Collection and linguistic processing of a large-scale corpus of medical articles

Simone Teufel

Computer Laboratory,
Cambridge University, UK
Simone.Teufel@cl.cam.ac.uk

Noemie Elhadad

Computer Science Department,
Columbia University, New York
noemie@cs.columbia.edu

Abstract

We have collected a large-scale corpus of electronic articles in the cardiology domain (85 million+ words) in the framework of a digital library project that tailors the presentation of online medical literature to both patients and healthcare providers. We describe the web-based and XML technologies we used for the collection, encoding and linguistic processing of the corpus. This resulted in a large-scale, high-quality, thoroughly marked-up resource which is used by many researchers in our project, in the areas of natural language processing, information retrieval and medical informatics. We show how the final use of the resource has influenced the design of its structural and linguistic encoding. The procedure we describe is general enough to be of use to researchers in a similar position wishing to compile, encode and linguistically annotate their own corpus from the web.

1. Motivation

Large-scale annotated corpus material is essential for the robust acquisition of statistical information in many aspects of natural language processing. The corpus we report about here was needed in our digital library (DL) project PERSIVAL (cf. the overview in (McKeown et al., 2001)). The goal of the project is to personalise searches of the medical literature by selecting and presenting tailored summaries of those documents that specifically match a patient under care. Searching databases of published medical results is a major activity for doctors in training, but also for experienced doctors in order to keep up with the latest results.

In our approach, scientific articles are compared to the patients' electronic patient record. Secure access to this record is provided via the Clinical Information System in place at New York Presbytarian Hospital (Clayton et al., 1992). Several researchers work on further NLP tasks such as the generation of fluent summaries from the relevant scientific articles, and others work on non-NLP modules such as search, human interface, medical knowledge modelling, and video summarisation.

The need for a large-scale corpus arose during the early stages of the project for three tasks:

- to train an information extraction module that selects relevant information from the document;
- to train a statistical module that identifies medical terms in the articles; and
- to provide real-world material for a large-scale evaluation of the information retrieval and search modules which are used for the patient-specific information filtering.

The corpus needed to be large-scale enough to address IR-relevant problems, and yet it needed to be analysed with a high degree of linguistic knowledge (POS-tags and medical terminology), due to the sophisticated processing necessary for tasks such as summarisation and context-sensitive presentation which is aware of which kind of information the user has seen before. Also, it was essential that as much structure of the article (in terms of paragraphs, sections, section headings, captions etc...) was captured as possible. This paper describes the sampling of appropriate articles, the technical steps involved in the collection, the XML encoding to linguistically process and represent the complex structure of scientific articles.

2. Data sampling

PERSIVAL focuses on the field of cardiology. According to the Ulrich's Periodicals Directory (Ulrich's, 2001), there are 312 journals in cardiovascular disease alone, which is one among the many subfields of cardiology. In the field of cardiology in general, there are at least 700 journals.

There are many journals of potential relevance to the field of cardiology. For example, in the narrow field of cardiac anesthesiology there are five regularly published scientific journals; but relevant information may appear in any of the 35 journals in anesthesiology, the 60 journals in cardiology, the 40 journals in cardiothracic surgery, on even in the more than 1000 journals in the general field of internal medicine.

Our data sampling was guided by two quality criteria, medical experts' judgement and citation-based impact factors from the Institute for Scientific Information (ISI). We obtained a list of 300+ potentially relevant journals (all cardiology-related plus some internal medicine journals) from the Columbia Health Library that are electronically available and for which Columbia has an access license. We asked three medical experts to rank the list by their likelihood of containing a cardiology article. We also asked them to rank the articles according to quality. This resulted in six lists (3 experts X 2 questions).

ISI calculates the impact factor of the journal based on its citation index, i.e. how many times articles in other journals cite articles in that journal. The assumption is that higher-quality journals get cited more overall.

