Journal of Universal Computer Science, vol. 17, no. 7 (2011), 1109-1133
submitted: 30/10/10, accepted: 15/3/11, appeared: 1/4/11 © J.UCS

A Comparison of Different Retrieval Strategies Working on Medical Free Texts

Markus Kreuzthaler, Marcus D. Bloice
(Institute for Medical Informatics, Statistics and Documentation
Medical University of Graz, Austria
{markus.kreuzthaler, marcus.bloice}@medunigraz.at)

Lukas Faulstich
(ID Information und Dokumentation im Gesundheitswesen GmbH & Co. KGaA (ID)
Berlin, Germany
L.Faulstich@id-berlin.de)

Klaus-Martin Simonic, Andreas Holzinger
(Institute for Medical Informatics, Statistics and Documentation
Medical University of Graz, Austria
{klaus.simonic, andreas.holzinger}@medunigraz.at)

Abstract: Patient information in health care systems mostly consists of textual data, and free text in particular makes up a significant amount of it. Information retrieval systems that concentrate on these text types have to deal with the different challenges these medical free texts pose to achieve an acceptable performance. This paper describes the evaluation of four different types of information retrieval strategies: keyword search, search performed by a medical domain expert, a semantic based information retrieval tool, and a purely statistical information retrieval method. The different methods are evaluated and compared with respect to its appliance in medical health care systems.
Key Words: information retrieval, health care, medicine, evaluation, text mining
Category: H.3.3, J.3, I.1.3, I.2.7

1 Motivation

In the field of medical informatics there exists a basic differentiation between standardized and non-standardized text [van Bemmel and Musen 1997]. Standardized text is an element of a predefined set of phrases or terms, where typically the set is called a catalog. On the other hand non-standardized text is produced when the user can freely enter terms, phrases or sentences and is therefore often called free text in scientific research and literature.

Patient information mostly consists of textual data and in particular free text makes up a significant amount of this data. For example a patient record of a 2 1/2 year old child can contain more than 300 documents and gaining access to the relevant information according to an information need can be a non-trivial task [Holzinger et al. 2007]. These free texts play an especially important role in diagnostic findings (for example in pathology reports or radiology reports), information exchange between medical professionals, clinical research, medical accounting, and quality management in general [van Bemmel and Musen 1997].

Medical professionals are often confronted with this kind of textual information and in order to make this data more accessible, usable, and useful, smart information retrieval systems that can operate on these free texts are essential [Sager et al. 1994, Hripcsak and Wilcox 2002, Huske-Kraus 2003, Cohen and Hersh 2005, Geierhofer and Holzinger 2007, Holzinger et al. 2008].

1.1 Medical Free Texts

To provide a better understanding of the structure of medical free texts a typical example from a pathology report written in German is shown below:

> MITTELGRADIGE CHRONISCHE GASTRITS (MAGENMUCOSA VOM CORPUSTYP, UEBERGANGSTYP) MIT MITTELGRADIGER AKTIVITAET, KOMPLETTER UND INKOMPLETTER (TYP III) INTESTINALER METAPLASIE, MITTELGRADIGER ATROPHIE DER TIEFEN DRUESEN. ANTEIL EINES TUBULAEREN MAGENSCHLEIMHAUTADENOMS (INTESTINALER TYP; MITTELGRADIGE DYSPLASIE; WHO: GERINGGRADIGE INTRAEPITHELIALE NEOPLASIE). HP NICHT NACHWEISBAR.

The above text epitomizes the types of challenges that are inherent in medical free text analysis:

Memo style text characteristics. The text resembles a memo more than an orthographically correct piece of text. From a doctors point of view, however, the semantic of the text takes precedence over proper grammar or correct spelling.

Domain specific knowledge. The text contains much domain specific knowledge and words that are only properly understood by a domain expert.

Frequent use of abbreviations. Medical texts often contain abbreviations. In the text above, WHO stands for World Health Organization (which has a classification scheme for diseases), and the doctor is documenting the classification as GERINGGRADIGE INTRAEPITHELIALE NEOPLASIE. Even more frequent are abbreviations that are only resolvable when the context is known. HP in the context of gastritis stands for helicobacter-pylorii, but is often also used as an abbreviation for haptoglobin.

Typing errors. Typing errors can further complicate matters when analyzing text, as is illustrated by the misspelling of the word gastritis as GASTRITS.

The remainder of the paper is organized as follows: Section 2 provides a literature research regarding this paper's topic and shows the novelty of this work. Following this, in Section 3, common methods and an overview of information retrieval evaluation are given. Section 4 describes the information retrieval strategies under test in detail. After this, in Section 5, the evaluation results plus a discussion of the results are depicted. The last Section 6 concludes the paper and presents future work and ideas that we want to further investigate in this topic.

2 Related Work

In this section an overview of recent research in the field of information retrieval systems working on medical free texts is given. The literature research also reflects the novelty of the work presented in this paper.

[Hersh and Hickam 1998, Killoran and Hersh 1999] made a literature research from 1966 to 1998 in terms of *information storage and retrieval*, *information systems* and *evaluation studies*. The authors developed a framework for analyzing the outcome of the research and estimated the impact of electronic information retrieval systems in the medical health care domain.

[Baujard et al. 1998] describe MARVIN (multi-agent retrieval vagabond on information networks), which was one of the first web-based retrieval architectures for medicine and health care. [Bin et al. 2001] compared medical domain specific web search engines such as their own MediAgent, with general purpose engines and found out that they have approximately the same retrieval performance.

[Houston et al. 2000] evaluated term suggestion methods on three different thesauri (MeSH, UMLS, and document based) and its impact on retrieval performance. They evaluated their approach using the CANCERLIT records, which is a pool of abstracts from biomedical journals.

[Volk et al. 2002] developed a concept-based cross language information retrieval methodology in the MUCHMORE project. They evaluated their approach on medical scientific abstracts written in English and German from the Springer website. Multi-language retrieval achieved about the same performance as German monolingual retrieval results.

In [Mao and Chu 2002] phrases were introduced instead of stems or concepts as a basis for the vector space model used for information retrieval. They evaluated their approach using the OSHUMED test collection, which is a subset of the MEDLINE database and showed that their approach yields a 16% increase in the 11 Point average retrieval accuracy over the stem-based vector space model.

[Hliaoutakis et al. 2006] tested different semantic similarity methods and evaluated them using WordNet and MeSH. The best semantic similarity method was applied to extend the vector space model and the according weighting scheme. They tested this new model on the OSHMUMED collection and described the performance improvement of their algorithm.

[Liu and Chu 2007] experimented with query expansion techniques of the original query and showed that by using a domain knowledge model (UMLS) the expanded query in comparison with a scenario specific weighting system has a positive effect on the overall retrieval performance. A vector space model was used as a retrieval model and OSHUMED as a test collection.

[Moskovitch et al. 2007] compared concept-based search, context-sensitive and full text search, by use of their Vaidurya architecture. For evaluation purposes the National Guideline Clearinghouse (NCG) collection was used and they defined 13 information needs. Concept based retrieval outperformed full text search and they found that the more ontological elements used, the better the results.

[Abdou and Savoy 2008] evaluated 7 different vector-space schemes and three probabilistic models with help of the MEDLINE corpus. They compared the different weighting schemes and found that the I(n)B2 probabilistic model performed best. Further, they achieved an improvement of up to 13.5% in the mean average precision when including MeSH terms in the retrieval process.

