionally, we were in need of anatomic and pathologic terms for our approach. A corpus of approximately 6.800 anatomic structures was generated from an anatomic dictionary [Dauber, 2005]. This dictionary (in German) offers a rough allocation of anatomic structures to anatomic regions. More precision in finding such structures can be reached by using synonyms, however, the gathering of which is extremely time consuming. Efforts have been made to start with a synonym enhancement for the anatomic data at the IMI, which has been used for the calculation. On the other hand,

a Pathology database has been set up manually due to a lack of accessible resources. These corpuses represent the domain specific database sources for the statistic calculation and can be maintained in special application modules.

4.3 Feature extraction

Our methodology for the identification of proximity is based on the calculation of significant co-occurrence. This term describes the cumulative joint appearance of pairs of words in a defined environment which, in our case, is determined to be a sentence.

4.3.1 Basic Algorithm and Methodology

The calculation of significant co-occurrences is based on the *Poisson distribution.* In accordance with [Biemann et al., 2004], [Heyer et al., 2006], the original formula can be simplified for two different ranges of the input parameters (see also figure 1).

Hereby, *a* stands for the number of sentences containing term *A*, *b* for the number of sentences containing term *B*, *n* for the number of all sentences and *k* for the number of sentences containing both terms.

Be $\lambda = \frac{a \cdot b}{n}$ then: $sig(A,B) = \frac{-\log\left(1 - e^{-\lambda}\sum_{i=0}^{k-1}\frac{1}{i!}\cdot\lambda^{i}\right)}{\log n}$

If $\frac{(k+1)}{\lambda} > 2.5$ then: $sig(A,B) \approx \frac{\lambda - k\cdot\log\lambda + \log k!}{\log n}$

If $k > 10$ then: $sig(A,B) \approx \frac{k\cdot(\log k - \log\lambda - 1)}{\log n}$

Figure 1: The three formulas to calculate significant co-occurrences of two terms

Due to performance reasons, we decided to implement the calculation of the significant co-occurrences independently, instead of re-using existing text mining modules. The different ways to calculate the co-occurrence allow the usage of the fastest algorithm for each diagnose, as the simplified formulas (the last two in the figure) require less time and processing power.

Listing 1 shows a JAVA code example of the calculation class which handles all of the required steps for the statistical analysis.

```
1 /**
2 * fact calculates a mathematical function in the formula.
3 */
4 static double fact(double n) {
5 if (n <= 0)
6 return 0;
7 if (n == 1)
```

```
8 return 1;
9 else
10 return (n * fact(n - 1));
11 }
12
13 /**
14 * sigsum sums up a function in the formula.
15 */
16 static double sigsum(double k, double lambda) {
17 double result =0;
18 System.out.println("k: "+k);
19 System.out.println("lambda: "+lambda);
20 for(double i=1; i<=(k-1);i++){
23 result += Math.pow(lambda,i)/fact(i);
24 }
25 return result;
26 }
27
28 /**
29 * sig calculates the significance for the given parameters.
30 */
31 static double sig(double a, double b, double k, double n){
32 double lambda = (double)(a*b)/n;
33 double decision = (k+1)/lambda;
34 double sig =0;
35 System.out.println("func lambda: "+lambda);
36 System.out.println("func decision: "+decision);
37 System.out.println("sigsum: "+sigsum(k,lambda));
38 if (decision > (double)2.5){
39 if (k > (double)10){
40 System.out.println("k>10");
41 sig = k*(Math.log(k)-Math.log(lambda)-1)/Math.log(n);
42 } else{
43 System.out.println("k<10");
44 sig = (lambda-k*Math.log(lambda)+Math.log(fact(k)))/Math.log(n);
45 }
46 } else{
47 System.out.println("dec<2.5");
48 if (k < 200){
49 sig = -Math.log(1-Math.pow(Math.E,-lambda)*sigsum(k,lambda))/
Math.log((double)n);
50 } else{
51 System.out.println("Math.log(k): "+Math.log(k));
52 System.out.println("Math.log(lambda): "+Math.log(lambda));
53 //System.out.println("Math.log(n): "+Math.log(n));
54 sig = k*(Math.log(k)-Math.log(lambda)-1)/Math.log(n);
55 }
56 //sig = k*(Math.log(k)-Math.log(lambda)-1)/Math.log(n);
57 }
58 return sig;
59 }
```

Listing 1: Code example from class DBMANAGER.JAVA

4.3.2 Performance issues

Of course we had to consider pre-processing steps of the MRI diagnoses in order to complete calculations on the initial text corpus within a reasonable period of time. These steps include a pre-selection of anatomic or pathologic expression which occur in the diagnostic corpus to speed up calculation later on.

