Text Mining Services to Support E-Research

Brian Rea, Sophia Ananiadou

National Centre for Text Mining, School of Computer Science, University of Manchester
{brian.rea, sophia.ananiadou} @manchester.ac.uk

Abstract

In recent years the developments and opportunities created for e-Science infrastructure have promised technological support for the ever growing area of text mining applications and services. The computationally expensive tools have previously only been usable on small scale systems but are now being developed for much larger scale tasks thanks to alternative models of processing and storage. In this paper, we provide an overview of a selection of the tools available at the National Centre for Text Mining, ways they can enhance document collections, improve on traditional search techniques and how complex combinations of the tools can interact to provide advanced solutions for researchers in the UK academic community. We finish with an account of two case studies currently being investigated; examining how these processes can be extended and customised to meet the needs of domains outside the boundaries of e-Science.

1. Introduction

Recent developments in grid technologies combined with the use of distributed text mining algorithms have led to a new era of document analysis. (Carroll 2005) Existing techniques can now be used on a practical scale and previously unfeasible tasks have suddenly become achievable. This is happening at the same time that information overload is reaching new strengths, with wide scale acceptance and use of web logs, wikis, news feeds and more recently grey literature through institutional repositories and virtual research environments. Traditional keyword based search is now barely sufficient to meet the needs of a researcher and therefore we must re-examine the models of information access that we have in place now, and will require in the near future. (Rice, McCreadie et al. 2001)

It is important to keep in mind that different groups of people will use text mining systems with different aims and objectives in mind. Flexible methods of interaction must therefore be made available within a system to enable users to discover appropriate and relevant documents, irrespective of the user's inherent knowledge, or lack thereof, of the area they are investigating. This can be achieved through an appreciation of the correlation of terms within documents, reducing the need to rely upon user knowledge of domain specific terminology or co-occurrence of traditional keywords.

Following these ideas, this paper examines work currently being undertaken at the National Centre for Text Mining[1] (NaCTeM) alongside future plans for the integration of tools, techniques and services into the e-Infrastructure for use by the UK academic community. We begin in Section 2, with an introduction to text mining and some of the key techniques we use for enriching and annotating document collections. This is followed in Section 3 by a more detailed examination of the techniques we are developing to enhance searching. Section 4 discusses the tools we can create through combination of these techniques. This section also introduces two case studies and future plans for extending these services out from firmly tested bio-medical domains to the social sciences.

2. Text Mining for Data Enrichment

Traditionally a reader would be able to examine a collection of documents and extract appropriate facts, gain insight into methodology and discover relationships between topics and concepts. This has not changed, but the sheer amount of new material being published in continually more specialized fields has meant that it has become impractical to examine all of the literature in close detail. The field of text mining begins to address this situation by assisting the reader through automated processing systems that are undeterred by the information explosion. In essence it follows the same process as the reader, discovering facts, patterns of information and relationships, but uses often quite different methods due to the

[1] http://www.nactem.ac.uk

lack of a natural insight and understanding of the text (Hearst 1999); (Ananiadou and McNaught 2006) (Hirschman, Park et al. 2002; Jensen, Saric et al. 2006). The benefit is that these new techniques can often find patterns or relationships buried in documents, which may have gone overlooked or otherwise been extremely difficult to discover. This section looks at some of the techniques involved in annotating the text as a precursor to analysis.