[Trieschnigg et al. 2009] were evaluating 6 different MeSH (Medical Subject Heading) classification systems. For evaluating the classification performance they used the TREC Genomics collection. They also measured the impact on information retrieval tasks using a subset of MEDLINE. Therefore, a user's query is annotated with MeSH concepts.

[Fautsch and Savoy 2010] adapted the term frequency-inverse document frequency (tf-idf) weighting scheme for the vector space model to different domains in information retrieval. They tested their approach on different corpora also including the GENOMICS track, a collection of abstracts and citations from publications in the biomedical domain. The results show the positive impact of their weighting scheme in contrast to the standard tf-idf weighting scheme.

[Mu et al. 2010] investigated research on different search strategies observed when adding a term browser and a tree browser (MeSH) to a normal search browser to support the user in fulfilling their medical information need. They evaluated their system by having 30 persons fulfill 3 information needs from the OSHUMED corpus.

All of the research papers presented above share one common aspect; namely, that the gold standard corpora they used for the purpose of information retrieval testing were all drawn from biology literature or from abstracts in scientific literature. These texts *do not* have the typical difficulties that free texts in computer based patient records have (Section 1.1). Therefore, for the purposes of our research, we created our own gold standard using medical free texts. Of interest to us was how various retrieval strategies, which differ in their core implementation, perform relative to one another on these text types.

3 Evaluation of Information Retrieval Systems

The standard procedure used to measure information retrieval effectiveness comprises of the following three elements [Harter and Hert 1997, Zeng et al. 2002]:

- A document collection.
- A test suite of information requests, expressible as queries.
- A set of judgments for each query-document pair, which defines each pair as either relevant or not relevant.

The common approach to information retrieval system evaluation is based on the exact notion of relevant and non-relevant documents. In the context of information retrieval, relevance describes how well a retrieved set of documents (or a single document) meets the information need of the user. In other words, with respect to a user information need, a document in the test collection is given a binary classification as either relevant or non-relevant [van Bemmel and Musen 1997]. This decision is referred to as a gold standard or ground truth judgment of relevance. The test collection should be a sample of the kinds of text that will be encountered in the operational setting of interest [Wingert 1986], and its relevance is assessed according to an information need [Robertson and Hancock-Beaulieu 1992]. An important aspect of this is that a query for an information retrieval tool is not the information need. Rather an information need can be expressed in terms of a query language for an information retrieval tool. For example, some standard test collections that are often used by information retrieval researchers are the different tracks from the Text REtrieval Conference and GOV2 (a very large web page collection).

Once a test collection to be used as a basis to test an information retrieval system has been chosen, a metric for the system comparison must be decided upon. Basically, this can be separated into two groups of pure statistical performance measures; namely, metrics for *unranked retrieval results* and *ranked retrieval results*. Classical information retrieval metrics that are widely used in literature are the Recall, Precision, Fallout, and F-Measure metrics. A lot of effort has been invested into finding new evaluation measures over the past few years, one of the most famous recently introduced being bpref [Buckley and Voorhees 2004]. Other common information retrieval metrics that are concerned with ranked retrieval results are R-Precision, Precision at k, Mean Average Precision (MAP), and Normalized Discounted Cumulative Gain (NCDG). A good explanation and overview of these and other current information retrieval metrics can be found in standard text books [Baeza-Yates et al. 1999, Manning et al. 2008]. However, other factors do exist that should also be considered when evaluating an information retrieval system. [Saracevic 1995] identified six different levels of information retrieval evaluation:

- Engineering level
- Input level
- Processing level
- Output level
- Use and user level
- Social level

The human factor and human information behavior, in the context of information retrieval systems, are especially important factors to consider when developing such systems. Such systems must be capable of satisfying user needs. Although getting the right answers according to an information need is one of the most important parts of an information retrieval system, human factors should also be considered [Lew et al. 2006].

3.1 The Gold Standard

Due to the lack of available gold standards in the field of medical free texts [Kreuzthaler et al. 2010], we had to create our own. We therefore took a significant sample, which reflects the diversity of the text base, out of a pool of pathology reports. The text samples were anonymized and tagged with ICD (International Classification of Diseases) Codes from two independent medical professionals. If there was a disagreement about a tag, a third expert was consulted. Currently our gold standard, which was used on the different information retrieval strategies under test, consists of 3542 diagnosis texts. A typical example from this pool is shown in Section 1.1.

4 Retrieval Strategies under Test

This section describes in detail the different retrieval strategies that were used for a comparative performance evaluation. Therefore we outline the principles of the search strategy of a medical domain expert and distinguish this strategy from a pure keyword search. Afterwards, a technical description of the semantic based information retrieval tool is provided. The section is concluded by describing the mathematical principles of the implemented statistical retrieval method.

4.1 Medical Domain Expert and Key Word Search

For scientific research purposes, physicians often have information needs such as, for example: *Return all diagnose texts which comprise an intestinal neoplasm.* One of the main challenges encountered by medical domain experts is the translation of the information need to a query language for a particular information retrieval tool that is able to fulfill the request. Despite searching for fields which contain standardized text, such as a date, search statements typically search for patterns in free text.

Contrary to a Web information search, there is a much higher recall and precision expectancy from a search performed within a medical environment. Furthermore, in scenarios other than medical research, it is of crucial importance that all documents necessary for the patient's treatment are at the disposal of the doctors and physicians within the shortest possible time as often, for example in an emergency, the time available for viewing the data is severely limited. A simple database and/or text search is frequently not a sufficient information system with which to support the doctors and physicians effectively.

As mentioned previously, a crucial aspect for the medical domain expert is the translation of the information need into a correct query statement with the important boundary condition to maximize precision and recall. Typically this is arranged by a query expansion which contains topographically and morphologically variations and sub terms of an actual query term. Therefore, two main factors are crucial:

Experience of the text sorts. As described in Section 1.1, medical free texts pose several linguistic challenges. Therefore it is of the utmost importance that the medical domain expert who fulfills the information need knows about the typical difficulties these text types pose.

Domain specific knowledge. Medical domain knowledge is needed for understanding the information need and to query expand the transformed information need.

To underpin the two statements made above, we want to provide an example of the translated information need shown previously and explain the query in more detail:

> *SELECT ID,DI FROM befunde JOIN details ON befunde.ID = details.BefundID WHERE (((befunde.DI) Like '%kolon%'Or (befunde.DI) Like '%colon%' Or (befunde.DI) Like '%darm%' Or (befunde.DI) Like '%re_t%' Or (befunde.DI) Like '%sigm%' Or (befunde.DI) Like '%duoden%' Or (befunde.DI) Like '%jejun%' Or (befunde.DI) Like '%hemi_olektomie%' Or (befunde.DI) Like '%ileum%' Or (befunde.DI)*

Like '%ileo%' Or (befunde.DI) Like '%appendix%' Or (befunde.DI) Like '%coe_um%' Or (befunde.DI) Like '%zoe_um%')) AND ((befunde.DI) Like '%karzinom%' Or (befunde.DI) Like '%adenom%' Or (befunde.DI) Like '%kar_inoid%' Or (befunde.DI) Like '%Car_inoid%' Or (befunde.DI) Like '%lymphom%' Or (befunde.DI) Like '%NHL%' Or (befunde.DI) Like '%sar_om%' Or (befunde.DI) Like '%myom%' Or (befunde.DI) Like '%neurom%' Or ((befunde.DI) Like '%tumor%' And (befunde.DI) Like '%neuroendokrin%') Or ((befunde.DI) Like '%intraepithel%' and (befunde.DI) Like '%neoplasie%'))

