A segmentation method for bibliographic references by contextual tagging of fields

Dominique Besagni, Abdel Belaïd, Nelly Benet

A Segmentation Method for Bibliographic References by Contextual Tagging of Fields

Dominique Besagni[1], Abdel Belaïd[2] and Nelly Benet[2],
[1]*URI, INIST-CNRS, 2 Allée du parc de Brabois, 54514 Vandoeuvre-lès-Nancy Cedex, France,*
[2]*LORIA-CNRS, Campus scientifique, BP 239, 54506 Vandoeuvre-lès-Nancy Cedex, France,*
besagni@inist.fr, abelaid@loria.fr

Abstract

In this paper, a method based on part-of-speech tagging (PoS) is used for bibliographic reference structure. This method operates on a roughly structured ASCII file, produced by OCR.. Because of the heterogeneity of the reference structure, the method acts in a bottom-up way, without an a priori model, gathering structural elements from basic tags to sub-fields and fields. Significant tags are first grouped in homogeneous classes according to their grammar categories and then reduced in canonical forms corresponding to record fields: ``authors'', "title", "conference name", "date", etc. Non labelled tokens are integrated in one or another field by either applying PoS correction rules or using a structure model generated from well-detected records. The designed prototype operates with a great satisfaction on different record layouts and character recognition qualities. Without manual intervention, 96.6% words are correctly attributed, and about 75,9% references are completely segmented from 2500 references.

1. Introduction

The "bibliographic references" mentioned in this paper correspond to the citations mentioned at the end of scientific publications. It is one of the structural elements of a standard scientific article that can be used for analysis.

The foundation in the 1960s at Philadelphia (USA) of the Institute for Scientific Information (ISI) by Eugene Garfield was instrumental in turning the citation in a unit of measure. Used at first only as a tool for information retrieval, the citation has become an important criterion because it allows to distinguish among different publications those which received the approbation of the scientific community. By the same token, the citation is also used to appraise scientific journals especially with the impact factor calculated as the average number of citations a paper receives over a period of 2 years [1].

The Institute for Scientific and Technical Information (INIST) of the French National Centre for Scientific Research (CNRS) has undertaken an experiment of digitisation of these references. This is done especially because of the interest of that structural element in the field of information retrieval and in the field of scientific information analysis: citation and co-citation analysis [2].

In collaboration with LORIA Laboratory, INIST engaged a structuring program for these bibliographic references. The objective is to identify the different fields in these citations such as: "authors", "title", "publication date", etc. Because of the structure complexity, a bottom-up method based on a part-of-speech tagging was investigated, using some dictionaries and some patterns recognized in the citation. The field location is based on many syntactic and statistic aspects such as the position regularity, the tag occurrence in some fields, etc.

In the literature, we identified a similar work done at the NEC Research Institute as part of the CiteSeer system [3]. The Autonomous Citation Indexing (ACI) uses a top-down methodology applying heuristics to parse citations. This approach employs some invariants considering that the fields of a citation have relatively uniform syntax, position and composition. It uses trends in syntactic relationships between fields to predict where a desired field exists if at all.

Even though this method is reportedly accurate, its functioning is not explicit enough to measure its efficiency on OCR output.

In this work, we propose a method based on text coding which in turn is based on Part of Speech tagging (PoS) [4,5]. The idea of this method, employed in language processing and text indexing, is to reassemble nouns in nominal syntagms representing the same information. The nouns are given by a specific morphological tagging.

This method can be applied in reference recognition for field identification by reassembling in the same

Proceedings of the Seventh International Conference on Document Analysis and Recognition (ICDAR'03)
0-7695-1960-1/03 $17.00 © 2003 IEEE

syntagm ``title'', ``authors'', etc., words having similar tags. The process of tagging consists of three stages~ : tokenisation, morphological analysis, and syntactical grouping and disambiguation. The tokeniser isolates each textual term and separates numerical chains from alphabetic terms.