2.1 Part-of-Speech Annotation

The process of annotating documents with linguistic information is difficult as computers do not have the insight of the language models used by humans. This therefore needs to be built up in stages, starting with basic word classes or parts-of-speech as they are more commonly known, such as a noun, verb, preposition, etc. There are numerous algorithms for performing this analysis, all of which essentially resemble a lookup in a table, or prediction based on rules, prior experience or statistics. Part-of-speech (POS) tagging is often one of the first techniques used in language based text mining and is used to add appropriate linguistic knowledge to text in order to assist further analysis by other tools. When we know the lexical class of a word it is then much easier to predict relationships between this and neighbouring words. (Toutanova 2003; Tsuruoka, Tateishi et al. 2005; Yoshida 2007)

The first stage of this process involves tokenising the text by splitting it into a sequence of single word units and punctuation. This includes a resolution of the commonly occurring problems such as abbreviations, quotations, contractions or sentence boundaries, which can often cause errors with POS tagging algorithms. At this point it is possible to introduce linguistic stemming into the annotation, which predicts the base form of a word to assist in later analysis or searching. (Hull 1996)

In order to keep a high accuracy it is recommended that any software used is trained on sample annotated texts before use on new domains or different types of text. With this process being in the early stages of the whole text mining chain any errors at this stage may grow exponentially so it is important to have a POS tagger be as accurate as possible for the document collection you are working with. (Yakushiji, Miyao et al. 2005; Yoshida 2007)

2.2 Syntactic Annotation

Linguistic parsing is a method of predicting the grammatical analysis of a sentence using linguistic information provided by POS taggers. Shallow parsers identify only the main grammatical elements in a sentence, such as noun phrases and verb phrases, which are useful for most text mining tasks. Deep parsers, however, generate a complete representation of the grammatical structure of a sentence. This is of particular use when extracting information hidden within sets of clauses from complex sentences. (Yakushiji, Miyao et al. 2005)

Using parsing for syntactic annotation is a computation intensive task, though one that is essential for many of the more advanced tools. As an example consider the following sentence:

The MP discussed the policy with the ambassador.

Though this is quite a simple sentence there are a number of ways this could be parsed, all equally valid in terms of the grammars used. It could either denote that the MP was talking with the ambassador about the policy. An alternative would be that the MP was talking to a person or persons unknown about a policy for dealing with the ambassador. When clauses and conjunctions become involved, as they often do, there are often many more possibilities all of which must be discovered and considered before deciding upon the most appropriate option.

2.3 Term Discovery Techniques

The default option for most search engines is currently for key words in documents. This is often sufficient if you know precisely what you are looking for and the language used to describe it. As is often the case in modern research, subject specialization and technical jargon make this process more complex. Term discovery can be used on a document collection to identify multi-word units of significance, which we refer to as terms.[2] This is especially useful for documents that contain many complex multi-word terms, such as scientific research papers. (Frantzi and Ananiadou 1999); (Maynard 2000); (Spasic 2005); (Ananiadou and Nenadic 2006).

When this is combined with information retrieval algorithms it is possible to

[2] http://www.nactem.ac.uk/software/termine

automatically construct lists of related terms to a given search, ranked by a calculated importance. This can be particularly beneficial for browsing/searching within a domain the user is not familiar or where many variants[3] exist of individual terms, such as biosciences.

2.4 Named Entity Recognition

Named-entity recognition is similar to term discovery as it identifies items belonging to a given semantic category, however to gain this extra semantic information the techniques involved are different. Common examples of this include the names in a document, such as the names of people or organisations. Some systems are also able to recognise dates and expressions of time, quantities and associated units, percentages, and so on.

This method can use either manually defined templates or machine learning approaches upon annotated documents to identify the patterns that surround the entity under investigation.[4] These can then be used on larger collections of documents to identify new occurrences of named entities. (Kumar, De Beer et al. 2006); (Li and Liu 2005); (Morgan, Hirschman et al. 2004); (Eriksson, Franzen et al. 2002); (Zhou and al. 2004)

3. Text Mining to Improve Search

Now that the key techniques for enrichment have been established in section two it is useful to see how this can be used for improving search and access to documents. By looking at these alternative models it is possible to move beyond keyword searches to a more conceptual approach removing the need for knowledge of the exact language being used to describe information of interest in text.