Intestine		*Neoplasm*	
Expression Syntax	*Terms to hit*	*Expression Syntax*	*Terms to hit*
%kolon%, %colon%	Kolon, Colon	%karzinom%	Karzinom
%re_t%	Rektum, Rectum	%adenom%	Adenom
%sigm%	Sigmoideum	%kar_inoid%	karzinoid, karcinoid
%duoden%	Duodenum	%car_inoid%	carzinoid, carcinoid
%jejun%	Jejunum	%lymphom%	Lymphom
%hemi_olektomie%	Hemikolektomie, Hemicolektomie	%sar_om%	Sarkom, Sarcom
%ileum%	Ileum	%myom%	Myom
%ileo%	Ileo	%neurom%	Neurom
%appendix%	Appendix	%tumor%	Tumor
%coe_um%	Coecum, Coekum	%neuroendokrin%	neuroendokrin
%zoe_um%	Zoekum, Zoecum	%intraepithel%	intraepithel
		%neoplasie%	Neoplasie
		%NHL%	Non-Hodgkin Lymphoma

Table 1: Query terms used by the medical domain expert.

As can be seen in Table 1, the expert searches for all linguistics variations of *intestine* where the different parts of what the intestine consists of are typically mentioned. This is also true of the word *neoplasm*, where different linguistic variations and forms of neoplasms are searched for.

In contrast to the expert who translates the information need according to their experience about the text type and the domain specific knowledge that they have, we define a keyword search in this paper as a simple, non-expanded text pattern search.

4.2 Semantic-Based Information Retrieval Tool

The semantic retrieval tool under test uses a linguistic processing pipeline to analyze documents. The result of this analysis are semantic representations that roughly assign each sentence a set of concepts from the WNC terminology (see below) occurring in the sentence. Negated occurrences of concepts are detected and excluded from query results. The extracted semantic representations are stored together with the original documents.

When used as a query language, Boolean expressions of terms in disjunctive form are supported. Internally, arbitrary propositional Boolean expressions over terms are used. Each term occurring in a query is expanded using the medical semantic network ID MACS® (MSN, see below). Expanded queries are matched against the stored semantic representations. Currently, only Boolean retrieval is supported. Matching documents are scored by taking into account the semantic distance between a query term and the matching document term measured as a path length within the MSN.

The domain knowledge used in the semantic retrieval system is modeled in the form of the medical semantic network ID MACS® (MSN). It uses the Wingert Nomenclature (WNC) as its medical terminology. The WNC is based on the German version of SNOMED developed by Friedrich Wingert. Although its main focus is on German, it, to a lesser extent, supports several other languages including English and French. The MSN forms a simple ontology whose concepts are organized in a taxonomy (isA-hierarchy) and a merology (anatomical partOf-hierarchy). Further relations between concepts are modeled by labeled edges. The MSN is divided into several subdomains, including:

- topography (i.e., anatomical concepts)
- morphology (e.g., fracture, fever)
- function (e.g., respiration)
- diseases (e.g., glaucoma)
- agents (e.g., pathogens, pharmaceutical substances)

Currently, the MSN contains more than 90,000 terms and 300,000 unique relations.

Input documents are analyzed using a linguistic processing pipeline; the text content is extracted from the documents, sectioned and split into words. Known abbreviations are expanded using a large database of medical abbreviations collected by ID. Since many abbreviations are ambiguous, a context dependent algorithm to disambiguate such cases is used.

Since medical language, and in particular German medical language, contains many compound words such as "Ot| o|rhin|o|laryng|o|log|y" or

"Gran|ulos|a|epi|theli|om|e", morphological analysis is absolutely necessary. The methods developed by Wingert [Wingert 1985a] and Goettsche are used for morphological segmentation. A large set of hand crafted segmentation rules tailored to the medical domain ensures that most correctly spelled words in typical input documents can be segmented.

Words that cannot be successfully segmented are likely to be typos or ad-hoc abbreviations (clippings), e.g., "insuff" for "insufficiency". This is particularly true of medical notes that are used for internal communication, where a significant percentage of such words exist. A large word list and a correction algorithm based on a sophisticated edit-distance is used to assign proper correction and completion candidates to those words.

Since sentence segmentation requires classification of periods as abbreviation and/or sentence terminators, this step is carried out after the abbreviations and clippings have been identified.

Medical documents often report negative findings, such as the exclusion of a possible diagnosis for the symptoms of the patient, and it is therefore important to understand their syntactic structure. Hence a chunk parser is used to detect basic phrases (chunks) without necessarily providing a complete syntactic analysis of the whole sentence. The parser flags negated phrases, quantitative expressions and secondary information. The phrase structure returned by the parser is then used as input for the indexation algorithm (based again on the work of Wingert [Wingert 1985b] and Goettsche) that identifies terms from the WNC terminology in the linguistically analyzed text. This algorithm in particular takes care of the frequent multi-word terms common in medical language such as "aortocoronary bypass", "otitis media", "vena cava inferior" etc. Furthermore it is robust with respect to continuous/discontinuous formulations of the same term such as "tonsillektomie" and "ectomy of tonsils". The indexation algorithm works recursively and thus also provides robustness with respect to synonyms such as "excision of tonsils". The result of this indexation is a graph-based semantic representation that connects the words and phrases of a sentence with the concepts they are referring to. Furthermore, syntactic relations between the phrases are used to add relations between the extracted concepts. This semantic representation is stored together with the original document and is searched during query processing. In particular, negative occurrences of concepts such as diagnoses explicitly ruled out by the physician are recognized by the parser and are excluded from retrieval results.

Figure 1 gives an overview of the linguistic processing pipeline that describes the steps that are performed from the document to its semantic representation.

The query language follows a simple grammar, namely:

```
Query::= Disjunction
Disjunction::= Conjunction | Conjunction ";" Disjunction
```

