RETRIEVING HIERARCHICAL TEXT STRUCTURE FROM TYPESCRIPT SCIENTIFIC ARTICLES – A REQUIREMENT FOR E-SCIENCE TEXT MINING

Ms. T. Tamilselvi
Computer Science and Engineering,
Jayam College of Engineering and Technology,
Dharmapuri, Tamilnadu, India

Dr. G. Tholkappia Arasu
Principal,
AVS Engineering College,
Salem, Tamilnadu, India

Abstract*— Despite the growth and development of the web in scientific publishing, there remain significant obstacles to the application of computer based text processing technologies. One obvious obstacle is the relative paucity of freely and publicly available full-text articles. Such obstacles have resulted in a large concentration of text processing research on the relatively small amount of suitable material that is currently available, notably MedLine abstracts. In this paper, we discuss a processing framework (PTX) for scientific documents guided by two main principles. We start from the de facto position that most published material is available in PDF, a layout or document appearance format. For text processing, the (hierarchical) structure of the text is required. Secondly, we believe that the most likely users of scientific text processing will be scientists exploring literature within a particular specialism. Consequently, the framework can and should exploit (in a modular fashion) knowledge about that specific literature. The framework is being developed in the context of two E-Science projects: FlySlip and CitRAZ. The former is developing tools to aid human database curation of Drosophila genetics literature. The latter is combining Argumentative Zoning with citation information in order to help improve both citation indexing and text summarisation.

Keywords—Hierarchical, FlySlip, CitRAZ.

I. INTRODUCTION

Despite the growth and development of the web in scientific publishing, there remain significant obstacles to the application of computer based text processing technologies. One obvious obstacle is the relative paucity of freely and publicly available full-text articles. Indeed, the reluctance of traditional scientific publishers to embrace an open access philosophy has led to the development of whole new journals which are both devoted to open-access and managed by teams from outside traditional publishing. However these remain small players in the market. Another obstacle is that material, especially archive material, is often available only in document formats such as pdf which are primarily oriented to visual display. The obtacles have resulted in a large concentration of text processing research on the relatively small amount of suitable material that is currently available, notably MedLine abstracts. In this paper, we discuss a processing framework (PTX) for scientific documents guided by two main principles. We start from the de facto position that most published material is available in PDF, a layout or document appearance format. For text processing, the (hierarchical) structure of the text is required. Secondly, we believe that the most likely users of text processing for scientific articles will be scientists exploring literature within a particular specialism. Consequently, the framework can and should exploit (in a modular fashion) specialism-specific knowledge. Entirely general purpose “pdf to html/xml/word...” converters generally perform poorly on scientific texts precisely because they lack this sort of knowledge. The framework is being developed in the context of two E-Science projects: FlySlip [Briscoe et al 2005] and CitRAZ [Teufel 2004]. In the former we are using an archive of scientific articles forming part of the FlyBase bibliography of publications concerned with the biology and genetics of Drosophila [Drysdale et al 2005]; in the latter an archive of ACL computational linguistics papers is being processed by means of Argumentative Zoning [Teufel 1999] with special attention to citations and references .

The output of the framework is a hierarchical document structure that divides text into sections such as “title”, “authors”, “abstract”,, “main body”, “references”. Sections can also subdivide. For example, scientific articles often have particular named divisions such as “Materials and methods”. The lower levels of structure are paragraphs, sentences and words. Parts of the structure can also be stylistically marked for example as italics, bold and super and sub-scripts.

In section 2 we motivate our work, re-highlighting the distinction between textual layout (e.g. pages, columns, lines) and hierarchical layout (chapters, section, paragraphs), and demonstrating the importance of hierarchical text structure for language processing. In section 3 we describe the general PTX

ISSN 2348 – 9928
Doi:01.0401/ijaict.2014.07.32 Published on 05 (5) 2017

framework for recovering text structure including the journal specific modules. The framework consists of optical recognition over pdf (generating page layout xml) followed by a journal style processor for recovering text structure. We compare the method to several other document processing frameworks, including bioRAT [Corney et al 2004] and Mars [Thoma 2001]. In section 4, we evaluate our framework and report results. At this moment, using our most well-developed journal template on a previously unseen test set of 28 documents (from Journal of Biological Chemistry) containing 2296 text structure tags (paragraph level and above), we obtain 95.9% precision, 95.5% recall (95.7% harmonic F-measure). This was achieved from a development set of 8 documents (with very similar accuracy figures). Error analysis indicates most remaining problems arise from optical recognition errors which may be able to be compensated for by subsequent processing. We also report results for reference processing, stylistic recognition and Greek character recognition. In section 5, we discuss future directions for the work and draw summary conclusions

