A novel approach to semantic indexing combining ontology-based semantic weights and in-document concept co-occurrences.

Brian Jackson, Werner Ceusters
Language and Computing nv, Hazenakkerstraat 20, B9520 -Zonnegem, Belgium

Abstract:

Many approaches exist for ad-hoc information retrieval systems. These include key word searching, a variety of statistical models, and semantic indexing. A key issue for proper indexing is finding those concepts that are highly relevant in a given document. We have developed a novel approach to automatic semantic indexing based on a medico-linguistic ontology combined with a mathematical model. In this paper, we present empirical evidence that our method, contrary to a pure statistical approach, can be applied to very small documents, independent from any pre-existing corpus. At the other hand, it requires a large ontology that captures in detail a large portion of the semantics of a domain..

INTRODUCTION

Natural Language Understanding is considered one of the most complex problems in artificial intelligence. Up to now, a computer is not yet capable of really understanding the true meaning of ordinary human language. The necessary background knowledge is so extensive and complex that even given recent advances in the field, this knowledge cannot yet be fully computationally represented. However ! Under certain circumstances it is possible to have a computer understand natural language to a level that is sufficient for a specific purpose. Medical language, as a sub-language of ordinary human language [1], is a field that complies in an excellent way with the 'specific circumstances' required: a closed world with restricted domains and disciplines easily separated from each other, a relatively uniform terminology, and the availability of numerous descriptions (textbooks, classifications, …). In addition, there are many tasks to be performed by physicians in which medical natural language understanding applications can offer assistance such as clinical coding or document retrieval. Also automatic background procedures such as alert triggering based on clinical guidelines are possible today when the right technology is in place.

The problem

One basic requirement in any relevant natural language understanding application is to identify in a running text those "components that carry meaning". Second, it is important to assess how relevant these components are in the context of the entire document. Stated otherwise, the first deals with finding all the issues that are "touched upon" in the document (let us call that the "substances"), while the second concentrates more on the "topic(s)" of the document. We used specifically the phrasing "components that carry meaning", instead of "words", "terms" or "concepts" as these terms only make sense with respect to specific approaches. Statistic based systems that do not possess explicit domain knowledge, can only identify words or multi-word units in texts, and project these on implicitly constructed concepts that are mathematically justifiable, but that do not necessarily correspond with metaphysical reality. Such systems, intrinsically, are capable in finding those components that qualify as topic markers, but are poor in identifying all components. Concept- or ontology based systems on the other hand use explicitly defined concepts to which words, terms or phrases are attached as known grammaticalisations in a specific language. Concept-based approaches whose basic mechanism relies on phrase identification with subsequent concept matching, tend to identify many more components, but are less performant in finding the topics.

Both the substances and topics of a document are equally important because they relate to different information needs. Somebody interested in recent advances in liver cancer therapy will more likely find an answer in documents whose topic is "liver cancer therapy". However, if that person wants to write a review paper on liver cancer therapy, he will also find valuable information in papers dealing with radiotherapy or chemotherapy that discuss their application in liver cancer just briefly. Hence, the problem addressed in this paper is how to build a system that can deal effectively with both

information needs: being exhaustive in finding substances, and being able to classify these substances such that the highest ranked substances are those that define the topic of a document. As an additional requirement, it is desirable to develop a system which is independent from the size of the documents to analyse, and that does not rely on the pre-processing of large corpora.

The technology used

Central in our approach to automated medical natural language understanding is LinKBase®, a large scale medical ontology. LinKBase® contains approximately one million language-independent medical and general-purpose concepts, linked to natural language terms in several languages, including English [2, 3]. These concepts are linked together into a semantic network like structure using approximately 350 different link types for expressing formal relationships. These relationships are based on logics dealing with issues such as mereology and topology [4, 5], time and causality [6] and models for semantics driven natural language understanding [7, 8]. It is very important to note that in LinkBase® the formal subsumption relationship covers about 15% of the total number of relationships amongst concepts. As such, LinkBase® is a much richer structure than terminological systems in which term-relationships are expressed as strictly "narrower" or "broader". LinkBase®, or at least relevant extractions from it, is the driving force behind all our applications.