We compiled an initial quality-ranked list ("wish list") for our project using a simple scoring system to merge these seven lists. For each list, a journal at rank r would receive a score of $\frac{1}{r}$; that is, the highest ranking journal on each list received a score of 1, the next highest ranking a score of $\frac{1}{2}$ and so on. We then summed the seven scores per journal. *Circulation* was by far the highest ranking journal.

Finally, the "wish list" was restricted to those journals which were electronically available in the right format (we required full text of the article in HTML, as pdf does not contain explicit information about the external structure). Only 5% of the journals on our list were electronically available in HTML; thankfully, the highest ranking journal, *Circulation*, was amongst these. Most of the content available in this format is recent (since 1993). Increasingly more content will become electronically available in the future, with many electronic journals providing backdating stepwise from 1993 into the past.

3. Collection

The 22 journals returned by our sampling method are edited by different publishers. Due to copyright issues, it was essential that we obtained research copyright for the electronic versions of all articles before download began. For the journals that are publicly available, we implemented a set of perl scripts to spider ("harvest") HTML pages, fitting the different formats of the crawled web sites. For the journals which were proprietorily owned, but which we had electronic access rights to, we used a specialised downloading protocol. Download was spread out over several weeks and resulted in a corpus of 29,890 articles (more than 85 Million words). Our corpus is described in Figure 1. We invented a naming scheme that encodes article name and appearance date in an unambiguous fashion.

3.1. Collecting journals freely available on the Web

Some publishers, such as Medscape makes certain journals, for instance the American Heart Journal, freely available on the Web. We implemented publisher-specific perl scripts that parse the table of contents of the different issues of a journal, select the HTML links to the full-text HTML version of the articles, download them, and automatically assign a file name following our naming convention to them.

3.2. Collecting journals with proprietory access

The majority of the other journals are not freely downloadable; they are distributed through the OVID publishing company. The collection of these journals was provided by consultation with OVID via a special protocol called Jumpstarts. The highest ranked journal *Circulation* provides almost 50% of the collection.

4. XML Encoding

The documents are originally encoded in different, publisher-specific, purely presentational HTML formats. However, a unified and abstract format such as XML is preferrable, as it can encodes the complex structure of medical articles in the same form for all articles. We wrote a series of perl scripts which convert the different idiosyncratic HTML markup schemes into a single XML markup which we designed and for which we wrote a document type definition (DTD) describing the articles' logical structure.

5. Linguistic processing

We then further annotate the articles automatically by applying a pipeline of different processors to it, using the tool TTT (Grover et al., 1999). The texts go through the following processing steps: first, they get tokenised using a regular grammar over characters. Then, sentence boundaries are identified. POS-tagging is performed. After this processing, the tokenisation and POS-information can be read off from the W-tag (for word) in Figure 3. The attribute C (for category) gives the POS tag, using the UPenn tagset. Next, a regular grammar for identification of noun phrases is applied, in order to find potential terms.

6. Terminology determination

In the next step, medical terms are identified. We currently have two different term identifiers: one is based on a robust lookup in the medical database UMLS (Lindberg et al., 1993; Humphreys et al., 1998), the other is a statistical term identifier which takes term association and term frequency across domains (general vs. medical English) into account. Figure 3 shows the UMLS version. After terms are identified, coordinated terms are resolved, such that "left and right ventricular pressure" is turned into two terms "left ventricular pressure" and "right ventricular pressure" (we call these logical terms, as they aren't present in the surface text, but represent the real medical terms). Our XML encoding of such "non-linear" terms encodes the logical terms in an attribute of the empty element NLTERM (for "nonlinear term"). This has the advantage of makding the logical terms accessible for successive processing, i.e. term comparisons and article reranking, while "hiding" them for the raw text, by leaving the textual representation untouched. This complies with one fundamental philosophy of XML-encoding, namely that any kind of analysis performed on XML-data should never change the data itself, but be added in the form tags and attributes.

Acronyms receive a similar treatment. They are looked up in UMLS, and if they are found, their resolved name is attached to the term as an (invisible) attribute.