II. DOCUMENT STRUCTURE

2.1 Hierarchical Document Structure

Hierarchical document structure can be distinguished from its presentational structure. Presentational structure describes the appearance of the document using concepts such as page, (numbered) lines, font changes, indentation and so forth. Hierarchical document structure divides text into logical sections such as “title”, “authors”, “abstract”, “main body” and “references”. Sections can also subdivide. For example, scientific articles often have particular named divisions such as “Materials and methods”. The lower levels of this linguistic structure may be taken to be paragraphs, sentences and words. The traditional publishing process stresses presentational structure to ensure overall document consistency and visual attractiveness. For example, the submission guidelines for this All Hands Conference stipulate some structural aspects: that the paper shall have a title, an abstract and show the names and affiliations of authors; otherwise, the guidelines are presentational, stating the required font types and sizes, the required spacing beyond different elements and so forth. This ensures the published proceedings have a uniform appearance. Indeed the only supplied electronic aid is an example document in a proprietary format (Microsoft Word) which instantiates the guidelines. The submission format, as in most scientific publishing is essentially pdf, another at least semi-proprietary format. The result is that computer based analysis of the supplied documents becomes a more difficult task. Enhanced search over the proceedings and even simple re-presentation of the Conference proceedings in a different style or for a different medium becomes a much more significant task

The hierarchical structure is important for us for a number of reasons. First, our objective is to apply natural language processing techniques which can a) discern structures over words (such as noun phrases, verb phrases and attachments) and b) can make use of information such as which part of the document a use of a phrase appears in. In some information processing tasks, recovery of these structures may not be necessary. For example, if the processing task is recovery of whole documents (Information Retrieval) then it remains an open question whether NLP techniques can add value to word based metrics (such as string equivalence, word-stem equivalence, words plus their relations in an ontology) which can be computed without requiring any higher level structures such as the sentences, paragraphs or sections that they appear in [Lewis & Sparck -Jones 1996].

Information Extraction however typically does require at least part of the sentence that words appear in. A “standard” extractor will consist of some sort of pattern matching over structures of words in a sentence built by a partial parser. Such a parser might determine, for example, that phrases denoting a particular gene and protein are subject and object respectively of a verb denoting the “expresses” relationship. In the FlySlip project we are applying the RASP [Briscoe and Carroll 2002] statistical parser to the full text of genetics articles. One use of this is to build chains of co-reference between mentions of genes in a text.

For our gene named entity recognizer, we are building both a dictionary based tagger and an HMM. Whilst the latter model only uses relatively local context in its decision-making, it is clearly desirable to train it with good contexts and not those resulting from poor recognition of sentence boundaries.

Higher level structures are essential for indicating argumentative zones in a document, for example that the first two paragraphs re-state known facts and give background but that the third outlines a new hypothesis. Intelligent search can also be helped by information about document structure. Patterns that match within titles and headings may be much more useful to highlight for readers (including research scientists and others e.g. database curators) than those that match only within footnotes or reference lists.

In the FlySlip project, we also wish to re-present a document to database curators in a linguistically enhanced interactive document processing environment which links the elements within it directly to the database of recorded facts that they are creating from it. In this way, the source material and the curated material provide two views of the same data for the curator. Our primary purpose is to aid the process of database

ISSN 2348 – 9928
Doi:01.0401/ijaict.2014.07.32 Published on 05 (5) 2017

curation itself but clearly there are also interesting possibilities in extending the approach so that database users themselves can use this sort of linkage.

The CitRAZ project employs information about citations and references in order to improve summarisation and create intelligent citation indexes. Citations are being classified based on their linguistic context and on their position in the text, similarly to the way in which rhetorical relations within the text are identified in the process of Argumentative Zoning [Teufel 1999]. For example, two current classes (the precise set remains under development) are based on and weakness, as in "The methodology was first proposed in [3]", or "This data proves problematical for [7]". As a prerequisite for this kind of processing, citation and reference information needs to be extracted from the text in a manner which relies on its hierarchical structure. For example, the algorithm for discerning the function of a citation depends in part on what part of the text it appears in.