4.4 The Web Application

A basic end user access control system is used for logging purposes. Maintenance operations must be executed as administrator. End user actions include registration, editing of the registration information, login, logout and observing the login history. The application itself offers modules for the following functionality:

- Diagnoses can be listed and filtered according to two terms;
- Anatomic terms can also be filtered;
- The location for each anatomic term is indicated;
- The synonyms module shows all available synonyms for each anatomic term. These synonyms cover small parts of terms, but will increase in future;
- The menu option "ADD PATHOLOGY" provides a dialogue to add a term;
- Significant co-occurrents are listed in module COOCCURRENTS. Additionally new calculations can be started;
- A maintenance module provides splitting of the diagnoses as well as calculation of the occurrence of the single terms;

4.5 Core Functions and Advantages

For performance purposes, the administrator can initiate a pre-calculation of the occurrence of each single term. Thus the number of sentences and anatomic terms for the statistic calculation can be reduced. The splitting of the diagnoses is the second method of improving performance during the calculation. Additionally the split sentences are reduced by excluding sentences with a character length < 15. Thereby abbreviation sentences are most likely eliminated.

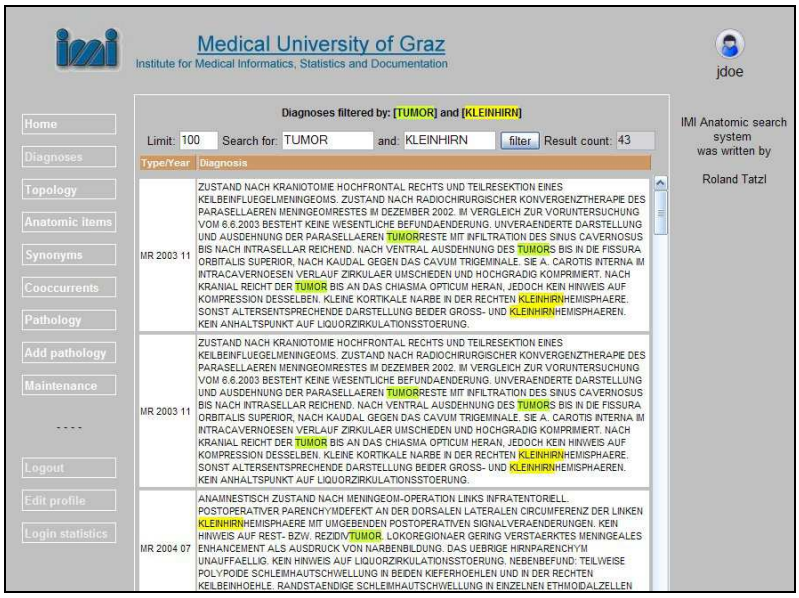

Figure 2: Filtering diagnoses according by the terms "TUMOR" and "KLEINHIRN"

Search for topological relations is based on MRI diagnoses in a heterogeneous context (e.g. cranial and spinal MRI diagnoses), as visualized in figure 2. The table of diagnoses can be filtered manually and retrieved in a list. Simple IR-techniques, such as query term highlighting, are used to visualize the results. These functions support medical experts on comparing the results for the calculated pairs of terms.

Anatomic and pathologic terms can be edited in separated dialogues. Because of a lack of a Pathology reference corpus, authoring functionality has been implemented for the application in order to manually add, modify or remove expressions. The processing of the medical free text is based on the formulas described in subsection 3.1. In order to improve the overall performance during calculation, the query considers only sentences which contain any of the anatomic or pathologic terms and only terms which were previously identified in the diagnose corpus.