One such application is TeSSI®, designed for *Terminology-Supported Semantic Indexing*. In order to perform semantic indexing, TeSSI® first segments a document into individual words and phrases. It then matches words and phrases in the document to individual LinKBase® concepts via the associated terms. This step introduces ambiguity, since some concepts have terms in common. To resolve cases of ambiguity, TeSSI® uses domain knowledge from LinkBase® to identify which concept out of the set of concepts that are linked to a homonymous phrase best fits with the meaning of the surrounding terms in the document. Figure 1 shows the output of TeSSI® at the end of this stage. Identified words and phrases are underlined. Figure 2 shows for the same text, the results obtained by a generic statistics-based phrase extractor that does not enjoy the wealth of a rich domain ontology such as TeSSI®.

In the next step, TeSSI® uses the relationships between concepts identified in the document and the domain knowledge in LinkBase® to infer additional concepts which do not explicitly occur in the document. The end result of that process is a graph structure in which nodes correspond to concepts present (or infered to be implicitly present) in the document, and arcs to semantic relationships derived from the domain ontology or co-occurrence relationships derived from the position of terms in the document. The arcs are weighted according to semantic distance in LinkBase® and term proximity in the document. The nodes are weighted based on their occurrence in the document.

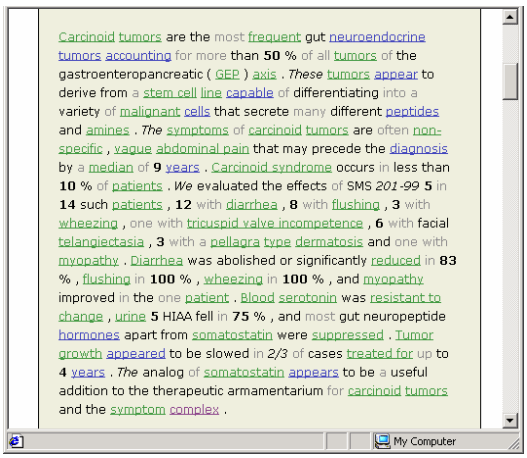

Fig. 1 : Phrase-identification results by TeSSI®

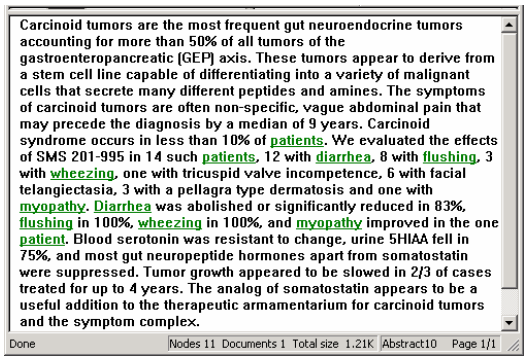

Fig. 2: Phrase-identification by a generic statistics based indexer.

Having identified all the medical (and many non-medical) concepts in a document, TeSSI® then ranks these concepts in order of their relevance to the document as a whole, hence identifying the topic(s). Relevance scores are on a scale of 0 to 100, with 100 representing the most relevant concept. To determine these scores, TeSSI® uses a constraint spreading activation algorithm on the constructed graph [9]. In this way, semantically related concepts "reinforce" each others' relevance rankings. The rationale for this algorithm stems from the observation that the concepts in any particular document will vary in their semantic independence from each other. For example, a document might contain one mention each of "heart failure," "aortic stenosis," and "headache." The first two of these concepts are clearly more closely related to each other than to the third. An indexing

system based entirely on term or concept frequency will treat these three concepts independently, thus assigning them all the same relevance. Yet intuitively, based on this limited description, the document has twice as many mentions of heart disease as of headache. TeSSI® takes advantage of its underlying medical ontology to more accurately represent this type of phenomenon.