Due to the different existing standards for specifying bibliographic information in reference lists, which often vary even within a single paper, marking up this information is not an entirely trivial task. We therefore employ a dedicated parser to process the "references" section produced by PTX. Previous experience with other text extraction programs was less successful mainly because of the lack of accessible accurate typographical information. The reference parser often had to determine reference boundaries based on textual considerations alone. By employing typographical information, PTX can detect reference boundaries, thus improving the accuracy of the reference parsing and preventing errors from being propagated across references. Direct access to typographic information is also important for processing citations in running text, for instance in the case of footnotes, which need to be accurately detected and mapped to the right location in the text (thus requiring superscript information) in order to supply the full context for citations which occur in footnotes or in their vicinity.

2.2 Text stylistics

Although text style (italics, bold, super and sub-scripts) are part of document appearance, they are also of importance to us for two reasons. First, style often functions as an indicator of structure. A single line paragraph in italics, for example, often functions as a heading. There may or may not also be graphical clues such as centered text. Secondly, we are interested in exploiting certain reasonably well-adhered to conventions in the genetics literature – for example, that a standard gene abbreviation is italicized; that superscript suffixes may indicate alleles. Standard text extraction technology often performs poorly on these sorts of features of text. OCR appears to function reasonably well.

OCR also performs much more adequately for some foreign character-sets such as Greek characters. This is particularly important for gene named entity recognition in genetics.

III. PTX – A PROCESSING FRAMEWORK

3.1 The general framework

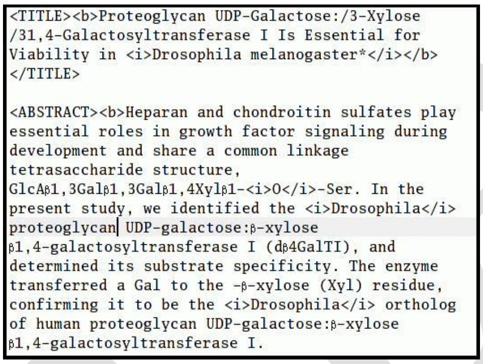
The general framework consists in three main processing stages. First, an OCR system (currently OmniPage Pro 14; we hope to make the system independent of this particular engine however) is used to recognize character-level information from input PDF documents. The output of OCR is an XML document encoding layout information page-by- page. Within pages, there are graphical regions or zones (detected using auto-zoning, i.e. without pre -prepared document or style specific templates) which in turn contain paragraphs and then lines and then characters. Zones may also include tables of information and images. Almost all elements may have various attributes including for example positioning information, and stylistic markers such as "non-bold superscript". The second stage is a generic processing stage that a) filters some of the OCR output b) summarizes it where appropriate (e.g. generating words from characters) c) detects and corrects certain characteristic OCR errors in its zone boundaries. The result is "Intermediate XML" or IXML.

The third stage is the application of a journal specific template. A template consists of pieces of program code (Perl, in fact) which are executed during an essentially top-down left to right traversal of the parse tree of the IXML document structure. For example, one can attach program code to the "zone template" which is executed when an opening zone tag is encountered. Different code is executed upon encountering an open paragraph tag. The code can read information about the object currently being processed (the zone, the paragraph, the line and so forth) and also any information stored away during processing so far. For example, in many journals, as in the style file for this conference paper, a change of zone is a reliable indicator of a change from the abstract to the main body text – simply because the abstract width covers the whole page, whereas main body text is in two-column format. To encode this in PTX requires not just knowing that we have a change of zone (information from the IXML) but that PTX currently believes it is processing the abstract. In other styles, the change from abstract to main text may not be indicated by a change in layout but only by a stylistic cue such as a move from bold to non-bold characters and possibly only by the contents of a text line (such as "1. Introduction").

ISSN 2348 – 9928
Doi:01.0401/ijaict.2014.07.32 Published on 05 (5) 2017

The value of PTX lies in its making the right sort of information available at the right points so that different sorts of rules like this can be expressed. However, we do not pretend to have anything like the last word on this and it is clear to us from some of our more Byzantine templates that the right level of abstraction is clearly missing at points. PTX also functions partly as a correction mechanism for OCR errors. For example, OCR output commonly fails to spot a small region of 2 column output at the foot of a page dominated by a picture and a full-width long caption. It also not uncommonly treats a full-width page heading and the second column of two column text as a single region. These errors can often be spotted and corrected by inspection of graphical co-ordinates.