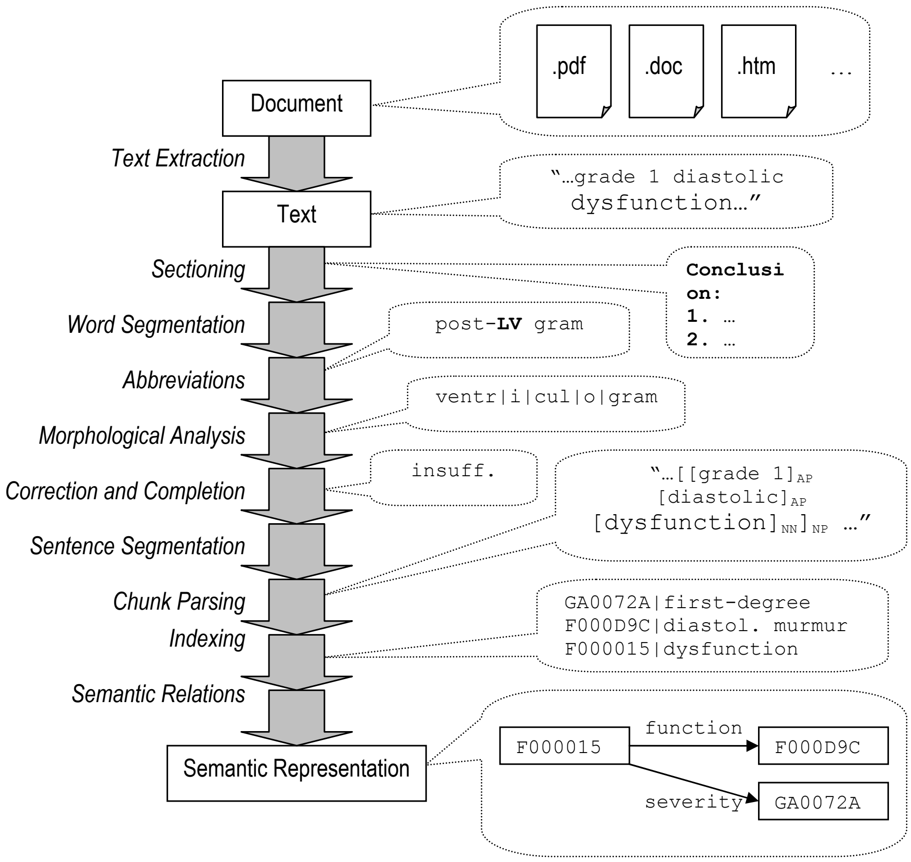


Figure 1: Linguistic Processing Pipeline.

```
Conjunction::= Atom | Atom ","Conjunction
Atom::= Term | "!" Term
```

Thus a query forms a Boolean expression in disjunctive form over search terms. Semantic query expansion has been discussed in several previous works [Aronson et al. 1994, Efthimiadis 1996, Kingsland et al. 1993]. The approach is as follows: each search term is indexed (using the linguistic processing methods described above) and replaced by the identifier of the WNC concept matching the term. These concept identifiers are called WNC indices. If the search term refers to a combination of several concepts in the WNC (e.g., Gastroparesis= Stomach + Paresis), the search term is replaced by a conjunction of the WNC indices. On the other hand, if a conjunction of search terms can be indexed as a single WNC concept (e.g., "inflammation, esophagus" = Esophagitis), then this conjunction is replaced by the single WNC index.

Each WNC index in the query is then replaced by a disjunction of this WNC index and indices of related WNC concepts. A concept is related if it is more specific with respect to the taxonomy or merology of the MSN. A maximum distance within the MSN can be specified. In [Faulstich et al. 2008] some experimentation results, regarding the optional inclusion of more general concepts, are described.

The expanded query is matched against the stored semantic representations sentencewise. The semantic distance between a WNC concept matched in a sentence and the WNC concept corresponding to the original search term is used for scoring. The document score is computed as the maximum of its sentence scores. In addition, partial matches of sentences within the same document may combine to a complete match, yielding a lower score.

4.3 Latent Semantic Analysis

The most common statistical retrieval methods working on free text are Latent Semantic Analysis (LSA) [Landauer and Dumais 1997, Landauer et al. 1998], Probabilistic Latent Semantic Analysis (PLSA) [Papadimitriou et al. 2000, Hofmann 2001] and Latent Dirichlet Allocation (LDA) [Blei et al. 2003]. In this paper we chose LSA as a first approach to analyze whether statistical retrieval methods are applicable for free text retrieval in the field of medicine.

Latent Semantic Analysis is both a theory and a method for both extracting and representing the meaning of words in their contextual environment by application of statistical analysis to a large amount of text. LSA is basically a general theory of acquired similarities and knowledge representations, originally developed to explain learning of words and psycholinguistic problems [Landauer and Dumais 1997], [Landauer et al. 1998]. The general idea was to induce global knowledge indirectly from local co-occurrences in the representative text. Originally, LSA was used for explanation of textual learning of the English language at a comparable rate amongst schoolchildren.

The most interesting issue is that LSA does not use any prior linguistic or perceptual similarity knowledge; i.e., it is based exclusively on a general mathematical learning method that achieves powerful inductive effects by extracting the right number of dimensions to represent both objects and contexts. The fundamental suggestion is that the aggregate of all words in contexts in which a given word does and does not appear provides a set of mutual constraints that largely determines the similarity of meaning of words and sets of words to each other. For the combination of Informatics and Psychology it is interesting to note that the adequacy of LSA's reflection of human knowledge has been established in a variety of ways [Foltz et al. 1998]. For example, the scores overlap closely to those of humans on standard vocabulary and subject matter tests and, interestingly, it emulates human word sorting behavior and category judgments [Landauer and Dumais 1997]. Consequently, as a practical outcome, it can estimate passage coherence and the learnability of passages, and both the quality and quantity of knowledge contained in an textual passage (originally these were student essays).

LSA is primarily used as a technique for measuring the coherence of texts. By comparing the vectors for two adjoining segments of text in a high-dimensional semantic space, the method provides a characterization of the degree of semantic relatedness between the segments. LSA can be applied as an automated method that produces coherence predictions similar to propositional modeling, thus having potential as a psychological model of coherence effects in text comprehension [Foltz et al. 1998].

Having t terms and d documents one can build a $t \times d$ matrix X, forming a vector space model [Salton et al. 1975, Boerjesson and Hofsten 1975], where typically this matrix is very sparse. Often the terms within this matrix are weighted according to term frequency-inverse document frequency (tf-idf) [Salton et al. 1975, Salton and Yang 1973]. The main method now is to apply the singular value decomposition on X. Therefore X can be disjointed into three components $X = TSD^T$. T and D^T are orthonormal matrices with the eigenvectors of XX^T and X^TX respectively. S contains the roots of the eigenvalues of XX^T and X^TX.

Reducing the dimensionality can now be achieved by step-by-step elimination of the lowest eigenvalue with the corresponding eigenvectors to a certain value k. A given Query q can now be projected into this space by applying the equation:

$$Q_k = q^T T_k S^{-1} \tag{1}$$

Having Q_k and the documents in the same semantic space, different similarity measures can now be applied. The so called cosine similarity between a document in the semantic space and a query Q_k is often used, for example. Having two document vectors V_1 and V_2 in the k dimensional space the cosine similarity is defined as:

$$cos(\phi) = \frac{V_1 \cdot V_2}{\|V_1\| \, \|V_2\|} \tag{2}$$

5 Evaluation Results

In this section we provide a description of the evaluation results of the different tested retrieval strategies and discuss the accomplished results. All the evaluation results were run against the developed gold standard, which is described in Section 3.1. Due to the diversity of the different retrieval strategies, different evaluation metrics had to be used (Section 3).

5.1 Defined Information Needs

We chose nine different information needs to test the different information retrieval strategies. Four of them contain an inflammation:

Information Need Appendicitis. *Full Text:* Return all diagnosis texts which comprise an appendicitis. *Information Need Translation Medical Domain Expert:* %appendi_itis% *Information Need Translation IR Tool:* appendicitis *Information Need Translation Keyword:* %appendicitis%

Information Need Colitis. *Full Text:* Return all diagnosis texts which comprise a colitis. *Information Need Translation Medical Domain Expert:* %colitis%, %kolitis% *Information Need Translation IR Tool:* colitis *Information Need Translation Keyword:* %colitis%

Information Need Gastritis. *Full Text:* Return all diagnosis texts which comprise a gastritis. *Information Need Translation Medical Domain Expert:* %gastritis% *Information Need Translation IR Tool:* gastritis *Information Need Translation Keyword:* %gastritis%

Information Need Hepatitis. *Full Text:* Return all diagnosis texts which comprise a hepatitis. *Information Need Translation Medical Domain Expert:* %hepatitis%, %Nash% *Information Need Translation IR Tool:* hepatitis *Information Need Translation Keyword:* %hepatitis%

Five information needs contain a neoplasm and a location, where we decided to put in a synonym of the location (colon, dickdarm) and a synonym of the neoplasm (neubildung, neoplasie, tumor).