The relevance ranking algorithm is nonlinear, and so the behavior cannot be described analytically. It is, however, important to characterize the behavior in order to normalize and optimize the rankings for incorporation into information retrieval systems and other applications.

METHODS

We used the OHSUMED corpus [10] for our testing. This corpus consists of approximately 350,000 Medline abstracts, along with 106 physician queries. In addition, the corpus contains relevance judgments of "definitely" and "probably" relevant for the abstract-query pairs, as determined by expert reviewers. From the queries, we selected one containing five distinct medical concepts, namely, "pancreas", "liver", "carcinoid", "treatment" and "research". In the corpus, there are 29 documents for which this query was judged "definitely" relevant. We processed all 29 documents using TeSSI®. We then constructed a set of larger documents by concatenating these to form 28 larger documents as follows: the two smallest documents, the three smallest, … the 28 smallest, and finally all 29 documents concatenated together, and we analyzed each of these concatenation products with TeSSI® as well. The concatenation of the documents was performed in order to be able to plot the performance of TeSSI® as a function of the length of the documents.

Of the five concepts in the query, "research" was not explicitly mentioned in any of the 29 abstracts, and so was excluded from further analysis. For each of the remaining four concepts in the query, and for each document, we determined the highest relevance ranking for that concept, its IS-A descendents according to LinkBase®, and its partonomy descendents. For example, if "liver" achieved a relevance score of 20, and "lobe of liver" received a score of 40, then for this document, we would use the score of 40 for our analysis. A typical information retrieval system will return a ranked list of documents in response to a user query. Ideally, the documents which best match the query should occur at the top of the list, those less well matched should occur at the bottom, and documents unrelated to the query should not show up at all. The position of a particular document on this ranked list is determined by the relevance of the query term(s) in this document, relative to the relevance of the query term(s) in all other documents.

For each of the four query concepts studied, we then calculated the relative rank of that concept in each document on a scale of 0 to 1. For example, if TeSSI® identified 50 concepts in a particular document, and determined that "liver" was the third most relevant concept of the 50, then the relative relevance of liver in that document would be (50-3+1)/50 = 0.96. If the query concept is not identified in the document at all, then the relative rank is zero.

For each of the four concepts, we then plotted three sets of values: the relative rank of the concept in each individual document (Fig. 3, triangle data points), the relative rank within the concatenated documents (Fig. 3, diamond data points), and the cumulative average of the relative ranks within the individual documents (Fig. 3, squared data points).

Finally, we processed nine documents by a pure statistical generic indexer which had no additional domain knowledge. This indexer uses intra-document word co-occurrence information to automatically derive meaningful concepts. We then compared the behaviour of TeSSI® with this system. The idea was not to show the superior behaviour of TeSSI® (this would be an unfair comparison on account of the absence of domain knowledge in that system), but rather to find out how such a system behaves with respect to document length. To that end, we analysed the number of substances (terms + phrases) recognised by TeSSI® for each of the nine documents (independent with respect to the relevance ranking) and compared these figures with the numbers of substances (called "nodes") in the statistic indexer. We calculated for both systems the ratio of found substances per word in the documents. This ratio does not take into account the number of individual words per meaningful phrase, and as such is no direct measure for recall. A document of 10 words, in which two phrases are found having a respective length of 2 and 4 words, would get a concept/word ratio of 2/10 = 0,2, whereas the percentage of "words explained by phrases", would be (2+4)/10 = 0,6. This approach frees us from analysing individual phrase-lengths, while it does not influence the comparison of the two systems as they are treated in the same way.