We keep several XML versions of our corpus, one where the scientific text has only its sections, headings and paragraphs marked; one where sentences are identified and the text is POS-tagged, and the third version has additionally all terminological terms marked. This corresponds to three increasingly information-rich DTDs. Figure 2 reproduces the most informative of these, i.e. the version containing medical terms, whereas Figure 3 shows an extract of the corresponding xml file.

In this encoding, Words (W element) can be wrapped into terms (TERM). Attributes associated with terms (or nonlinear terms, NLTERM) include: their CUI (unique concept identifier, from UMLS), the section where this term appears, whether or not it is an acronym (if it is, the attribute ACROSTING contains the resolved name), the term's semantic type (as defined by UMLS), and the attribute EXCLUSION, which encodes whether the sentence the term occurs in describes exclusion criteria. This information

Journal name	No. of articles	No. of words
Journal of the American College of Cardiology	1,816	8,332,759
Journal of the American College of Surgeons	453	1,466,532
American Heart Journal	926	1,907,560
American Journal of Cardiology	3,135	10,352,558
American Journal of Hypertension	643	2,189,571
American Journal of Medicine	821	2,507,360
American Journal of Surgery	570	1,558,415
Atherosclerosis	1,030	4,830,901
Annals of Thoracic Surgery	3,000	7,966,535
Annals of Vascular Surgery	216	673,044
Cardiovascular and Interventional Radiology	169	382,524
Circulation	13,516	34,340,009
Cardiovascular Surgery	304	909,855
European Journal of Cardio-Thoracic Surgery	1,003	1,295,982
International Journal of Cardiology	258	733,577
Journal of Clinical Anesthesia	249	751,540
Pediatric Cardiology	364	659,772
Trends in Cardiovascular Medicine	100	431,691
Thrombosis Research	715	2,512,886
World Journal of Surgery	496	2,039,592
Total	29,784	85,842,699

Figure 1: List of Journals in our Corpus

is useful for successive matching of articles with patient records. If a terms appears in a piece of text describing exclusion criteria, or in a section with direct negation, the term is considered a negative match, i.e. it does *not* describe the patient population of the article.

Besides identifying medical terms within the text, we also find their corresponding values (when present) using a grammar of regular expressions.

In the following piece of XML-encoded text, for instance, "*patients with LVEF of > 40%*", we see that "*40%*" was identified as the value of the term "*LVEF*". The correspondance between term and value is encoded with the attribute BELONGS-TO of the VALUE element. The ACROSTRING attribute of the TERM element also tells us that "*LVEF*" means "*left ventricular ejection fraction*":

```
   <TERM CUI="C0030705" ID="T-
327" SECTION="Discussion" SEM-
TYPE="Patient or Disabled Group @"><W
C='NNS'>patients</W><SEMCLASS TYPE='Patient
or Disabled Group'/></TERM>
   <W C='IN'>with</W> <W C='DT'>an</W>
<TERM CUI="C0809096" ID="T-328" SEC-
TION="Discussion" ACROSTRING="Left ventric-
ular ejection fraction" SEMTYPE="Organism
Attribute @"><W C='NNP'>LVEF</W><SEMCLASS
TYPE='Organism Attribute'/></TERM>
   <W C='IN'>of</W> <W C='GT'>></W><VALUE
BELONGS-TO='T-328'><W C='CD'>40</W><W
C='SYM'>%</W></VALUE>
```

In order to help in the debugging of writing tokenisation rules, or the writing of grammar rules for value detection and exclusion criteria, we use a mechanism which turns XML into HTML, for easy browsing of the output in a regular Web-browser. The colour feature of Cascading Style Sheets (CSS-2) is used to mark the output of each particular processing step.

7. Use of the resource in our Digital Libraries project

The resource is currently used by three sub-groups in the project: for reranking of articles by relevance to a given patient (Teufel et al., 2001), for patient-specific summarisation (Elhadad and McKeown, 2001) and by providing the right terminology for patient-specific video search and summarisation (Ebadollahi et al., 2001a; Ebadollahi et al., 2001b).