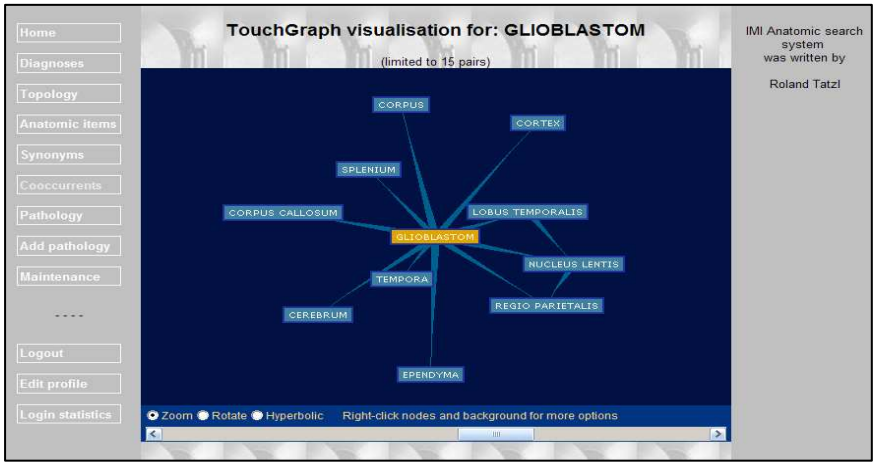

Figure 3: Tough graph visualization for the pathological term "GLIOBLASTOMA"

The resulting pair-list is shown in a table in descending order of significance for each expression. For further investigations, the most promising pairs can be used to filter the diagnoses in order to evaluate the results. For visualization purposes, a *Touchgraph* applet was implemented which enables the medical expert to estimate the proximity at a glance (see figure 3). In addition to the basic relations between the pathologic expression and the anatomic structures, the interconnection significance among the anatomic structures is calculated, to show their potential proximity within the corpus of diagnoses. This visualization clearly shows the topologic relationship.

5 Results and Experiences

The results of the calculation must be analyzed in their special context. Each pair of terms implicates specific anatomic-pathologic issues, and, in addition, the meaning of the co-occurrence must be interpreted individually.

5.1 Results for an Exemplary Scenario

The results are discussed on the basis of the malign brain cancer glioblastoma, which generally occurs in the group of middle-aged men and is located most likely in the corpus callosum and the temporal lobe. According to [Schlegel et al., 2003] and [Hopf et al., 1999] the percentage of glioblastoma in all intracranial tumors is 15 to 20%. In addition to the location information the age of the patients could easily be retrieved inside the clinical information system in order to check the occurrence of the tumor in male or female age history. The peak incidence of glioblastoma is reported to be in middle adult life (56 to 60 years) but no age group is known to be exempt. The incidence is higher in men (ratio of approximately 1.6:1) [Ropper & Brown, 05].

An analysis in a linguistic database (http://wortschatz.uni-leipzig.de) showed a very low occurrence in common language sources, such as newspaper articles and books, and emphasizes the importance of a well maintained domain specific database.

The result set for the calculation (refer again to figure 3) showed 10 pairs of terms for glioblastoma including synonyms. Finally, figure 4 shows the detailed result set for glioblastoma.

Pathology	Anatomic item	a	b	k	n	sig	date
GLIOBLASTOM	Tempora	45	328	14	15731	2.467	2007-12-08 23:01:39.0
GLIOBLASTOM	Lobus temporalis	45	63	3	15731	0.736	2007-12-08 23:01:39.0
GLIOBLASTOM	Splenium	45	34	2	15731	0.564	2007-12-08 23:01:39.0
GLIOBLASTOM	Regio parietalis	45	17	1	15731	0.318	2007-12-08 23:01:39.0
GLIOBLASTOM	Cerebrum	45	31	1	15731	0.26	2007-12-08 23:01:39.0
GLIOBLASTOM	Nucleus lentis	45	37	1	15731	0.243	2007-12-08 23:01:39.0
GLIOBLASTOM	Corpus callosum	45	41	1	15731	0.234	2007-12-08 23:01:39.0
GLIOBLASTOM	Corpus	45	54	1	15731	0.209	2007-12-08 23:01:39.0
GLIOBLASTOM	Ependyma	45	67	1	15731	0.191	2007-12-08 23:01:39.0
GLIOBLASTOM	Cortex	45	87	1	15731	0.17	2007-12-08 23:01:39.0

Figure 4: Calculation details for the Glioblastoma result set

We emphasize, that a highly significant co-occurrence does not prove the affliction of an anatomic structure with the paired disease. Also, a co-occurrence suggests a relation for further investigation. The most significant result suggests the co-occurrence with TEMPORA and SPLENIUM, whereby both locations are well known and located in the most significant decile of the result set. The distribution of the result set is shown in figure 5. The threshold for the upper decile is 0.564.