RESULTS

Phrase recognition

Table 1 shows the results of the phrase recognition phases of both TeSSI® and the pure statistical system. TeSSI®'s phrase recognition capabilities are clearly independent from the number of words in the documents, averaging around 0,351. For the statistical system, we found an increase in the number of nodes found as compared to document

length. The number of nodes found doubled from the smallest document in which nodes were found towards the largest document, reaching a level of 0,073. This is 5 times less than TeSSI®.
However, because of this upward trend, we performed an additional forecast calculation to find out whether or not the statistical system would be able to reach the same performance as TeSSI® when there would be "enough" words in the document. This forecast analysis shows that there is indeed a continuous increase, but that there is an asymptotic maximum towards 0,075.

Relevance ranking

The concept "carcinoid tumor" occurred in 27 of the 29 documents. The relative rank of carcinoid tumor in the individual documents averaged 0,75, whereas the relative rank within the concatenation of all 29 documents was 1,00. The concept "treatment" occurred in 21 of the 29 documents. The relative rank of that concept in the individual documents averaged 0,49, whereas the relative rank in the concatenation was 0,94. The concept "liver" occurred in 6 of the 29 documents. The relative rank of liver in the individual documents averaged 0,06, whereas the relative rank in the concatenation was 0,72. The concept "pancreas" occurred in 3 of the 29 documents. The relative rank of pancreas in the individual documents averaged 0,06, whereas the relative rank in the concatenation was 0,66.
The results are plotted in Fig. 3. Most indexing systems are based on frequency of a search term within a document. In such a system, the relative ranking of a concept within a concatenated document would be expected to be similar to the average relative ranking of the concept within the individual documents. TeSSI®, however, functions in such a way as to reinforce the relevance ranking of concepts based on their semantic similarity to other concepts in the document. As a result, the relative relevance of each query concept within the concatenated documents was substantially higher than the cumulative average relative relevance in the individual documents. This effect was seen with all four query concepts studied. The effect was greatest with liver and pancreas, which were the least common of the four concepts identified in this document set. These results are consistent with the original design of TeSSI®, namely, to compute relevance scores based on a combination of frequency of the index concept and the frequency of semantically related concepts. They also form empiric support for deriving the normalization function necessary for using TeSSI® as the foundation for an information retrieval system.