7.1. Reranking

The automatically identified medical terms and values are used for patient-specific search, which is implemented as a comparison between patient records and scientific articles. A patient record is a complex of reports, some of which contain text (e.g. a patient's prior medical history), some of which have predefined fields filled in, and some of which consist of tables of numbers (e.g. microbiology lab reports and blood chemistry panels). A patient record can contain upwards of 100 of such reports. A specialised term weighting algorithm measures the similarity between a particular patient record and all articles returned by a search engine. This turns a general search into a patient-specific search, selecting only those articles that are relevant to the patient. The system and an evaluation are described in (Teufel et al., 2001).

7.2. Text Summarisation

The summariser for technical articles in PERSIVAL takes as input a specific patient record and a list of reranked articles that are relevant to the the patient. It generates a summary that reports on the findings of the different input articles and identifies commonalities and differences accross them (Elhadad and McKeown, 2001).

The summariser takes advantage of the XML encoding of the input articles (especially the identification of the different sections in the articles) to pinpoint the locations in the articles to look for salient information. Typically, the Abstract and the Results sections contain findings that could be relevant for the patient in question, and that may be included in the summary.

The summariser then uses information extraction techniques to extract templates for the relevant findings. The article with the marked medical terms and their associated values form an ideal semantically preprocessed input. A set of templates is applied to extract the findings reported in the articles. Generation techniques are then used to combine and order the different extracted templates into a coherent and fluent summary.

Because our collected corpus contains a large amount of articles, we used it as a training and testing corpus for learning information extraction templates in a supervised manner.

7.3. Video Summarisation

When videos of different cardiology tests relevant to the patient in question are available in the system in electronic form, PERSIVAL allows for a summary of the main findings of the videos. The medical terms used in the patient records as well as their corresponding values are identified using our technology and are linked to different views of the video summary. This provides the user with a good semantic correspondance between text and videos, and proves to be useful especially for doctors in training (Ebadollahi et al., 2001a).

8. Conclusion and future work

Our procedure has proven to be workable in practice, resulting in a large-scale, high-quality, thoroughly marked-up resource which is used by many subgroups of a large-scale project. This procedure was also rather resource-efficient, using half-time effort of 2-3 persons over a 4 week period. Several subprojects need further linguistic processing like POS-tagging, sentence boundary detection, term identification, value identification, determination of negative context and acronym resolution, which the pipeline presented here readily provides. Having such complex information encoded robustly for a vast volume of information was facilitated greatly by XML technology, for instance the possibility to check validity of each output step, and the possiblity to browse output as HTML in a regular browser (i.e. without having to program any specific output device), with the currently checked steps presented to the developer in colour.