Information Need Adenokarzinom, Colon. *Full Text:* Return all diagnose texts which comprise a adenokarzinom in the colon. *Information Need Translation Medical Domain Expert:* %kolon%, %colon%, %dickd%, %re_t% %sigm% %asc% %desc% %trans% %flex% %adeno_ar_inom% *Information Need Translation IR Tool:* adenokarzinom, colon *Information Need Translation Keyword:* %adenokarzinom%, %colon%

Information Need Adenokarzinom, Dickdarm. *Full Text:* Return all diagnose texts which comprise an adenokarzinom in the colon. *Information Need Translation Medical Domain Expert:* %kolon%, %colon%, %dickd%, %re_t% %sigm% %asc% %desc% %trans% %flex% %adeno_ar_inom% *Information Need Translation IR Tool:* adenokarzinom, dickdarm *Information Need Translation Keyword:* %adenokarzinom%, %dickdarm%

Information Need Neubildung, Darm. *Full Text:* Return all diagnose texts which comprise an intestinal neoplasm. *Information Need Translation Medical Domain Expert:* Described in detail in Section 4.1 and Table 1 *Information Need Translation IR Tool:* neubildung, darm *Information Need Translation Keyword:* %neubilding% %darm%

Information Need Neoplasie, Darm. *Full Text:* Return all diagnose texts which comprise an intestinal neoplasm. *Information Need Translation Medical Domain Expert:* Described in detail in Section 4.1 and Table 1 *Information Need Translation IR Tool:* neoplasie, darm *Information Need Translation Keyword:* %neoplasie% %darm%

Information Need Tumor, Prostata. *Full Text:* Return all diagnose texts which comprise a prostata neoplasm. *Information Need Translation Medical Domain Expert:* %prostata%, %ar_inom% %neoplasie% %prostata% *Information Need Translation IR Tool:* tumor, prostata *Information Need Translation Keyword:* %tumor% %prostata%

5.2 LSA Pre-Processing and Similarity Measures

No specialized text pre-processing was accomplished considering the fact of processing medical free text and their typical challenges but a standard processing chain supported by Java Lucene[1] was used. The built-in Lucene German stop word list was applied to the processing chain, which resulted in a 5569 x 3542 Term Document Matrix. The tf-idf (Section 4.3) weighting scheme was applied to the matrix.

An as of yet unknown issue arises when regarding the degree of the dimensionality reduction when applying SVD. Therefore, we evaluated the area under the 11 Point Average Precision-Recall Graph for different feature space dimensions to get the best insight as to what dimensionality size fits best to the retrieval method. Our experiments have shown that the maximum performance is achieved at a feature space dimensionality of 400 (Figure 2). Increasing the feature space dimensionality above 400 slightly decreases the retrieval performance in our setup. Therefore, having a feature space dimensionality of 400 resulted in a dimensionality reduction of about 93%. All transformed information needs were mapped into this space, thereby representing one point in the feature space. The ranking of the documents was achieved by applying the cosine similarity measure (Section 4.3).

[1] http://lucene.apache.org

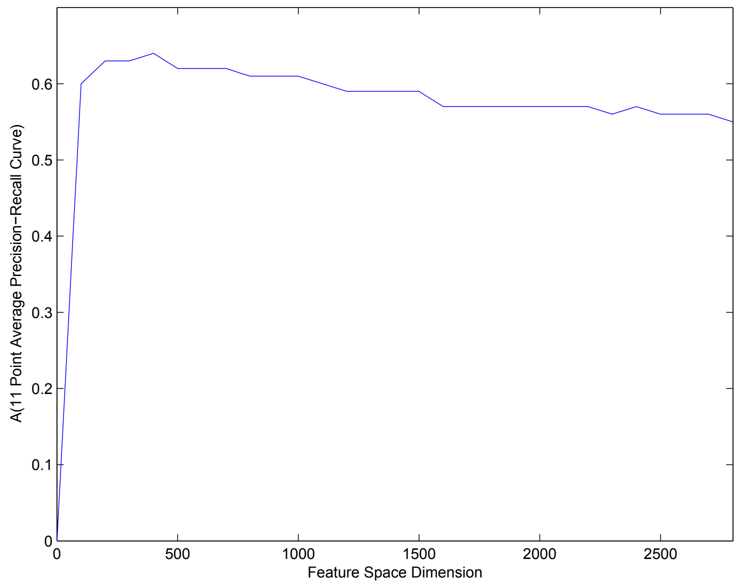


Figure 2: Estimating the feature space dimension.

5.3 Results

To get comparable performance values, taking into consideration the fact that unranked retrieval results (Medical Domain Expert, Keyword, IR Tool One), are compared with ranked results (LSA), we decided to map the average precision recall values from the different retrieval strategies onto the 11 Point Average Precision-Recall Graph gained from applying LSA. The results are shown on Figure 3. The precision recall tables for the individual information needs are depicted in Appendix B. The mapping of the individual performance values on to the precision recall graph of the LSA method is shown in Appendix A.

It must also be mentioned that two semantic-based information retrieval tools were under test, which we refer to as *IR Tool One* and *IR Tool Two* respectively. Despite the fact that they use the same linguistic processing pipeline for interpreting medical texts and query terms (Section 4.2), IR Tool One employs a proprietary retrieval algorithm, while IR Tool Two builds on the Apache Lucene text search engine library. IR Tool Two has a strong hierarchical ranking of the retrieval results, but the amount of returned results is limited to 50. Conversely, for IR Tool One, the number of search results are not limited to any particular amount and the results are not strongly hierarchically ordered, so we evaluated this information retrieval tool with the precision and recall performance measure.

Getting comparable performance values between LSA and IR Tool Two we

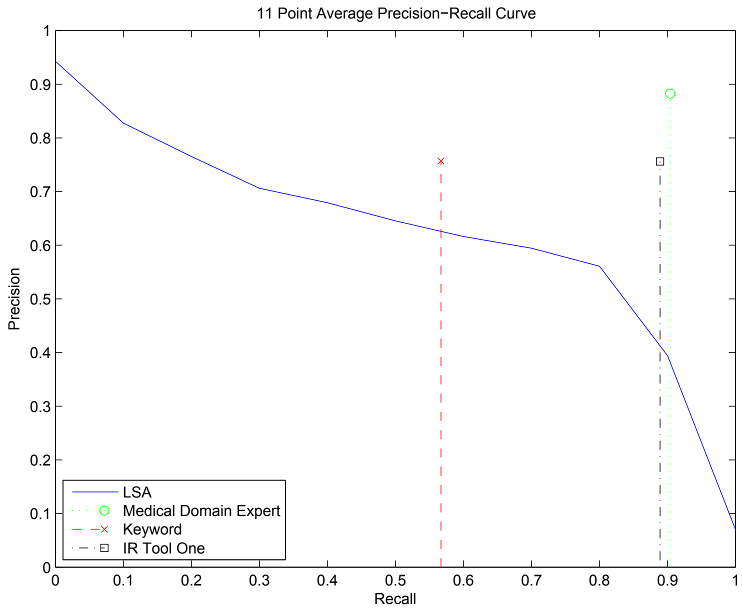


Figure 3: 11 Point Average Precision-Recall Graph

chose to calculate the mean average precision at the very first 50 search results. The results of this evaluation are depicted in Table 2.