3.1 Similarity Calculation

Many information retrieval (IR) systems use simple Boolean logic to find matches in the document collection. Though this can provide accurate results, there is no inherent concept of partial matching, either it matches the whole query, or it is not returned as a result. A different model of IR considers mapping each document into a term space, with each dimension representing an index term, and the distances representing some measure of contribution of that index term to the document. These measures range from simple frequency of occurrence or normalised frequency compared to common usage to manually designated weights. The vector space model offers greater flexibility in querying with partial matching automatically built in. (Baeza-Yates and Ribeiro 1999)

Any search in the vector space is essentially a similarity match between the query vector and the document vector. The similarity score then also acts as a ranking score for relevance giving a sorted list of the most similar documents to the users query. This similarity is generally calculated using the cosine of the angle between the two t-dimensional vectors, representing the document and the query. (Witten and Frank 1999) This is a very computationally intensive method and has been a problem in previous research, in that a similarity check would need to happen for every document in the system, every time a search is made. For small collections this is fine, but for larger collections such as Medline abstracts (~16 Million) this is taken outside of the realms of being practical at runtime. (MEDLINE 2004) This is especially the case when comparing the documents with each other i.e. 16 million document vectors compared with all other 16 million document vectors.

3.2 Dimensionality Reduction

Given that it is impractical to search the entire search space due to computational limitations it is essential that we reduce the term space to something more manageable. The difficult part of this is keeping the flexibility of the language whilst reducing the overall number of dimensions. One way to view this is to consider a concept space rather than the term space where a number of synonymous terms map directly to a single concept. This is possible with a technique called latent semantic indexing which uses the matrix theory of singular value decomposition to create a mapping matrix. Essentially when a query is performed the query is mapped onto this matrix, mapping the terms to their abstract concepts, this is then compared with the pre-mapped documents making the search space much smaller. (Witten and Frank 1999); (Homayouni, Heinrich et al. 2005)

The benefits of this technique include noise reduction and removal of a large proportion of redundancy in the document set, whilst

[3] http://www.nactem.ac.uk/software/acromine/
[4] http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/tagger/

increasing the overall efficiency with less run-time processing involved.

3.3 Online or Offline Processing

Each of the techniques discussed in this section takes a great deal of processing to produce the required results. Traditionally this has been fine for small collections of documents but has been a limiting factor in work with larger collections such as Medline. The Medline collection consists of around 16 million biomedical journal references and all but the most basic searches have been impractical until recently. In previous experiments (carried out on around 14 million references) it was calculated to have ~71 million sentences. If each sentence took one second to process, it was estimated to take around two years to process them all on a single machine. (Miyao, Ohta et al. 2006)

With recent developments at NaCTeM we have managed to speed up the processing involved by around a factor of 10. We are currently in the process of preparing NaCTeM tools to allow the use of computational resources available to the UK community, i.e. HPC and grid services. This will enable us to process gigabytes of raw text in a reasonable time by scaling over thousands of processors.

Once this has been done it is relatively easy to discover terms or use the other techniques on much smaller scale clusters. The benefit of having this done before opening the service is to allow for very fast access to results rather than the long delays one would expect to have for the amount of processing involved. It also reduces the need for duplicating effort and allows common queries to be cached. As we look at the services in section three, it will become apparent that not all processing can be done ahead of time as they rely upon the results of a given search, but even then the user will be working with a much smaller document collection making the processing more efficient, though some distribution of effort will be inevitable.

3.4 Search Expansion

Whilst keyword search has become a standard and offers useful results, it requires that the user and the author of any given document use the same terminology to describe the same concepts. This is clearly not always the case, even after taking into account spelling variants and mistakes. It is often the case that for any given concept there are multiple synonyms and unfortunately the same is true for a single terms mapping to multiple concepts, though this is rarer in practice. It is therefore impractical to expect a user to know all possible forms of a term and enter them all each time they wish to search for documents on a given subject.