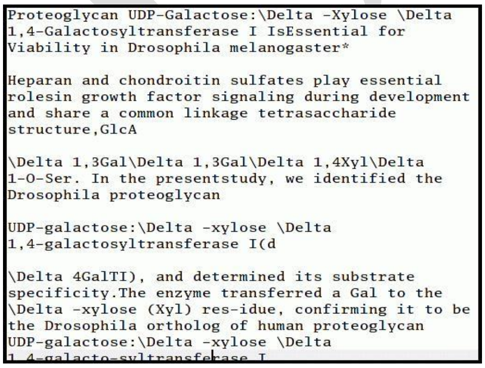
PTX is therefore essentially a programming framework and its utility should be assessed by how well it facilitates the task in hand: extracting document structure from layout structure, and how well it helps writers, readers and program maintainers. As an example Figures 1 and 2 below show an example PTX output for a title and the beginning of an abstract and, for comparison, how one rather more basic text extraction tool treats the same text. The PTX output, whilst not perfect (the Greek beta in the title is mis- recognized; but otherwise is correct even down to the italicized zero in the suffix 1-0-Ser) represents a substantial improvement.



```
<TITLE><b>Proteoglycan UDP-Galactose:/3-Xylose
/31,4-Galactosyltransferase I Is Essential for
Viability in <i>Drosophila melanogaster*</i></b>
</TITLE>

<ABSTRACT><b>Heparan and chondroitin sulfates play
essential roles in growth factor signaling during
development and share a common linkage
tetrasaccharide structure,
GlcAβ1,3Galβ1,3Galβ1,4Xylβ1-<i>O</i>-Ser. In the
present study, we identified the <i>Drosophila</i>
proteoglycan UDP-galactose:β-xylose
β1,4-galactosyltransferase I (dβ4GalTI), and
determined its substrate specificity. The enzyme
transferred a Gal to the -β-xylose (Xyl) residue,
confirming it to be the <i>Drosophila</i> ortholog
of human proteoglycan UDP-galactose:β-xylose
β1,4-galactosyltransferase I.
```

Fig 1: Example PTX output



```
Proteoglycan UDP-Galactose:\Delta -Xylose \Delta
1,4-Galactosyltransferase I IsEssential for
Viability in Drosophila melanogaster*

Heparan and chondroitin sulfates play essential
rolesin growth factor signaling during development
and share a common linkage tetrasaccharide
structure,GlcA

\Delta 1,3Gal\Delta 1,3Gal\Delta 1,4Xyl\Delta
1-O-Ser. In the presentstudy, we identified the
Drosophila proteoglycan

UDP-galactose:\Delta -xylose \Delta
1,4-galactosyltransferase I(d

\Delta 4GalTI), and determined its substrate
specificity.The enzyme transferred a Gal to the
\Delta -xylose (Xyl) res-idue, confirming it to be
the Drosophila ortholog of human proteoglycan
UDP-galactose:\Delta -xylose \Delta
1,4-galacto-syltransferase I
```

Fig 2: Unprocessed ps2ascii output

3.2 Related work

BioRAT is an Information Extraction engine designed to enable researchers to find research papers, but can also read them itself and extract key facts from them for display. The interface searches (principally via PubMed) for pdf documents, downloads them, converts them to text and then applies Information Extraction patterns (using the GATE architecture [Cunningham et al 2002]) to find items of interest. The function of the system differs from that of FlySlip in that our objective is partly to re-present the document to database curators to aid them in their work and this function may best be served by highlighting various facets of text (genes of interest, phrases, sentences or paragraphs of interest) rather than directly placing candidate facts of the form “A is related by B to C” in a table. BioRat is principally aimed at research scientists who may navigate their topic using extracted facts only or who may, when something of interest arises, revert to the originating source. There may also be an interesting difference in the optimal trade-off in precision and recall between database curators and scientific researchers.