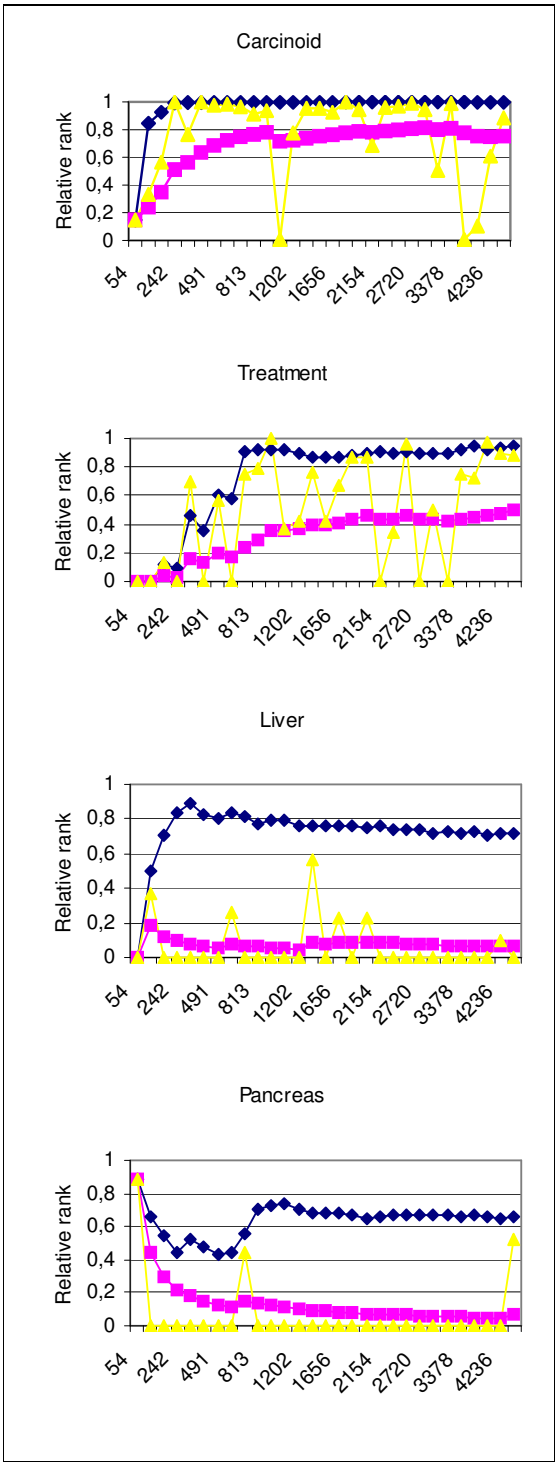

Fig. 3 : Relative relevance ranking of four concepts by TeSSI® in individual documents (triangles), concatenated documents (diamonds) and cumulative over individual documents (squares). The X-axis shows the number of words in the concatenated documents. The individual documents are increasingly ordered with respect to the number of words that they contain.

DISCUSSION

Term weighting is an important issue in document retrieval. The simplest and oldest term weighting measure is Term Frequency (tf), that just takes into account the number of times a term occures in a text [11]. It has been proven to improve recall, but specific attention must be given in removing terms that do not carry meaning, such as function words. Another frequently used measure is Inverse Document Frequency (idf), that takes into account term occurrence over a collection of texts [12], or the combination of both by simple multiplication [13].

In these older studies, "terms" actually meant "words". Because much more meaning can be attached to phrases than to individual words, studies have been conducted to see whether or not idf-related weights applied to phrases would positively influence recall and retrieval. Most were disappointing. For an overview, see [14]. As phrased by Arampatzis et al. : "*One explanation of why NLP has not had more successes in document retrieval is that it does not go far enough. First, the currently available NLP techniques suffer from lack of accuracy and efficiency, and second, there are doubts if syntactic structure is a good substitute for semantic content. The evidence so far suggests further investigation and better modeling*." [15]. Indeed, most studies concentrated on finding better statistics- or syntax- based phrase extraction techniques. In our view, more efforts should go into applying deep semantics. But even here, there is still no conclusive evidence that the use of large thesauri such as WordNet [16] can improve document retrieval as many studies contradict each other [17]. Even this is not a surprise to us. Indeed, a closer look on how thesauri or ontologies are used in these evaluations, reveals that in most cases, only hierarchical relationships are exploited, mostly because associative relationships are absent in the ontologies used. Our findings with TeSSI® suggest that these relationships are extremely important. When "pancreas" and "inflammation" occur together in a sentence, it is of little help to have an ontology that just represents "pancreatitis" as a more narrower term than "inflammation". It is only when "pancreatitis" is related to "pancreas" as well, that the co-occurrence of "pancreas" and "inflammation" in a sentence can be boosted by ontological evidence. Statistical systems can discover these kinds of relationships, but they require massive amounts of text and our study suggests that there is an upper limit.

More close to our developments comes the National Libray of Medicine's indexing initiative. However, features such as word sense disambiguation and full text processing (rather than only abstract processing) are work in progress as reported in [18, 19]. It would be interesting to compare the results of TeSSI® with that system as soon as these features are available. Because that system makes use of UMLS, and because UMLS only contains associative relationships at the level of its Semantic Network®, and not at the level of each individual concept, relevance ranking might not be as accurate.

CONCLUSION

We presented here an empiric characterization of a novel semantic indexing mechanism based on combining a very large domain ontology with in-document phrase co-occurrence. The results and graphs illustrate the manner in which relevance rankings are raised in the setting of semantically related concepts occurring in the same document. They also illustrate that this effect is seen at very low absolute concept frequencies, and without the need for having a large corpus, or large documents. On the other hand, it requires a very large domain ontology with a dense network structure.