5.4 Discussion

As can be seen from Figure 3 the medical domain expert outperforms the other retrieval methods, achieving high precision at a high recall level. Interestingly, the semantic based information retrieval tool achieves approximately the same recall level as the medical domain expert while having a lower precision value. This performance result is good remembering the fact what effort the medical domain expert has to make to translate the information need into a query string (Section 4.1). In contrast to this, the input for the information retrieval tool is short and clear so therefore less effort has to be made to transform the information need to the query language understood by the information retrieval tool.

Keyword search, as it is defined in Section 4.1, has a high precision value but a lower recall value. This result is clear when considering the fact that information needs that can be described by using these keyword(s) will achieve a high precision value. So, if documents are found they will be relevant but the recall level will generally suffer. Looking at Figure 3, keyword search achieves approximately the same precision as IR Tool One but a far worse recall. It is also

		IR Tool Two	*LSA*
Information Need	*Quantity*	*Average Precision*	*Average Precision*
Appendicitis	50	1	1
Colitis	50	0.596	0.800
Gastritis	50	0.988	1
Hepatitis	50	0.865	0.955
Adenokarzinom, Colon	50	0.891	0.664
Adenokarzinom, Dickdarm	50	0.806	0.652
Neubildung, Darm	50	0.742	0.588
Neoplasie, Darm	50	0.742	0.950
Tumor, Prostata	50	0.920	0.627
MAP		0.839	0.804

Table 2: Mean Average Precision for IR Tool Two and LSA.

possible that no search results are found at all when using the keyword search methodology as can be seen for the *Neubildung, Darm* information need (see Appendix B and Appendix A). In contrast to this, for this information need, IR Tool One has about the same precision recall levels as the medical domain expert, reflecting the semantic processing chain of the tool.

The LSA statistical retrieval method has, when compared to the other methods, a lower precision for all measured recall levels (Figure 3). However, as shown in Table 2, the performance is good compared to IR Tool Two for the top 50 documents returned. This result gives the impression that LSA is applicable for getting high precision values for a particular amount of search results but hard to use to achieve both high precision *and* high recall values, which is needed for example in clinical studies.

6 Conclusion and Future Work

In this paper we highlighted and compared a number of different retrieval strategies that work on medical free texts. Due to the lack of available gold standards [Kreuzthaler et al. 2010] in this area, we had to develop our own, which is a time consuming process. We evaluated the performance for a selected number of information needs in the field of medicine which were performed by a simple keyword search, a medical domain expert, two versions of a semantic based information retrieval tool, and a purely statistical retrieval method that treads the texts as a bag of words. The comparison of the different retrieval methods and their appliance for retrieval of medical free texts was evaluated solely by statistical evaluation measures.

Nevertheless, the following must be noted:

Gold Standard. The developed gold standard comprises only of German pathology diagnosis texts. Due to the fact that the retrieval strategies should be applicable in any field (e.g. radiology torax) it would be interesting to test them in other medical areas. As well as different areas, the feasibility of multi-language retrieval in medical free texts is of interest.

Negations. A difficult aspect in medical free text processing are the linguistic variations of negations. While both versions of the IR Tool handle negations quite well due to their parsing technology, LSA was applied on term vectors that did not distinguish between positive and negative occurrences of terms. As future work we plan to combine LSA with a more sophisticated text processing pipeline as used in the IR Tool.

Complexity. Physicians' information needs normally comprise of more complex information needs. Nevertheless, the information needs described in this paper form a base, so that when evaluated they reflect a trend for applicability. Typically, in the field of clinical research, not one data pool is searched, rather, several sources that form a pool of up to a few million diagnosis texts are used.

Future work in this area will concentrate on how to enhance the recall of statistical retrieval methods by using other statistical retrieval models (PLSA, LDA). Also, an evaluation of the benefit of using a terminology such as SNOMED CT to enhance the statistical retrieval process will be performed.

With regard to gold standards, another important question that arises is how to get access to a larger pool of freely available gold standards in the medical domain. Extending this point further, how should the issue of being able to access tagged medical *objects* be tackled? Future research aims to discuss the creation of a framework and a proof of concept for a collaborative annotation service that guarantees a certain level of quality of the annotated objects.

We are also interested in how well neural networks perform annotation tasks on medical free texts, which could be further on be used to semi automatically generate gold standards. In this context, Informatics for Integrating Biology and the Bedside (i2b2) has to be mentioned because they started recently a series of natural language challenges which work on de-identified clinical records. A first data set [Uzuner et al. 2007] is available (889 unannotated, de-identified discharge summaries) and a second one will be available in November 2010. Another idea to get an open available gold standard would be to bring in this retrieval challenge as a Text REtrieval Conference track.

However, for now, the work in this paper and the evaluation results presented therein, constitute our contribution towards the important field of information retrieval in medical health care systems.

References

[Abdou and Savoy 2008] S. Abdou and J. Savoy. "Searching in Medline: Query expansion and manual indexing evaluation"; *Information Processing & Management*, 44(2):781–789, 2008.

[Aronson et al. 1994] A.R. Aronson, T.C. Rindflesch, and A.C. Browne. "Exploiting a large thesaurus for information retrieval"; In *Proceedings of RIAO*, volume 94, pages 197–216. Citeseer, 1994.

[Baeza-Yates et al. 1999] R. Baeza-Yates, B. Ribeiro-Neto, et al. "Modern information retrieval"; Addison-Wesley Reading, MA, 1999.

[Baujard et al. 1998] O. Baujard, V. Baujard, S. Aurel, C. Boyer, and RD Appel. "Trends in medical information retrieval on Internet"; Computers in Biology and Medicine, 28(5):589–601, 1998.

[Bin et al. 2001] L. Bin et al. "The retrieval effectiveness of medical information on the web"; International journal of medical informatics, 62(2-3):155–163, 2001.

[Blei et al. 2003] D.M. Blei, A.Y. Ng, and M.I. Jordan. "Latent dirichlet allocation"; The Journal of Machine Learning Research, 3:993–1022, 2003.

[Boerjesson and Hofsten 1975] E. Boerjesson and C. Hofsten. "A vector model for perceived object rotation and translation in space"; Psychological Research, 38(2): 209–230, 1975.

[Buckley and Voorhees 2004] C. Buckley and E.M. Voorhees. "Retrieval evaluation with incomplete information"; In Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval, pages 25–32. ACM, 2004.

[Cohen and Hersh 2005] A.M. Cohen and W.R. Hersh. "A survey of current work in biomedical text mining"; Briefings in Bioinformatics, 6(1):57, 2005.

[Efthimiadis 1996] E.N. Efthimiadis. "Query expansion"; Annual review of information science and technology, 31:121–187, 1996.

[Faulstich et al. 2008] L.C. Faulstich, F. Müller, A. Sander, R. Pitzler, M. Errath, and A. Holzinger. "Semantisches Retrieval medizinischer Freitexte"; In 53. Jahrestagung der Deutschen Gesellschaft für Medizinische Informatik, Biometrie und Epidemiologie (gmds), Stuttgart, 2008. German Medical Science GMS Publishing House.

[Fautsch and Savoy 2010] C. Fautsch and J. Savoy. "Adapting the tf idf vector-space model to domain specific information retrieval"; In Proceedings of the 2010 ACM Symposium on Applied Computing, pages 1708–1712. ACM, 2010.