Current web-technology, however, allows us to further improve one step of the procedure, namely the compilation from HTML to XML. Currently, this step consists of a set of perl-scripts. A shorter and more elegant procedure involves the use of an HTML to XHTML converter (which, for instance, automatically ensures that all tags are closed), followed by an XSLT script (for a description of the XSLT language, cf. (Kay, 2001)).

```
<!ELEMENT PAPER (TITLE,AUTHORS,FILENO,APPEARED,KEYWORDS?,
   (ABSTRACT|STRUCT-ABSTRACT),BODY?,REFERENCES?)>
<!ELEMENT TITLE (#PCDATA|TERM|NLTERM|VALUE|W|REF|REFA)*>
<!ELEMENT REFLABEL (#PCDATA)>
<!ELEMENT AUTHORS (#PCDATA|AUTHOR)*>
<!ELEMENT AUTHOR (#PCDATA)*>
<!ELEMENT FILENO (#PCDATA)*>
<!ELEMENT APPEARED (#PCDATA|DATE|JOURNAL|VOL|ISS|YEAR)*>
<!ELEMENT KEYWORDS (#PCDATA|W)*>
<!ELEMENT JOURNAL (#PCDATA)*>
<!ELEMENT YEAR   (#PCDATA)>
<!ELEMENT VOL (#PCDATA)>
<!ELEMENT ISS (#PCDATA)>
<!ELEMENT ABSTRACT  (#PCDATA|A-S)*>
<!ELEMENT STRUCT-ABSTRACT  (#PCDATA|DIV)*>
<!ELEMENT A-S (#PCDATA|W|TERM|NLTERM|VALUE|REF|REFA)*>
<!ATTLIST A-S
        ID  ID  #IMPLIED>
<!ELEMENT BODY (DIV)*>
<!ELEMENT DIV  (HEADER?, (DIV|P|IMAGE)*)>
<!ATTLIST DIV
        DEPTH  CDATA  #REQUIRED>
<!ELEMENT HEADER (#PCDATA|W|TERM|NLTERM|VALUE|REF|REFA)*>
<!ATTLIST HEADER
        ID  ID  #IMPLIED>
<!ELEMENT P (#PCDATA|S|IMAGE|REF|REFA)*>
<!ELEMENT IMAGE (#PCDATA|CAPTION)*>
<!ATTLIST IMAGE
        ID  ID  #IMPLIED>
<!ELEMENT CAPTION (#PCDATA|W|TERM|NLTERM|VALUE|REF|REFA)*>
<!ELEMENT S  (#PCDATA|W|TERM|NLTERM|VALUE|REF|REFA)*>
<!ATTLIST S
        EXCLUSION (YES|NO) "NO"
        ID  ID  #IMPLIED >
<!ELEMENT TERM (#PCDATA|W|SEMCLASS)* >
<!ATTLIST TERM   ID  ID  #IMPLIED
        CUI        CDATA  #IMPLIED
        SECTION    CDATA  #IMPLIED
        ACROSTRING CDATA #IMPLIED
        EXCLUSION  (YES|NO)  "NO"
        SEMTYPE    CDATA  #IMPLIED>
<!ELEMENT NLTERM (#PCDATA)>
<!ATTLIST NLTERM
        ID        ID        #REQUIRED
        SEMTYPE   CDATA     #REQUIRED
        CUI       CDATA     #REQUIRED
        T         CDATA     #REQUIRED
SECTION   CDATA     #IMPLIED
EXCLUSION (YES|NO)  "NO">
<!ELEMENT VALUE (#PCDATA|W)* >
<!ATTLIST VALUE
        BELONGS-TO   CDATA #IMPLIED>
<!ELEMENT SEMCLASS  EMPTY>
<!ATTLIST SEMCLASS
TYPE         CDATA    #IMPLIED>
<!ELEMENT W (#PCDATA)>
<!ATTLIST W
        C        CDATA  #IMPLIED>
<!ELEMENT REF (#PCDATA)*>
<!ATTLIST REF
        SELF      (YES|NO)  "NO"
        C         CDATA     "NNP">
<!ELEMENT REFA (#PCDATA|SURNAME)*>
<!ATTLIST REFA
        SELF      (YES|NO)  "NO"
        C         CDATA     "NNP">
<!ELEMENT REFERENCES (P|REFERENCE)*>
<!ELEMENT REFERENCE (#PCDATA|W|REF|REFA)*>
```