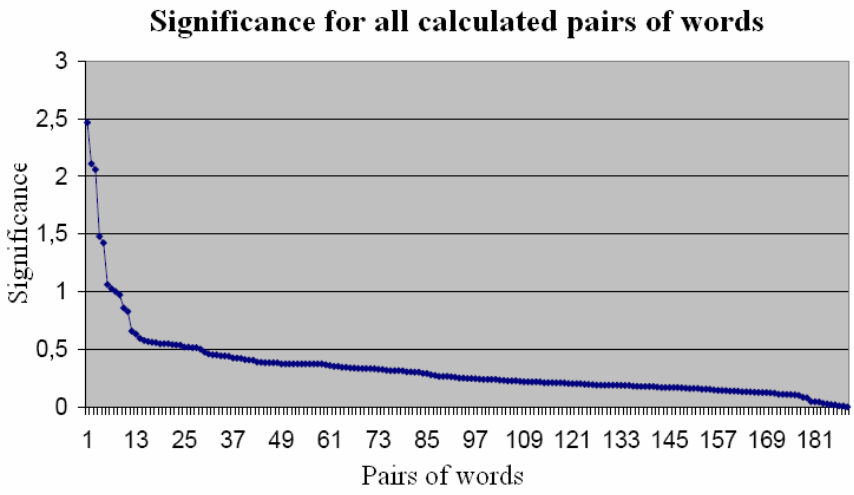

Figure 5: Statistical distribution of the result set

The second group of results does not make sense, because the associated anatomic terms CORPUS or EPENDYMA (thin epithelial membrane lining the ventricular system of the brain and the spinal cord canal) are too common and do not implicate worthwhile location information. The third group showed a combination of terms which are less well known and might be an interesting hint for further investigation.

5.2 Discussion of Experiences and Problematic Aspects

The resulting set of the calculations are listed in descending order according to the significance. The overall statistical evaluation says that the upper decile contains the most significant co-occurrences with a high probability. Some of the less significant results include interesting combinations of terms which may point out valuable new conclusions. Finally, syntax highlighting enables the observer to find the terms of the query quite easily in the result set of diagnoses. Despite such experiences, we also identified the following problematic aspects for applying statistical text mining techniques on diagnoses: (1) The group of results which are not useful is an evidence of the weakness of the anatomic reference corpus. Expressions with a too widespread meaning should not be considered in order to reduce the wrong results. (2) The amount of diagnoses must be increased. Examples from professional common language research show that about 5 million sentences or more are required to validate our approach. (3) The diagnoses are specific for responsible radiologists. Therefore, they are not thoroughly comparable, as different experts tend to use other terminologies. (4) Words with only few characters (like “OS” or “COR”) are not suitable for searching purposes. Thus, they have to be eliminated. (5) Acronyms and abbreviations (dotted) cause difficulties when splitting the diagnoses into sentences.

5.3 Remedies for the problems

As a result the future work will include a reduction of useless words in the anatomic reference corpus by implementing a relevance ranking. The increase of the amount of diagnoses is no problem at all except the protection of privacy. In order to identify the differences between radiologists additional calculation strategies must be introduced.

5.4 Opportunities

Nevertheless, we find that our methodology for analyzing medical diagnoses comprises a promising approach: The analyzing of medical text corpora is fully computer driven and fully automated. The overall calculations require from a few minutes up to some hours, depending of the computational hardware and the amount of data. Thus, this method can be applied on text corpora at any time, e.g. to use other, more accurate anatomic and pathologic expressions for old diagnoses. Secondly, our tool is of interest for clinical professionals in order to support them at their daily tasks, for instance during pre-analyzing diagnoses. We also implemented some functions to support general medical experts in their daily work in order to reduce their cognitive load (see subsection 3.3). Finally, this kind of text mining algorithm could be also valuable for other application areas.

6 Conclusion

We emphasize the importance of computer-based methods in medical documentation and the automatization of clinical processes, including analyzing diagnoses. In this context, we developed a methodology for text mining in medical text corpora and implemented a tool to evaluate our idea. The outcome of the calculations showed valuable results although based on a relatively low number of sentences. Observing all the diagnoses, generated in a hospital daily, will definitely improve the diagnostic value. However, we still have no proof that the anatomic structure is affected by the related disease, however, our experiences encouraged us to carry out further research efforts on these co-occurrences. Mining in large amounts of textual medical information can reveal new patterns for various questions. There will be a continued need for new mining assistants solving problems which are not even known today. We identified a huge benefit for the administration in identifying trends and developments in time to come to the appropriate decisions. The appropriate information presentation to the end users is a central future challenge, in order to keep their cognitive load in an optimum level, providing cognitive performance support.