Through a combination of term discovery and named entity recognition, alongside information from terminologies and ontologies, it is possible to construct a dictionary of the many normalised forms of terms. When this dictionary is built into the query parsing algorithms of an information retrieval system any query can be automatically expanded to include any other forms of query terms. The results of each of the expansions can then be transparently merged into a single result set. This can be processor intensive, but lends itself to parallel processing allowing for efficient run-time searching of otherwise quite time consuming tasks.

A further technique combines this approach with models of browsing for information. The search is expanded as above but before the results are presented, the user may choose to narrow down their collection using a generated list of significant terms. For example an expanded search on ‘heart attack’ could return a number of suggestions based upon term discovery across the result set including ‘cardiac surgery’, ‘risk factors’ and ‘blood pressure’. In essence this combines the best of both approaches, keeping the scope of the search as wide as possible whilst allowing the user to drill down in the results based upon their information need, rather than an assumed goal.[5]

3.5 Associative Search

This process relies upon the similarity between documents in the term space to discover sets of texts with similar content to a number of documents provided by the user. This is relatively straight forward for a single document as we already have the document similarities calculated across the whole collection, but when multiple documents are used as a ‘query’ this becomes more complicated.

For this to work we must examine the area of term space occupied by the query collection and use this to map back to the document space as shown in figure 1 below. Document similarity can be used for this, but in order to

[5] http://www.nactem.ac.uk/software/ctermine/

find the best matching set of documents it is essential that this be calculated differently by comparing the combined term vector of the query collection to the rest of the documents.

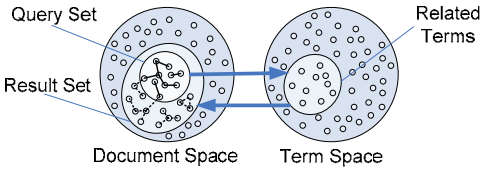

Fig. 1: Document-Term Mappings

Each dot in figure 1 represents a document in document space, or a term in term space. The lines connecting the dots represent a high level of similarity between the documents. As can be seen this method not only returns directly similar documents but may also identify new areas that could be of interest, which would have otherwise gone undiscovered.

4 Applications and Case Studies

4.1 Topic Visualisation

This service brings together some of the key ideas in document clustering, classification and summarisation to provide a graphical overview of a domain or result set from a query. Based around the browsing model of information access, this tool automatically identifies a number of important topics within a collection and clusters the documents around those topics. Finally the documents within that topic are then searched to identify a coherent theme which is then used as a human readable summary of the contents. (Osinski 2005)

The standard interface to this tool displays the title and abstracts of the documents alongside a list of clusters in a side panel for ease of viewing. The more advanced version currently under development displays the clusters graphically in document space and can identify areas of overlap between clusters. These overlapping areas contain sets of documents discussing the interrelationship between the clustered topics in the collection. This makes it very easy to quickly see the big picture of current research in a given area. (Mima 2006)[6]

[6] http://www.biz-model.t.u-tokyo.ac.jp/users/mima/en/index.html

The solutions provided by this service improve with accuracy when more documents are considered in the clustering process. This essentially filters out the noise in the documents leaving the important concepts behind. It must be stated however that this technique is limited due to its computational intensive nature and we are looking into alternative methods in order to make this a real solution for larger document collections.

4.2 Information Extraction

This is the process of automatically obtaining structured data from an unstructured natural language document. Often this involves defining the general form of the information that we are interested in as one or more templates, which are then used to guide the extraction process. Information extraction (IE) systems rely heavily on the data generated by natural language processing (NLP) systems. (Navarro, Martinez-Barco et al. 2005) (Andreasen, Jensen et al. 2002); (Alphonse 2004); (Huang, Chuang et al. 2005)

Information extraction includes the earlier stages, mentioned above, of term discovery, named entity recognition and pattern matching techniques, which identify and extract complex facts from documents. Such facts could be relationships between entities or events. (McNaught and Black 2006). Put simply, we can characterize information extraction as follows:

Take a natural language text from a document source, and extract the essential facts about one or more predefined fact types. Represent each fact as a template, whose slots are filled on the basis of what is found from the text. A template is a “form” which, when filled, conveys a fact. The form has labels, one of which says what type of fact it represents, and the others (the slots) identify the attributes that make up the fact. We will be interested here mainly in simple facts, and events. An example of a simple fact is:

Dr_X is a Programme Manager for Programme_Y at JISC

A template for the report of the death of a famous person could have slots for the name, the age, the place, the date, and the cause of death. Slots are typically filled by named entities or, in more complex representations, other facts. Increasingly, IE is being applied to scholarly papers, where the entities of

importance will include the objects of the domain, e.g. substances, artefacts, apparatus, locations, results and cited papers. [7]

4.3 Document Summarisation

Within manual summarisation techniques there are two main methods of summarisation, abstractive and extractive. Abstractive methods are often based upon an understanding of the content of a document and use this to rewrite a much shorter version of the original text. Extractive methods draw upon the perceived importance of sections of texts, to use only the most significant sentences to form the summary. As the result is based upon whole sentences or sections of text, it is much more likely to be readable than machine constructed abstracts. (Radev 1999)

Extractive approaches, whilst possible with current technology, face similar language based issues as other text mining tools such as synonymy and variations, but can equally be resolved in order to reduce redundancy in the output. Further to these issues, a summary of multiple documents is much harder to construct. There is a higher chance of repetition across potentially inconsistent language usage and multiple writing styles. With different authors who may hold contrasting views and contradictory information. (Goldstein 2000; Mani 2001)

This service is becoming increasingly more essential just to keep up to date with ongoing developments in a research area. Few people have the time to scrutinise everything published in a given area, simply because of the time it takes to read and consume the throughput of modern publishing models. As this rate of publication increases it will be necessary to use summarisation techniques to keep a detailed overview of developments and, in combination with other services, link back to anything of particular relevance or interest.

4.4 Case Studies

At NaCTeM we know it is important to keep abreast of the latest developments in the field and ensure that our technology is cutting edge. We also appreciate that all of this would be worthless if we did not balance it with user steering and feedback. We feel it is vital to develop tools that solve everyday problems facing researchers in the UK and as such actively consult with user communities to establish future potential collaborations that are mutually beneficial. Two such projects are described below, which take the core technologies of NaCTeM and apply them to real world challenges.

4.4.1 Systematic Reviews for Social Sciences

Working closely with the Evidence for Policy and Practice Information[8] (EPPI) Centre and the National Centre for e-Social Science[9] (NCeSS) the ASSERT[10] project is looking to extend NaCTeM technologies to support systematic reviews for the social sciences.

The project uses search expansion services to assist in creating a search strategy that is as complete within the area of interest as possible. It carries out these searches across multiple document repositories for extra coverage and then hands over to a screening phase for removing duplicate documents, identifying irrelevant studies and limiting the final result to a representative and complete set of texts within the investigation domain. At this stage a combination of advanced information extraction and summarisation techniques are used to locate evidence of interest to the final report before combining it into a readable form for the reviewer to analyse.

The ASSERT project has been designed to offer assistive technology rather than replace the role of the reviewer. By working alongside current workflows we hope to reduce the time and effort it takes to complete a review of what can be more than 20,000 papers and reports. Current areas of testing include ‘mental health rehabilitation’ and ‘walking and cycling schemes’. Both topics will show how text mining can assist in social sciences research and to enable transfer of e-Science techniques to e-Social Science applications.

4.4.2 BBC News Feeds

We have recently started a promising pilot project with the British Broadcasting Corporation and their research and technology group looking at how text mining can improve search and access of BBC online news feeds. The project is divided into two phases, the first looking at leveraging search expansion and

[7] http://nactem2.mc.man.ac.uk/medie

[8] http://eppi.ioe.ac.uk/cms/

[9] http://www.ncess.ac.uk/

[10] http://www.nactem.ac.uk/assert/

narrowing technology to improve traditional access to resources. The second phase looks at the benefits of using topic visualisation and clustering technology as an alternative approach.