Figure 2: DTD for POS-tagged corpus, including medical terms

While full citation linking is not performed yet, the encoding of the corpus would enable us to add this level of analysis without a great deal of work. We have developed a citation parser, based on a version delivered with TTT, which finds citations in running text. A symbolic grammar in parallel parses the citation list at the end of each paper and identifies author names, titles and other bibliographic information of the cited articles. The pipelined architecture of the system makes it nevertheless easy to detect all cited authors' names along with citation parsing, which is

```
<TITLE> <W C=’NNS’>Effects</W> <W C=’IN’>of</W> <TERM CUI="C0126174" ID="T-0" SECTION="Title" SEMTYPE="Organic
Chemical Pharmacologic Substance @"><W C=’NNP’>Losartan</W><SEMCLASS TYPE=’Organic Chemical’/><SEMCLASS
TYPE=’Chemical’/></TERM> <W C=’CC’>and</W> <TERM CUI="C0006938" ID="T-1" SECTION="Title" SEMTYPE="Pharmacologic
Substance Amino Acid, Peptide, or Protein @"><W C=’NNP’>Captopril</W><SEMCLASS TYPE=’Amino Acid, Peptide, or
P’/></TERM> <W C=’IN’>on</W> <W C=’NNP’>Left</W> <W C=’NNP’>Ventricular</W> <W C=’NNP’>Volumes</W> <W C=’IN’>in</W>
<W C=’NNP’>Elderly</W><TERM CUI="C0030705" ID="T-2" SECTION="Title" SEMTYPE="Patient or Disabled Group @"><W
C=’NNP’>Patients</W><SEMCLASS TYPE=’Patient or Disabled Group’/></TERM> <W C=’IN’>with</W> <TERM CUI="C0018801"
ID="T-3" SECTION="Title" SEMTYPE="Disease or Syndrome @"><W C=’NNP’>Heart</W> <W C=’NNP’>Failure</W><SEMCLASS
TYPE=’Disease or Syndrome’/></TERM> <W C=’CM’>:</W> <W C=’NNS’>Results</W> <W C=’IN’>of</W> <W C=’DT’>the</W> <W
C=’NNP’>ELITE</W> <W C=’NNP’>Ventricular</W> <W C=’NNP’>Function</W> <W C=’NNP’>Substudy</W> </TITLE>

<AUTHORS>Marvin A. Konstam, MD, Richard D. Patten, MD, Ignatious Thomas, MD, Tarik Ramahi, MD, Kenneth La Bresh, MD, Steven Goldman, MD, William
Lewis, MD, Alan Gradman, MD, K. Stanley Self, MD, Vera Bittner, MD, William Rand, PhD, Debra Kinan, RT(N), John J. Smith, MD, PhD, Tim Ford, MD, Robert
Segal, MD, James E. Udelson, MD, Tufts University, New England Medical Center, Boston, Mass; Medical Research Institute, Sidell, La; Yale University, New
Haven, Conn; Brown University, Pawtucket, RI; Veterans Administration Hospital, Tucson, Ariz; University of California at Davis Medical Center, Sacramento,
Calif; Western Pennsylvania Hospital, Pittsburgh, Pa; Self Center, P.C., Fairhope, Ala; University of Alabama at Birmingham, Birmingham, Ala; Merck &; Co, Inc,
West Point, Pa.</AUTHORS>

<FILENO>ahj_139_06_0921</FILENO> <APPEARED> <JOURNAL>Am Heart J</JOURNAL> <VOL>139</VOL> <ISS>6</ISS> <YEAR>2000</YEAR>
</APPEARED>

<STRUCT-ABSTRACT> <DIV DEPTH=’1’> <HEADER ID=’H-0’> <W C=’NN’>Background</W> </HEADER> <P> <S ID=’S-0’> <W
C=’DT’>The</W> <W C=’NN’>mechanism</W> <W C=’IN’>by</W> <W C=’WDT’>which</W> <TERM CUI="C0003018" ID="T-4"
SECTION="Abstract" SEMTYPE="Pharmacologic Substance Biologically Active Substance Amino Acid, Peptide, or
Protein @"><W C=’NN’>angiotensin</W><SEMCLASS TYPE=’Amino Acid, Peptide, or P’/></TERM> <W C=’DASH’> - </W> <W
C=’VBG’>converting</W> <TERM CUI="C0014432" ID="T-5" SECTION="Abstract" SEMTYPE="Pharmacologic Substance @"><W
C=’NN’>enzyme</W> <W C=’NNS’>inhibitors</W></TERM> <W C=’VBP’>reduce</W> <W C=’NN’>mortality</W> <W C=’NNS’>rates</W>
<W C=’CC’>and</W> <TERM CUI="C0242656" ID="T-6" SECTION="Abstract" SEMTYPE="Pathologic Function @"><W
C=’NN’>disease</W> <W C=’NN’>progression</W></TERM> <W C=’IN’>in</W> <TERM CUI="C0030705" ID="T-7" SECTION="Abstract"
SEMTYPE="Patient or Disabled Group @"><W C=’NNS’>patients</W><SEMCLASS TYPE=’Patient or Disabled Group’/></TERM> <W
C=’IN’>with</W> <TERM CUI="C0018801" ID="T-8" SECTION="Abstract" SEMTYPE="Disease or Syndrome @"><W C=’NN’>heart</W>
<W C=’NN’>failure</W><SEMCLASS TYPE=’Disease or Syndrome’/></TERM> <W C=’VBZ’>is</W> <W C=’RB’>likely</W> <W
C=’VBN’>mediated</W> <W C=’IN’>in</W> <W C=’NN’>part</W> <W C=’IN’>through</W> <TERM CUI="C0687732 C0742770" ID="T-9"
SECTION="Abstract" SEMTYPE="Therapeutic or Preventive Procedure @Finding @"><W C=’NN’>prevention</W><SEMCLASS
TYPE=’Finding’/><SEMCLASS TYPE=’Therapeutic or Preventive’/></TERM> <W C=’IN’>of</W> <W C=’JJ’>adverse</W> <W
C=’NN’>ventricular</W> <W C=’VBG’>remodeling</W><W C=’FS’>.</W></S>
```