Database curators are likely to be more concerned not to miss facts expressed in papers (i.e. prefer higher recall) whereas a scientific researcher may be more concerned at the possibility of being overwhelmed by too many false positives. We anticipate that the FlyBase database curators will remain the main Information Extractors themselves and that they will use FlySlip system outputs as potentially insecure evidence in their own human extraction process. For PDF processing, BioRat initially used the standard unix tool ps2ascii, and subsequently an open source version of a commercial tool (JPedal) which no longer appears to include a text extraction facility in its freely downloadable version. Some of the advantages of starting with an OCR tool are that, apart from general reliability (tools that attempt to extract text from postscript and pdf are often reported to be fragile), the output simply includes much more information such as font sizes, italicization, boldness and positional information. This output enables much more reliable detection of structural components such as abstracts and headings.

Furthermore, we expect to be able to use features such as italicization to help us in some of our domain specific tasks such as gene name spotting. MEDLINE’s Mars system for extracting bibliographic references similarly uses commercial OCR followed by a filtering, analysis and correction mechanism. [Mathiak and Eckstein 2004] also report a possible investigation into using OCR instead of current attempts to specialize an open source pdf text extractor.

ISSN 2348 – 9928
Doi:01.0401/ijaict.2014.07.32 Published on 05 (5) 2017

IV. EVALUATION

To evaluate our processing we undertook a number of small experiments designed to test different features of interest: in structure discovery and in reference delimitation and in Greek character recognition.

For structure discovery, we built a Gold standard of texts marked up in xml with elements from the following set: TITLE, AUTHORS, ABSTRACT, HEADING and P. P (for paragraph) represents logical or "linguistic" paragraphs containing complete sentences. Therefore, a sentence which continues from the bottom of one column to the top of the next, or from one page to the next, belongs only to one logical paragraph even if it is graphically split into several. Our Gold standard does not currently contain tables, captions or images. Also, the mark-up has no subdivisions. For example, our current purposes do not require us to individuate authors within an author list. At test-time we take a PDF document, run it through OCR and then through PTX with the appropriate journal style module. This generates XML with the above mentioned elements. Then we calculate precision, recall and harmonic F-Score over the opening and closing tags with respect to the Gold standard. An opening tag in the test set is a true positive if the (up to) first four words following it match one in the Gold Standard. Similarly a closing tag is a true positive (tp) if the (up to) last four words preceding it match. Otherwise a tag in the test-set is a false positive (fp) . Tags in the Gold standard not in the test -set are false negatives (fn). Precision (P) is tp/(tp+fp); Recall is tp/(tp+fn); F is 2PR/(R+P). The measuring process means that it is possible that scores at the structural level could be good, even though recognition within the structure is awful. In the worst case, every character except those in the four words beginning and finishing elements could be misrecognized without impairing our score. We have only found one circumstance of a structural error not being reported by our measure. In a case where OCR zoning
Misconstrues two columns as one, the beginnings and endings of paragraphs can be correct even though many (possibly all) sentences within the paragraph are incorrect. Fortunately these cases are quite rare. The measure is also somewhat insensitive to the amount of material (say, measured by numbers of sentences or words) successfully recognized. One large paragraph successfully recognized counts for no more than one small one.

For our most developed journal template, built using a development set of 8 papers from 2004, we achieved precision, recall and F-Scores of 97.5, 96.2 and 96.8 respectively on a test set of 18 further papers from 2004. (Errors rates on the development set are similar). We then selected a further 10 papers from 2003 and achieved 93.0, 94.2 and 93.6 respectively. The slightly increased error rate appears to be due to OCR errors in the second test set. For example, OCR misinterpreted one table in one paper as a sequence of text paragraphs leading to very low precision for that paper. On a less well developed template for a different journal we currently achieve precision, recall and F-Scores of 73.9, 90.4 and 81.3. The lower precision again reflects mostly a poorer handling of what is actually tabular data in the paper together with an evident OCR font recognition problem for citation strings (which differed from main text but were not "foreign characters"). Development of this template represents approximately one day's effort for an "expert" developer with previous experience of this work. There was also one further paper for which the system failed on account of a completely different format. It is not yet clear to us whether a back-off mechanism for generating poorer quality output rather than no output is worthwhile for our users. There may be a threshold of quality below which attempted use of the complete system could be pointless.

For reference lists in our ACL anthology, we examined a sample of 9 documents containing a total of 194 references. 190 references were correctly delimited in total. OCR alone would have correctly delimited 167. Subsequent processing by PTX therefore improved performance from 86% to 97.8%.