We are looking to trial both phases of development on multiple user communities, ranging from the general public, media researchers and news teams. We hope to investigate the role they could play in different usage environments and suggest how social science researchers can benefit from such interfaces. As this project develops more information will become available on our website.

5. Conclusions

NaCTeM has made significant progress in its first three years by developing a strong set of core technologies and through provision of scalable text mining services. Our next development stage focuses on the consolidation of these tools into practical, user driven applications and further services to assist in solving key challenges facing the UK academic community. As part of this process we intend to extend and customise our tools to assist in new areas of the sciences, social sciences and within the arts and humanities domains.

With a sustained growth in the number of publications and increased access becoming available through novel business models, new and original methods of document discovery need to be investigated to ensure continued access to relevant, appropriate and wide-ranging resources. Further to this, full paper processing is necessary for the discovery of new types of evidence from literature.

Text mining developments of this scope and scale have been previously considered impractical, but recent advances in grid and HPC technology are allowing distributed document resources to become a reality. In light of this a number of issues previously not considered are becoming apparent, such as intellectual property rights on derived data, which need to be addressed at this initial stage before any potentially dangerous precedents are set.

The specifics of developing text mining applications offer some potential challenges in terms of balancing generality and specificity with accuracy and speed, interface usability with functional complexity, and ensuring interoperability in a field with few standards. For example, text mining tools are increasingly released as open source but in order to be interoperable common infrastructure and annotation schemes have to be adopted. In order for any project to be successful and sustainable all of these issues must be considered carefully at the start of the project and throughout its lifetime.

This paper focuses upon a single aspect of the National Centre for Text Mining that of information retrieval based systems. The NaCTeM website holds more information on our other research activities including advanced information extraction and relationship mining, semi-automated ontology construction, analysis of sentiment and opinion in text and citation studies.

6. References

Alphonse, E. e. a. (2004). Event-based Information Extraction for the biomedical domain: the Caderige project. Workshop on Natural language Processing in Biomedicine
and its Applications (JNLPBA).

Ananiadou, S. and J. McNaught (2006). Text mining for biology and biomedicine. Boston, Mass.; London, Artech House.

Ananiadou, S. and G. Nenadic (2006). Automatic Terminology Management in Biomedicine. Text Mining for Biology and Biomedicine. S. Ananiadou and J. McNaught, Artech House Books.

Andreasen, T., P. A. Jensen, et al. (2002). Ontological Extraction of Content for Text Querying. Applications of natural language to information systems; Natural language processing and information systems, NLDB 2002, Stockholm, New York.

Baeza-Yates, R. and B. d. A. j. N. Ribeiro (1999). Modern information retrieval. Reading, Mass., Addison-Wesley Longman.

Carroll, J., Evans, R., and Klein, E. (2005). Supporting Text Mining for e-Science: the challenges for Grid-enabled Natural Language Processing. UK e-Science All Hands Meeting, Notingham, UK.

Eriksson, G., K. Franzen, et al. (2002). Exploiting Syntax when Detecting Protein Names in Text. Proceedings of Workshop on Natural Language Processing in Biomedical Applications - NLPBA 2002, Nicosia, Cyprus.

Frantzi, K. and S. Ananiadou (1999). "The C-value/NC-value Domain Independent Method for Multiword Term Extraction." Journal of Natural Language Processing **6**(3): 145-180.

Goldstein, J., Mittal, V., et Al (2000). "Multi-document summarization by sentence extraction." NAACL-ANLP 2000 Workshop on Automatic Summarization **4**: 40-48.

Hearst, M. (1999). Untangling Text Data Mining. Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL 1999)**:** 3-10.

Hirschman, L., J. C. Park, et al. (2002). "Accomplishments and challenges in literature data mining for biology." Bioinformatics **18**(12): 1553-1561.