The morphological analyser contains a transducer lexicon. It produces all the legitimate tags for words that appear in the lexicon. If a word is not in the lexicon, a guesser is consulted. The guesser employs another finite-state transducer which examines the context and decides to assign the token to ``title'' or to ``authors'' depending on prefixes, inflectional information and productive endings that it finds.

We applied this method with great success on table of contents in order to detect and structure their different articles [6]. Here, we show its adaptation on bibliographic records having more fields with more complicated internal fields.

In the following sections, we shall describe the data and the segmentation method, then we shall explain the different steps of that method (tagging, locating fields, modelling and correction) before relating the experiment carried out with it.

2. Data and method

The raw data, obtained by OCR, is a set of "well-formed" XML documents in which each reference from the same article is singled out (see figure 1).

```
<INFCOM fic="1998/refm278.dat">
<NUMACQ><CLEA>35400007110423</CLEA><CLE
B>0030</CLEB></NUMACQ>
<REFBIB copie="0" >1 American Cancer Society.
Cancer Facts and Figures-1997, American Cancer
Society: Atlanta, 1997.</REFBIB>
<REFBIB copie="0" >2 Bonnadonna G, Valgussa P,
Moliteri A, Zambetti M, Brambilla C. Adjuvant
cyclophosphamide, methotrexate, and fluorouracil in
node-positive breast cancer: The results of 20 years of
follow-up. N Engl Med 1995; 332: 901-906.</REFBIB>
<REFBIB copie="0" >3 Booser DJ, Hortobagyi GN.
Treatment of locally advanced breast cancer. Seminars
in Oncology 1992; 19: 278-285.</REFBIB>
<REFBIB copie="0" >4 Rouëssé J et al. J Clin Oncol
1986; 4: 1765-1771.</REFBIB>
<REFBIB copie="0" >5 Swain SM et al. Neoadjuvant
chemotherapy in the combined modality approach of
locally advanced non-metastatic breast cancer. Cancer
Res 1987; 47: 3889-3894.</REFBIB>
</INFCOM>
```

Figure 1. Example of data file.

However the different parts of that reference (authors, title, journal, date...) are not identified. The character set used in the data files is ISO-latin 1 (standard ISO 8859-1).The other alphabetical characters that do not belong to this character set are represented as character entities as defined by SGML (ISO 8879:1986). For example, "**Ş**" represents the uppercase Latin letter "S" with a cedilla.

The problems we encounter while segmenting a bibliographical reference in its different fields are of several kinds:

- those due to the digitisation: unrecognised characters, badly recognised characters (as the uppercase Latin letter "D" which sometime gives the Latin letter "I" followed by a right parenthesis) or even forgotten characters (as it is sometime the case for punctuation marks),
- those due to the heterogeneity of the data: the structure of a reference depends on the type of document it refers to and on the origin of the quoting article since the model of the citation depends on the journal in which it is published. Although on that point, it is to be noted that publishers don't enforce their own rules with the same rigorousness and the structure of a reference may vary greatly from one paper to another in the same journal.
- to that, you may add typing errors, omissions and sometime footnotes which have nothing to do with bibliographical references.

Still, there are a few regularities:

- within the same paper, references have the same structure (for the same type of quoted document),
- when authors' names are present (general case), they are always at the beginning of the reference,
- for the references to journal articles, a field like the date of publication can be found only in a very limited number of positions:
    - after the authors,
    - after the journal title,
    - after the pagination.

But always at the same place within a set of references. Likewise, the paper title is always before the journal title.

All this allows to describe very generic models of bibliographical references, but the uncertainties in the details make it hard to use a method based on a set of predefined models of what a reference is supposed to be.

To solve that problem, we use a method derived from one devised to recognize tables of content [1]. That method comprises three steps:

- a primary tagging of citation components,
- a syntactic analysis by searching for terms (based on pattern regularities and redundancies) and term associations (part of speech) revealing the field nature,
- a structural analysis which realizes a verification and correction task. Considering structural models generated from well analysed citations, this approach tries to correct the bad remaining citations.