Figure 3: XML representation of title, authors and first abstract sentence of the article “Effects of Losartan and Captopril on Left Ventricula Volumes in elderly patients with Heart Failure: Results of the ELITE Ventricular Function Substudy”, *Am. Heart J 139 (6)* (article ahj_139_06_0921; textual data from article in bold face)

a practical and robust way of performing citation-specific named entity recognition. The mention of author names in running text – which proved a frequent phenomenon in another genre we worked in (computational linguistics) – does not seem to be a as common in the cardiology domain. Almost all citation is done parenthetically, i.e. in brackets at the end of the sentence, rather than authorially, i.e forming a syntactically integral part of the sentence.

9. References

P. D. Clayton, R. V. Sideli, and S. Sengupta. 1992. Open architecture and integrated information at Columbia-Presbyterian Medical Center. *M.D. Computing*, **9**(5):297–303.

Sharam Ebadollahi, Shih-Fu Chang, and Henry Wu. 2001a. Echocardiogram video summarization. In *SPIE Medical Imaging*.

Sharam Ebadollahi, Shih-Fu Chang, and Henry Wu. 2001b. Indexing and summarization of echocardiogram videos. In *The scientific session of the annual conference of the American College of Cardiology*.

Noemie Elhadad and Kathleen R. McKeown. 2001. Towards generating patient specific summaries of medical articles. In *Proceedings of NAACL-2001 Workshop “Automatic Summarization”*.

Claire Grover, Andrei Mikheev, and Colin Matheson. 1999. LT TTT version 1.0: Text tokenisation software. Technical report, Human Communication Research Centre, University of Edinburgh.

B. L. Humphreys, D. A. B. Lindberg, H. M. Schoolman, and G. O. Barnett. 1998. The Unified Medical Language System: An informatics research collaboration. *JAMIA*, **5**:1–11.

Michael Kay. 2001. *XSLT Programmer’s Reference*. WROX, 2nd edition edition.

D. A. B. Lindberg, B. L. Humphreys, and McCray A. T. 1993. The Unified Medical Language System. *Methods of Information in Medicine*, **32**:281–291.

Kathleen R. McKeown, Shi-Fu Chang, Steve Feiner James Cimino, Carol Friedman, Luis Gravano, Vasileios Hatzivassiloglou, Steve Johnson, Desmond Jordan, Judith Klavans, André Kushniruk, Vimla Patel, and Simone Teufel. 2001. PERSIVAL, a system for personalized search and summarization over multimedia healthcare information. In *Proc. JCDL*.

Simone Teufel, Vasileios Hatzivassiloglou, Kathleen R. McKeown, Kathy Dunn, Desmond Jordan, Sergey Sigelman, and André Kushniruk. 2001. Personalized medical article selection using patient record information. In *Proc. Annual Symp. AMIA*.

Ulrich’s. 2001. Ulrich’s periodicals directory. Available at `http://www.ulrichsweb.com`.