Homayouni, R., K. Heinrich, et al. (2005). "Gene clustering by Latent Semantic Indexing of MEDLINE abstracts." Bioinformatics **21**(1): 104-115.

Huang, C. C., S. L. Chuang, et al. (2005). "Categorizing Unknown Text Segments for Information Extraction Using a Search Result Mining Approach." Lecture Notes in Computer Science(3248): 576-586.

Hull, D. A. (1996). "Stemming algorithms: A case study for detailed evaluation." Journal of the American Society for Information Science **47**(1): 70-84.

Jensen, L., J. Saric, et al. (2006). "Literature mining for the biologist: from information retrieval to biological discovery." Nature Reviews, Genetics **7**(February): 119-129.

Kumar, N., J. De Beer, et al. (2006). "Evaluation of Information Retrieval and Text Mining Tools on Automatic Named Entity Extraction." Lecture Notes in Computer Science: 666-667.

Li, X. and B. Liu (2005). Mining Community Structure of Named Entities from Free Text. Information & knowledge management; CIKM 2005, Bremen, Germany, Acm.

Mani, I. (2001). Automatic summarization. Amsterdam; Philadelphia, J. Benjamins Pub. Co.

Maynard, D., Ananiadou, S. (2000). Identifying terms by their family and friends. COLING, Luxembourg.

McNaught, J. and W. Black (2006). Information Extraction. Text Mining for Biology and Biomedicine. S. Ananiadou and J. McNaught, Artech house.

MEDLINE. (2004). "MEDLINE." from http://www.ncbi.nlm.nih.gov/PubMed/.

Mima, H., Ananiadou, S. & Katsushima, M. (2006). "Terminology-based Knowledge Mining for New Knowledge Discovery." ACM Transactions on Asian Language Information Processing, **5**(1).

Miyao, Y., T. Ohta, et al. (2006). Semantic Retrieval for the Accurate Identification of Relational Concepts in Massive Textbases. Coling/ACL, Sydney, Australia, Association for Computational Linguistics.

Morgan, A. A., L. Hirschman, et al. (2004). "Gene Name Identification and Normalization using a Model Organism Database." Journal of Biomedical Informatics **37**: 396-410.

Osinski, S., Weiss, D. (2005). "A Concept-Driven Algorithm for Clustering Search Results." IEEE Intelligent Systems **20**.

Radev, D., R., McKeown, K. (1999). "Generating natural language summaries from multiple on-line sources." Computational Linguistics **24**: 469-500.

Rice, R. E., M. McCreadie, et al. (2001). Accessing and browsing information and communication. Cambridge, Mass.; London, MIT Press.

Spasic, I., Ananiadou, S. & Tsujii, J. (2005). "MaSTerClass: a case-based reasoning system for the classification of biomedical terms." Bioinformatics.

Toutanova, K., Klein, D., Manning, C., Singer, Y. (2003). "Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network." 467-474.

Tsuruoka, Y., Y. Tateishi, et al. (2005). "Developing a robust part-of-speech tagger for biomedical text." Advances in Informatics, Proceedings **3746**: 382-392.

Witten, I. H. and E. Frank (1999). Tools for data mining: practical machine learning tools and techniques with Java implementations. San Francisco, Calif., London: Morgan Kaufmann; Booth-Clibborn.

Yakushiji, A., Y. T. Miyao, and, et al. (2005). Biomedical information extraction with predicate-argument structure patterns. First International Symposium on Semantic Mining in Biomedicine.

Yoshida, K. (2007). "Ambiguous Part-of-Speech Tagging for Improving Accuracy and Domain Portability of Syntactic Parsers." Proceedings of the Twentieth International Joint Conference on Artificial Intelligence.

Zhou, G. and e. al. (2004). "Recognizing Names in Biomedical Texts: A Machine Learning Approach." Bioinformatics **20**(4): 1178--1190.