We shall describe these three steps in the next sections and in section 6 the experiment carried out, knowing that for the time being we concentrate on segmenting citations from journal articles, the most frequent type of references and the main subject of bibliometric studies.

3. Primary tagging

Each textual element receives a tag from a predefined list (see table 1). Moreover, a number tag is followed by the number of digits (for example, "2003" is tagged "NM4") and a punctuation mark tag is followed by the punctuation mark itself ("-" is tagged "PU-"). An element may receive several tags because it can belong to different morphological categories. When no attribution is made, the unknown tag "UN" is assigned.

Table 1. Main primary tags.

Tag	Meaning	Tag	Meaning
AN	Alphanumeric string	IN	Expression "In:"
CC	Connector (and, & …)	IT	Initial
CWC	Common noun, initial capital	JM	Journal marker
CWL	Common noun, lowercase	NM*n*	Number (*n* digits)
CWU	Common noun, uppercase	PN	Proper name
EA	Expression "*et al.*"	PU*s*	Punctuation mark *s*
ED	Editor (Ed., Eds.)	UN	Unknown

The lexicons used by the tagger came from electronic resources available at INIST, as the PASCAL database for authors' name, journal titles and countries or electronic dictionaries for English or French nouns or prepositions.

4. Syntactic analysis

This is based on either 1) search of pattern regularities and redundancies or 2) term grouping in parts of speech. In both cases, each field identified receives a predefined tag (see table 2). In the first case, we noticed that these properties are very relevant within the same citation set, to locate some specific fields such as the "date", the "pagination", "citation identifier", etc. which came up at the same position, with the same structure and context. The detection of such regularities reinforced by high frequencies contribute to their easy location.

The approach carried out in the second case is less straightforward. Some grouping rules are needed to reveal the presence of some fields or sub-fields, rules that are handcrafted and selected by trial and error.

Table 2. Main secondary tags.

Tag	Meaning	Tag	Meaning
AU	Authors	PG	Pagination
DA	Date of publication	TIP	Article title
JN	Journal title	VOL	Volume number

In order to adapt to different writing styles and field structures, we employ different categories of grouping rules. Among them, we can quote:

- **Reduction rules:** leading to aggregate identical elements in the part of speech, such as two initials of the first name: IT + IT => IT
- **Forming rules:** initiating the field creation by associating complementary elements, such as initials and proper name: PN + IT => AU
- **Extending rules:** concatenating sub-fields separately recognized, such as author and "et al." to confirm the author field, or extension of the title from an initial nucleus composed of three nouns, by adding other surrounding nouns, connectors and prepositions to widen that field as much as possible: AU + EA => AU
- **Agglutination rules:** absorbing in some obvious contexts the "UN" terms, such as an unknown term between two authors: UN + PU- + PN => PN
- **Mixed rules:** combining forming rules to detect the potential candidates and regularities to select the best one. This is the case of the pagination the structure of which is very variable (see table 3). Besides, it may be preceded by pagination indicators like "p." and the hyphen may be missing.

Table 3. Pagination formats

numeric – numeric
alphanumeric – alphanumeric
alphanumeric – numeric
numeric
alphanumeric

5. Structural analysis

The syntactic analysis as showed in figure 2, has some limits in the field separation for different reasons: some terms are unknown, the title has a too complex structure, confusion between the publication year and the pagination, etc.

The idea of the structural phase is to exploit what was well recognized as a model for the remaining cases. So, the procedure adopted consists on searching for models and then using them for correction.

We proposed two kind of models: inter-fields and intra-fields in order to progressively correct the fields, first by searching for their limits (by the inter-field models) that are then confirmed by the intra-field models.