References

[Biemann et al., 2004] Biemann, C., Bordag, S., Heyer, G., Quasthoff, U. and Wolff, C.: "Language-independent methods for compiling monolingual lexical data", Computational Linguistics and Intelligent Text Processing, Springer-Verlag Berlin, Berlin, 2004, 217-228.

[Buenaga et al., 2006] Buenaga, M., Maña, M., Gachet, D. and Mata, J.: "The SINAMED and ISIS Projects: Applying Text Mining Techniques to Improve Access to a Medical Digital Library", Research and Advanced Technology for Digital Libraries, 2006, 548-551.

[Chen et al., 2005] Chen, H., Fuller, S., Friedman, C. and Hersh, W.: "Medical Informatics: Knowledge Management And Data Mining in Biomedicine", Springer, Berlin, Heidelberg, New York, (2005).

[Cohen and Hersh, 2005] Cohen, A. M. and Hersh, W. R.: "A survey of current work in biomedical text mining", Briefings in Bioinformatics, 6, (2005), 57-71.

[Dauber, 2005] Dauber, W.: "Feneis' Bild-Lexikon der Anatomie, 9th Edition", Thieme, Stuttgart, (2005).

[Feldman and Sanger, 2007] Feldman, R. and Sanger, J.: "The Text Mining Handbook: Adavanced Approaches in Analyzing Unstructured Data ", Cambridge University Press, Cambridge (2007).

[Fu et al., 2003] Fu, Y., Mostafa, J. and Seki, K.: "Protein association discovery in biomedical literature", Proceedings of the 3rd ACM/IEEE-CS joint conference on Digital libraries, (2003), 113-115.

[Gell, 1983] Gell, G.: "AURA - routine documentation of medical text", Methods of Information In Medicine, 22, (1983), 63-68.

[Gell et al., 1976] Gell, G., Oser, W. and Schwarz, G.: "Experiences with the AURA Free Text System", Radiology, 119, (1976), 105-109.

[Granitzer, 2006] Granitzer, M.: "KnowMiner: Konzeption und Entwicklung eines generischen Wissenserschliessungsframeworks", PhD Thesis TU Graz, Graz, (2006).

[Gregory et al., 1995] Gregory, J., Mattison, J. E. and Linde, C.: "Naming Notes - Transitions from Free-Text to Structured Entry", Methods of Information in Medicine, 34, (1995), 57-67.

[Hall and Walton, 2004] Hall, A. and Walton, G.: "Information overload within the health care system: a literature review", Health Information and Libraries Journal, 21, (2004), 102-108.

[Heyer et al., 2006] Heyer, G., Quasthoff, U. and T., W.: "Text Mining: Wissensrohstoff Text: Konzepte, Algorithmen, Ergebnisse", W3L, Bochum, (2006).

[Holzinger et al., 2007a] Holzinger, A., Geierhofer, R. and Errath, M.: "Semantic Information in Medical Information Systems - from Data and Information to Knowledge: Facing Information Overload", Procedings of I-MEDIA '07 and I-SEMANTICS '07, Graz, 2007a, 323-330.

[Holzinger et al., 2007b] Holzinger, A., Geierhofer, R. and Errath, M.: "Semantische Informationsextraktion in medizinischen Informationssystemen", Informatik Spektrum, 30, (2007b), 69-78.

[Holzinger et al., 2000] Holzinger, A., Kainz, A., Gell, G., Brunold, M. and Maurer, H.: "Interactive Computer Assisted Formulation of Retrieval Requests for a Medical Information System using an Intelligent Tutoring System", ED-MEDIA 2000, Montreal, (2000), 431-436.

[Hopf et al., 1999] Hopf, H. C., Deuschl, G., Diener, H. C. and Reichmann, H.: "Neurologie in Praxis und Klinik", Thieme, Stuttgart, (1999).

[Hotho et al., 2005] Hotho, A., Nürnberger, A. and Paaß, G.: "A Brief Survey of Text Mining", GLDV-Journal for Computational Linguistics and Language Technology, 20, (2005), 19-62.