We also undertook a small evaluation of stylistic recognition and Greek character recognition success by comparing the characters found by OCR with characters present in html versions of papers that were available to us. We examined a small sample of 4 papers and found 98 correct recognitions out of 134 Greek characters, a precision rate of 73.1%. Of course many incorrect t recognitions were predictable: y for gamma, a for alpha and so forth and we expect to be able to work around these errors. Our current impression is that the results are clearly usable. There were very few recall errors where a Greek character was interpreted where none was present except for the particular journal style mentioned above in which citation occurred in a different (English) font from the main text. More complex post-OCR correction mechanisms have been proposed [Le et al 2002] should our initial optimism be misplaced. Superscript recognition was excellent with 123 correct recognitions out of 126 in the target set and no false positives.

V. FUTURE WORK

Within CitRAZ and FlySlip we are currently processing thousands of papers from the ACL and FlyBase anthologies. Papers are pipelined through the OCR software into PTX, where journal style-specific templates enable reasonably reliable extraction of the text and its hierarchical structure.

ISSN 2348 – 9928
Doi:01.0401/ijaict.2014.07.32 Published on 05 (5) 2017

PTX is designed to be modular in that different style templates can be easily plugged in, assuming one knows the journal of the input pdf document. (This is not always sufficient information however – journals sometimes have special styles for certain types of article such as review articles or fast-track articles). However, the programming of new templates is not itself a trivial task. Whereas the marginal effort does indeed decrease with acquired experience, it still seems desirable to minimise or ideally even eliminate the manual work involved in this process. One possibility is simply to re-factor the templates so that useful common parts of processing can be more easily re-used. A second possibility is to abstract out certain features of templates (for example, the graphical position which indicates a likely footnote occurrence rather than main text continuation). Some of these features may be suitable for employment of a Machine Learning paradigm.

VI. CONCLUSIONS

We have described PTX, a document processing framework designed for retrieving text, document structure and certain text stylistics from the currently most prevalent format of online scientific literature. The system includes both generic components (e.g. OCR) and some journal specific components. We are highly encouraged that the level of performance it gives will be sufficient to support the natural language processing techniques we wish to employ in two different specialisms within scientific literature: genetics and computational linguistics. We have also explained the project contexts (FlySlip and CitRAZ) within which we are exploiting this work.

References

[1] E. Briscoe, R. Drysdale, S.Teufel and S. Micklem. (2005) FlySlip: Integrating Literature, Experiments and Curation in Drosophila Genomics Research 2005 (UK BBSRC grant 16291);http://www.cl.cam.ac.uk/users/av308/Pr oject_Index

[2] E. Briscoe & J. Carroll (2002) Robust Accurate Statistical Annotation of General Text. In Proc of 3rd Int. Conference on Language Resources and Evaluation (LREC2002) pp 1499-1504, Las Palmas, Canary Islands.

[3] D.P.A. Corney, B.F. Buxton, W.B. Langdon and D.T. Jones (2004) BioRAT: Extracting Biological Information from Full-length Papers. BioInformatics 20 3206-13

[4] H. Cunningham, D. Maynard, K. Bontcheva, V. Tablan. (2002) GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications. Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL'02). Philadelphia, July 2002.

[5] R.A. Drysdale, M.S. Crosby and The Flybase Consortium (2005). FlyBase: genes and gene models Nucleic Acids Research 33 D390-D395

[6] D.X. Le, S.R. Straughan and G.R.Thoma (2002) Greek Alphabet Recognition Technique for Biomedical Document. Proceedings of the 5th Int. Workshop on Document Analysis Systems, Berlin. pp 423-42

[7] D.D. Lewis & K. Sparck-Jones (1996) Natural language processing for information retrieval, in Communications of the ACM (CACM) 1996 Vol. 39, No. 1, pp 92-101

[8] B. Mathiak & S. Eckstein (2004) Five Steps to Text Mining in Biomedical Literature in Proceedings of the 2nd European Workshop on Data Mining and Text Mining in Bioinformatics

[9] S. Teufel (1999) Argumentative Zoning: Information Extraction from Scientific Text. Ph.D. thesis, School of Informatics, University of Edinburgh, Edinburgh, UK.

[10] S. Teufel. (2004) CitRAZ: Rhetorical Citation Maps and Domain Independent Argumentative Zoning (UK EPSRC grant GR/S27832/01)

[11] G. R. Thoma (2001) Automating the production of bibliographic records for MEDLINE. Internal R&D report, CEB, LHNCBC, NLM, 2001.