1. Costall B, Naylor RJ, Tvers MB: Recent advances in the neuropharmacology of 5 - HT3 agonists and antagonists. Rev Neurosci 1988;2(1):41 - 65.

2. Tyers MB: Pharmacology and preclinical antiemetic properties of ondansetron. Semin Oncol 1992;19(suppl 10):1 - 8.

3. Lesser J, Lip H: Prevention of postoperative nausea and vomiting using ondansetron, a new selective 5 - HT3 receptor antagonist. Anesth Analg 1991;72:751 - 755.

4. Bodner M, Poler SM, White P: Antiemetic efficacy of ondansetron after ambulatory surgery. Anesth Analg 1991;73:250 - 254.

5. McKenzie R, Tantisira B, Joslyn AF: Ondansetron. a selective serotonin type 3 (5 - HT3) antagonist, reduces nausea and vomiting in females following major gynecologic surgery. Anesth Analg 1992;74:520 - 522.

6. Alon E, Himmelseher S: Ondansetron in the treatment of postoperative vomiting: a randomized, double - blind comparison with droperidol and metoclopramide. Anesth Analg 1992;75:561 - 565.

7. Raphael JH, Norton AC: Antiemetic efficacy of prophylactic ondansetron in the laparoscopic surgery: randomized double - blind comparison with metoclopramide. Br Anaesth 1993;71: 845 - 848.

Figure 2. Tagged references after syntactic analysis. The same fields have identical colors

5.1 Inter-field modelling

We used a pair modelling revealing the association of consecutive fields. This gives the sequence of possible consecutive fields with their separators.

Tag 1	Tag 2	Percentage	Position	Simple separator	Double separator
AU	DA	97.83	1.00	(93	)(2 ,(2
DA	TIP	56.52	1.96		). 56
~~DA~~	~~AU~~	~~2.17~~	~~2.00~~		~~). 2~~
TIP	JN	56.52	2.62	. 56	
~~AU~~	~~TIP~~	~~2.17~~	~~3.00~~	~~: 2~~	
JN	VOL	65.22	3.20	, 60	
VOL	PG	95.65	3.89	, 93 . 2	

3. Building model

2. Selecting separators

1. Suppressing invalid couples

Figure 3. Consecutive couples parameters

Only punctuation marks are allowed between such couple of fields. When all the different couples are obtained, with their frequency and their separator if any, they are sorted accordingly to their relative position and their logical sequence is determined, keeping only the couple with a significantly greater frequency when confronted with several possibilities. Likewise, only the separators with a significantly high frequency are considered valid. The model is then built by stringing couples together like dominoes as shown in figure 3.

5.2. Inter-field correction

Using that model, incomplete fields are corrected if the surrounding fields are clearly identified and delimited. In such a case, we can extend the field on the right and/or the left until the gap is closed. Put to the extreme, we can deduce the presence of an utterly unrecognised field by the presence of the correct fields and separators around it.

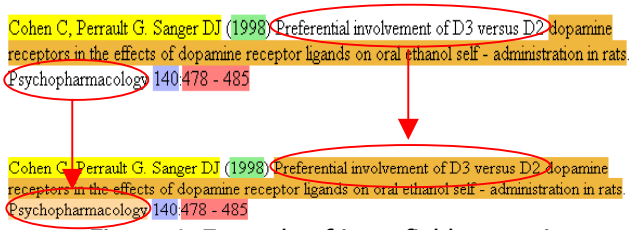

Figure 4. Example of inter-field correction

5.3 Intra-field modelling

Once obtained the field identity and limits, we try to find out the kind of elements used in the field and their structure in terms of sequence and separators. In the case of the field "authors", initials may be in front or behind the author's last name, each name or initial may be followed by a specific punctuation mark. The pattern may be different for the first author and the last author may be preceded by a connector. That is why we consider three different cases as shown in figure 5.