References

[1] N. Sager, C. Friedman, M.S. Lyman, Medcial Language Processing: Computer Management of Narrative Data. Reading, MA: Addison - Wesley, 1987.

[2] W. Ceusters, P. Martens, C. Dhaen, B. Terzic, LinkFactory: an Advanced Formal Ontology Management System. Interactive Tools for Knowledge Capture Workshop, KCAP-2001, October 20, 2001, Victoria B.C., Canada (http://sern.ucalgary.ca/ksi/K-CAP/K-CAP2001/).

[3] F. Montyne, The importance of formal ontologies: a case study in occupational health. OES-SEO2001 International Workshop on Open Enterprise Solutions: Systems, Experiences, and Organizations, Rome, 14-15 September 2001 (http://cersi.luiss.it/oesseo2001/papers/28.pdf).

[4] B. Smith, Mereotopology: a theory of parts and boundaries, Data and Knowledge Engineering 20 (1996), 287-301.

[5] B. Smith, A.C. Varzi, Fiat and Bona Fide Boundaries, in Proc. COSIT-97, Springer-Verlag 1997, 103-119.

[6] F. Buekens, W. Ceusters, G. De Moor, The Explanatory Role of Events in Causal and Temporal Reasoning in Medicine, Met Inform Med 1993, 32: 274 - 278.

[7] W. Ceusters, F. Buekens, G. De Moor, A. Waagmeester, The distinction between linguistic and conceptual semantics in medical terminology and its implications for

NLP-based knowledge acquisition. Met Inform Med 1998; 37(4/5):327-33.

[8] J.A. Bateman. Ontology construction and natural language. In Proc. International Workshop on Formal Ontology. Padua, Italy, 1993, 83-93.

[9] J.A. Hendler, Marker-Passing over Microfeatures: Towards a Hybrid Symbolic/Connectionist Model. Cognitive Science 1989 (1) 79-106.

[10] W. R. Hersh, C. Buckley, T. J. Leone, D. H. Hickam, OHSUMED: An interactive retrieval evaluation and new large test collection for research. In Proceedings of the 17th Annual ACM SIGIR Conference (1994), 192-201.

[11] H.P. Luhn, A statistical approach to mechanized encoding and searching of literature information. IBM Journal of Research and Development 1(4), 1957, 307-319.

[12] K. Sparck Jones, A statistical interpretation of term specificity and its application in retrieval. Journal of Documentation 28 (1), 1972, 11-21.

[13] G. Salton, C.S. Yang, On the specification of term values in automatic indexing. Journal of Documentation 29 (4), 1973, 351-372.

[14] C. Jacquemin. What is the tree that we see through the window: a linguistic approach to windowing and term variation. Information Processing and Management, 32(4), 1996, 445-458.

[15] A. T. Arampatzis, Th.P. van der Weide, P. van Bommel, C.H.A. Koster, Linguistically-motivated Information Retrieval. Technical Report CSI-R9918, Dept. of Information Systems, Faculty of Mathematics and Computing Science, University of Nijmegen, September 1999. (http://www.cs.kun.nl/~avgerino/Avi_Arampatzis/publications/HTMLized/encyclop/).

[16] C. Fellbaum (ed), WordNet: an electronic lexical database. MIT Press, Boston, 1998.

[17] J. Gonzalo, F. Verdejo, I. Chugur, J. Cigarran, Indexing with WordNet synsets can improve text retrieval. Proceedings of the COLING/ACL Workshop on Usage of WordNet in Natural Language Processing systems, Montreal 1998. (http://sensei.ieec.uned.es/~ julio/colacl98.ps)

[18] A.R. Aronson, O. Bodenreider, H.F. Chang, S.M. Humphrey, J.G. Mork, S.J. Nelson, et al. The NLM indexing initiative. *Proc AMIA Symp* 2000(20 Suppl):17-21.

[19] A.R. Aronson. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. Proc AMIA Symp. 2001;17-21.