[Foltz et al. 1998] P.W. Foltz, W. Kintsch, and T.K. Landauer. "The measurement of textual coherence with latent semantic analysis"; Discourse processes, 25:285–308, 1998.

[Geierhofer and Holzinger 2007] R. Geierhofer and A. Holzinger. "The evaluation of semantic tools to support physicians in the extraction of diagnosis codes"; In Proceedings of the 3rd Human-computer interaction and usability engineering of the Austrian computer society conference on HCI and usability for medicine and health care, pages 403–408. Springer-Verlag, 2007.

[Harter and Hert 1997] S.P. Harter and C.A. Hert. "Evaluation of Information Retrieval Systems: Approaches, Issues, and Methods"; Annual Review of Information Science and Technology (ARIST), 32:3–94, 1997.

[Hersh and Hickam 1998] W.R. Hersh and D.H. Hickam. "How well do physicians use electronic information retrieval systems?: A framework for investigation and systematic review"; *JAMA*, 280(15):1347, 1998.

[Hliaoutakis et al. 2006] A. Hliaoutakis, G. Varelas, E. Voutsakis, E.G.M. Petrakis, and E. Milios. "Information Retrieval by Semantic Similarity"; International Journal on Semantic Web & Information Systems, 2(3):55–73, July-September 2006.

[Hofmann 2001] T. Hofmann. "Unsupervised Learning by Probabilistic Latent Semantic Analysis"; Machine Learning, 42:177–196, 2001.

[Holzinger et al. 2007] A. Holzinger, R. Geierhofer, and M. Errath. "Semantische Informationsextraktion in medizinischen Informationssystemen"; Informatik-Spektrum, 30(2):69–78, 2007.

[Holzinger et al. 2008] A. Holzinger, R. Geierhofer, F. Mödritscher, and R. Tatzl. "Semantic Information in Medical Information Systems: Utilization of Text Mining Techniques to Analyze Medical Diagnoses"; Journal of Universal Computer Science, 14(22):3781–3795, 2008.

[Houston et al. 2000] A.L. Houston, H. Chen, B.R. Schatz, S.M. Hubbard, R.R. Sewell, and T.D. Ng. "Exploring the use of concept spaces to improve medical information retrieval"; Decision Support Systems, 30(2):171–186, 2000.

[Hripcsak and Wilcox 2002] G. Hripcsak and A. Wilcox. "Reference standards, judges, and comparison subjects"; Journal of the American Medical Informatics Association, 9(1):1, 2002.

[Huske-Kraus 2003] D. Huske-Kraus. "Text generation in clinical medicine-a review"; Methods of information in medicine, 42(1):51–60, 2003.

[Killoran and Hersh 1999] E. Killoran and W. Hersh. "Electronic information retrieval by physicians and medical librarians"; JAMA, 281(14):1272, 1999.

[Kingsland et al. 1993] L.C. Kingsland et al. "Coach: applying UMLS knowledge sources in an expert searcher environment"; Bulletin of the Medical Library Association, 81(2):178, 1993.

[Kreuzthaler et al. 2010] M. Kreuzthaler, M.D. Bloice, K.M. Simonic, and A. Holzinger. "On the Need for Open Source Ground Truths for Medical Information Retrieval Systems"; In International Conference on Knowledge Management and Knowledge Technologies, volume 10, pages 371 – 381, September 2010.

[Landauer and Dumais 1997] T.K. Landauer and S.T. Dumais. "A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge"; Psychological review, 104(2):211–240, 1997.

[Landauer et al. 1998] T.K. Landauer, P.W. Foltz, and D. Laham. "An introduction to latent semantic analysis"; Discourse processes, 25:259–284, 1998.

[Lew et al. 2006] M.S. Lew, N. Sebe, C. Djeraba, and R. Jain. "Content-based multimedia information retrieval: State of the art and challenges"; ACM Transactions on Multimedia Computing, Communications, and Applications (TOMCCAP), 2(1): 1–19, 2006.

[Liu and Chu 2007] Z. Liu and W.W. Chu. "Knowledge-based query expansion to support scenario-specific retrieval of medical free text"; Information Retrieval, 10 (2):173–202, 2007.

[Manning et al. 2008] C.D. Manning, P. Raghavan, and H. Schütze. "Introduction to information retrieval"; Cambridge Univ Pr, 2008.

[Mao and Chu 2002] W. Mao and W.W. Chu. "Free-text medical document retrieval via phrase-based vector space model"; In Proceedings of the AMIA Symposium, page 489. American Medical Informatics Association, 2002.

[Moskovitch et al. 2007] R. Moskovitch, S.B. Martins, E. Behiri, A. Weiss, and Y. Shahar. "A comparative evaluation of full-text, concept-based, and context-sensitive search"; Journal of the American Medical Informatics Association, 14(2): 164–174, 2007.

[Mu et al. 2010] X. Mu, K. Lu, and H. Ryu. "Search strategies on a new health information retrieval system"; Online Information Review, 34(3):440–456, 2010.

[Papadimitriou et al. 2000] C.H. Papadimitriou, P. Raghavan, H. Tamaki, and S. Vempala. "Latent semantic indexing: A probabilistic analysis"; Journal of Computer and System Sciences, 61(2):217–235, 2000.

[Robertson and Hancock-Beaulieu 1992] S.E. Robertson and M.M. Hancock-Beaulieu. "On the evaluation of IR systems"; Information Processing & Management, 28(4): 457–466, 1992.

[Sager et al. 1994] N. Sager, M. Lyman, C. Bucknall, N. Nhan, and L.J. Tick. "Natural language processing and the representation of clinical data"; Journal of the American Medical Informatics Association, 1(2):142, 1994.

[Salton and Yang 1973] G. Salton and CS Yang. "On the Specification of Term Values in Automatic Indexing"; The Journal of documentation, 29(4):351, 1973.

[Salton et al. 1975] G. Salton, A. Wong, and CS Yang. "A vector space model for automatic indexing"; Communications of the ACM, 18(11):620, 1975.

[Saracevic 1995] T. Saracevic. "Evaluation of evaluation in information retrieval"; In Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval, pages 138–146. ACM, 1995.

[Trieschnigg et al. 2009] D. Trieschnigg, P. Pezik, V. Lee, F. De Jong, W. Kraaij, and D. Rebholz-Schuhmann. "MeSH Up: effective MeSH text classification for improved document retrieval"; Bioinformatics, 25(11):1412, 2009.

[Uzuner et al. 2007] O. Uzuner, Y. Luo, and P. Szolovits. "Evaluating the State-of-the-Art in Automatic De-identification"; Journal of the American Medical Informatics Association, 14(5):550, 2007.

[van Bemmel and Musen 1997] J.H. van Bemmel and M.A. Musen. "Handbook of medical informatics"; Springer Berlin, 1997.

[Volk et al. 2002] M. Volk, B. Ripplinger, et al. "Semantic annotation for concept-based cross-language medical information retrieval"; International Journal of Medical Informatics, 67(1-3):97–112, 2002.

[Wingert 1985a] F. Wingert. "Morphologic analysis of compound words"; Methods of Information in Medicine, 24(3):155, 1985a.