[Jenssen et al., 2001] Jenssen, T. K., Laegrid, A., Komorowski, J. and Hovig, E.: "A literature network of human genes for highthroughput analysis of gene expression", Genetics, 28, (2001), 21-28.

[Le Quan et al., 2002] Le Quan, H., Sicilia-Garcia, E. I., Ji, M. and Smith, F. J.: "Extension of Zipf's law to words and phrases", Proceedings of the 19th international conference on Computational linguistics - Volume 1, (2002), 1-6.

[Leroy et al., 2003] Leroy, G., Chen, H., Martinez, J. D., Eggers, S., Flasey, R. R., Kislin, K. L., Huang, Z., Li, J., Xu, J., McDonald, D. M. and Ng, G.: "Genescene: Biomedical text and data mining, Proceedings of the ACM/IEEE-CS joint conference on Digital Libraries", (2003), 116-118.

[Liddy, 2000] Liddy, E. D.: "Interview with Gayle Curtis, Modem Media", (2000), online available: http://www.asis.org/Bulletin/Oct-00/liddy.html, last access: 2008-05-15

[Lovis et al., 2000] Lovis, C., Baud, R. H. and Planche, P.: "Power of expression in the electronic patient record: structured data or narrative text?", International Journal of Medical Informatics, 58, (2000), 101-110.

[Nasukawa and Nagano, 2001] Nasukawa, T. and Nagano, T.: "Text analysis and knowledge mining system", IBM SYSTEMS JOURNAL, 40, (2001), 967-984.

[Noone et al., 1998] Noone, J., Warren, J. and Brittain, M.: "Information overload: opportunities and challenges for the GP's desktop", Medinfo, 9, (1998), 1287-1291.

[Rajman and Besancon, 1998] Rajman, M. and Besancon, R.: "Text Mining: Natural Language Techniques and Text Mining Applications", In: Spaccapietra, S. and Maryansky, F., (eds.),

Data mining and reverse engineering: Searching for semantics, Chapmann and Hall, London, 1998, 50-64.

[Schlegel et al., 2003] Schlegel, U., Weller, M. and Westphal, M.: "Neuroonkologie", Thieme, Stuttgart, (2003).

[Sistrom and Honeyman-Buck, 2005] Sistrom, C. L. and Honeyman-Buck, J.: "Free text versus structured format: Information transfer efficiency of radiology reports", American Journal of Roentgenology, 185, (2005), 804-812.

[Smalheiser and Swanson, 1998] Smalheiser, N. R. and Swanson, D. R.: "Using ARROWSMITH: a computer-assisted approach to formulating and assessing scientific hypotheses", Computer Methods and Programs in Biomedicine, 57, (1998), 149-153.

[Srinivasan, 2004] Srinivasan, P.: "Text mining: Generating hypotheses from MEDLINE", Journal of the American Society for Information Science and Technology, 55, (2004), 396-413.

[Stephens et al., 2001] Stephens, M., Palakal, M., Mukhopadhyay, S., Raje, R. and Mostafa, J.: "Detecting gene relations from Medline abstracts", Pacific Symposium on Biocomputing, (2001), 483-496.

[Sullivan et al., 1999] Sullivan, F., Gardner, M. and Van Rijsbergen, K.: "An information retrieval service to support clinical decision-making at the point of care", British Journal of General Practice, 49, (1999), 1003-1007.

[Tan, 1999] Tan, A.-H.: "Text Mining: The state of the art and the challenges", PAKDD 1999 Workshop on Knowledge Discovery from Advanced Databases, Beijing, (1999), 65-70.

[Weeber et al., 2001a] Weeber, M., Klein, H., Berg, L. and Vos, R.: "Using concepts in literature-based discovery: Simulating Swanson's Raynaud-Fish Oil and Migraine-Magnesium discoveries", Journal of the American Society for Information Science and Technology, 52, (2001a), 548-557.

[Weeber et al., 2001b] Weeber, M., Mork, J. G. and Aronson, A. R.: "Developing a test collection for biomedical word sense disambiguation", Journal of the American Medical Informatics Association, (2001b), 746-750.

[Zingmond and Lenert, 1993] Zingmond, D. and Lenert, L. A.: "Monitoring Free-Text data using medical language processing", Computers and Biomedical Research, 26, (1993), 467-481.

[Zipf, 1949] Zipf, G. K.: "Human Behaviour and the Principle of Least-Effort", Addison-Wesley, Cambridge (MA), (1949).