First author	
Pattern:	PN PU, IT PU.
Separator:	PU,
Next authors	
Pattern:	PN PU, IT PU.
Separator:	PU,
Last author	
Pattern:	PN PU, IT PU.
Connector:	and

Figure 5. Example of a model for authors

5.4 Intra-field correction

Using that model, we check each reference for inconsistencies. In the example shown in figure 6, corresponding to the model of figure 5, the connector "and" indicates the position of the last author and the string "Wilson, J.M." fits the pattern of an author's name.

This means the word “Gene” cannot belong to that field and it is therefore excluded.

After that, a new iteration is still necessary to identify the field “Title” from the inter-field model and to test it with its intra-field model.

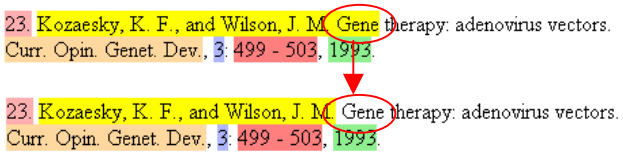

Figure 6. Example of intra-field correction

6. Experiment and results

The experiment was done on 140 journals of pharmacology. The digitisation of these bibliographical references was carried out by a subcontractor. The data set is made of 64 articles chosen at random from the original set. It contains 2,575 references.

We tagged them and we carried out the syntactic and the structural analysis up to the intra-field modelling. At each stage, the results can be visualised using a HTTP server and CGI scripts that highlight each recognised field with a specific colour as shown on figure 2.

Table 4. Results after inter-field correction

Fields	**Complete**	**Partial**	**Not found**	**Wrong**
Authors	90.2%	6.6%	0.3%	2.9%
Title	82.4%	15.4%	1.7%	0.4%
Journal	92.4%	2.9%	3.2%	1.5%
Date	97.7%	0.0%	2.3%	0.0%
Volume	93.6%	0.4%	5.8%	0.2%
Pagination	94.7%	0.6%	4.3%	0.4%
Whole Reference	75.9%	18.8%	0.0%	5.3%

In parallel, the complete data set was tagged by hand so we have a standard against which we can compare the results of our segmentation method. After the inter-field correction stage, 96.6% of words have been placed correctly in the right field while 0.5% have been wrongly attributed. Table 4 shows the results field by field and for the whole reference expressed as the percentage of references where a specific field is complete, incomplete, not found or erroneous. For the whole reference, this means all fields are complete, at least one is incomplete, none are found or at least one is wrong.

7. Conclusion

The method presented here works well on bibliographical references from most articles. For other sets of references with too few citations or too much heterogeneity, new algorithms will have to be devised to get round the problem like treating one document type at a time.

For the entire process, we are following new leads to improve each stage from tagging to correcting.

For the time being, the process cannot learn from previous use, neither can it use external sources of information to help solve a problem (INIST has more then 10 million bibliographical records on-line and counting). This might also increase the efficiency of the system.

8. References

[1] E. Garfield, "Citation analysis as a tool in journal evaluation", *Science*, 178 (4060), 1972, p. 471-479.

[2] H.G. Small, "Visualizing science by citation mapping", *J. Am. Soc. Inform. Sci.*, 50 (9), 1999, p. 799-813.

[3] S. Lawrence, C. L. Giles, K. Bollacker, “Digital Libraries and autonomous Citation indexing”, *IEEE Computer*, 32 (6), 1999, p. 67-71.

[4] L. Van Guilder,. “Automated Part of Speech Tagging : A Brief Overview”, http//www.georgetoown.edu /cball/Ling361/tagging_overview.html, 1997.

[5] Brill, Eric . 1992. A simple Rule-Bases Part of Speech Tagger, In Proceedings of the third Annual Conference on Applied Natural Language Processing, ACL, 1992.

[6] A. Belaïd, “Recognition of Table of Contents for Electronic Library Consulting”. *International Journal on Document Analysis and Recognition*. 2001. 4 (1), p. 35-45.