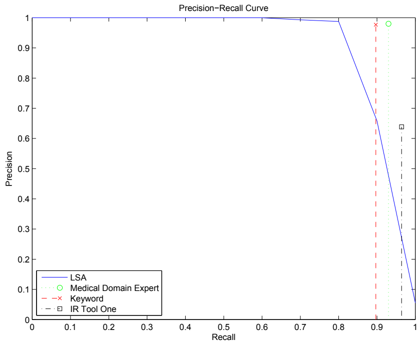
[Wingert 1985b] F. Wingert. "Automated indexing based on SNOMED"; Methods of information in medicine, 24(1):27–34, 1985b.

[Wingert 1986] F. Wingert. "An indexing system for SNOMED"; Methods of information in medicine, 25(1):22–30, 1986.

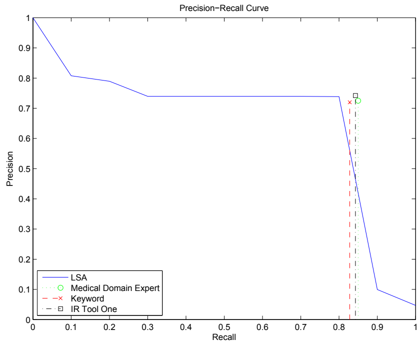
[Zeng et al. 2002] Q. Zeng, J.J. Cimino, and K.H. Zou. "Providing concept-oriented views for clinical data using a knowledge-based system"; Journal of the American Medical Informatics Association, 9(3):294, 2002.

A Precision Recall Graphs

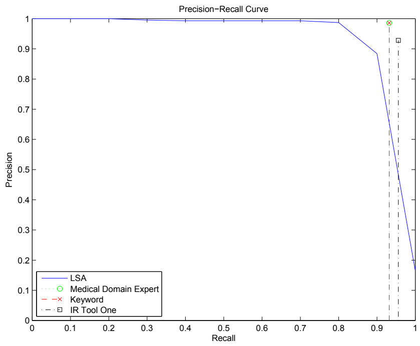
A.1 Inflammation



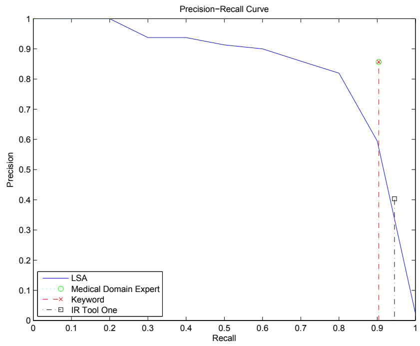
(a) Information Need Appendicitis



(b) Information Need Colitis



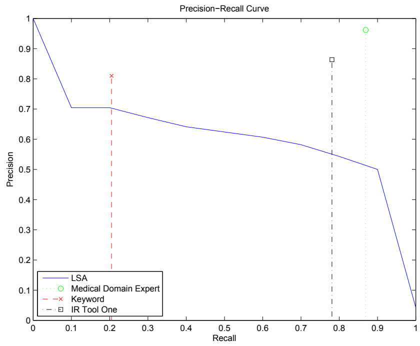
(c) Information Need Gastritis



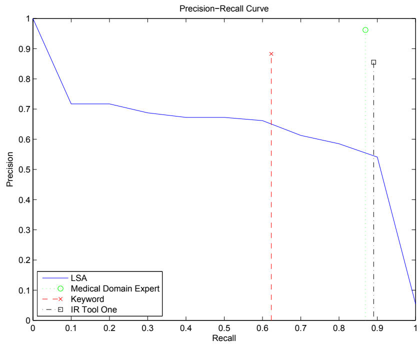
(d) Information Need Hepatitis

Figure 4: Precision Recall Graphs (1)

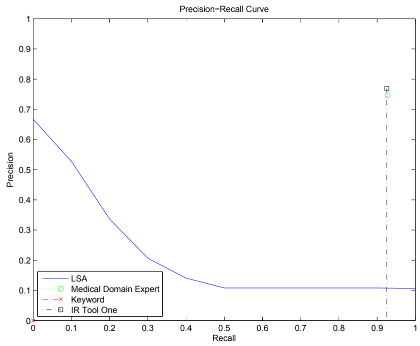
A.2 Neoplasm, Location



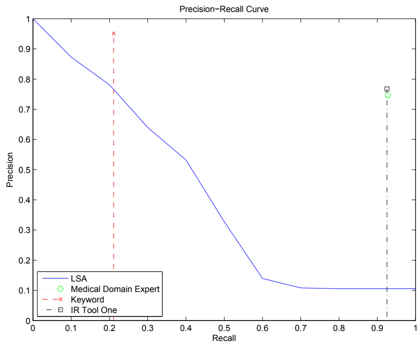
(a) Information Need Adenokarzinom, Colon



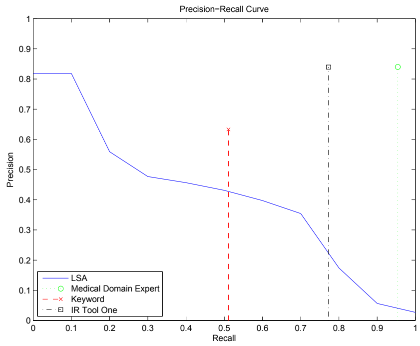
(b) Information Need Adenokarzinom, Dickdarm



(c) Information Need Neubildung, Darm



(d) Information Need Neoplasie, Darm



(e) Information Need Tumor, Prostata

Figure 5: Precision Recall Graphs (2)

B Precision Recall Tables

		Medical Domain Expert		
Information Need	*Quantity*	*Precision*	*Recall*	*LSA Precision at Recall*
Appendicitis	196	0.980	0.930	0.477
Colitis	140	0.725	0.850	0.419
Gastritis	567	0.986	0.932	0.654
Hepatitis	73	0.857	0.904	0.572
Adenokarzinom, Colon	146	0.962	0.869	0.513
Adenokarzinom, Dickdarm	146	0.962	0.869	0.555
Neubildung, Darm	374	0.746	0.927	0.106
Neoplasie, Darm	374	0.746	0.927	0.106
Tumor, Prostata	88	0.840	0.954	0.040
Arithmetic Mean	238	0.882	0.904	0.382
		Keyword		
Information Need	*Quantity*	*Precision*	*Recall*	*LSA Precision at Recall*
Appendicitis	196	0.977	0.897	0.668
Colitis	140	0.720	0.828	0.560
Gastritis	567	0.986	0.932	0.654
Hepatitis	73	0.857	0.904	0.572
Adenokarzinom, Colon	146	0.810	0.205	0.703
Adenokarzinom, Dickdarm	146	0.883	0.623	0.650
Neubildung, Darm	374	0	0	0.833
Neoplasie, Darm	374	0.951	0.211	0.776
Tumor, Prostata	88	0.633	0.511	0.427
Arithmetic Mean	238	0.757	0.567	0.649
		IR Tool One		
Information Need	*Quantity*	*Precision*	*Recall*	*LSA Precision at Recall*
Appendicitis	196	0.639	0.964	0.271
Colitis	140	0.742	0.843	0.465
Gastritis	567	0.928	0.956	0.482
Hepatitis	73	0.404	0.945	0.336
Adenokarzinom, Colon	146	0.864	0.781	0.550
Adenokarzinom, Dickdarm	146	0.855	0.890	0.545
Neubildung, Darm	374	0.769	0.925	0.106
Neoplasie, Darm	374	0.767	0.925	0.106
Tumor, Prostata	88	0.840	0.773	0.223
Arithmetic Mean	238	0.756	0.889	0.343