A Supervised Approach for Detecting Allusive Bibliographical References in Scholarly Publications

Anaïs Ollagnier, Sébastien Fournier, Patrice Bellot

A Supervised Approach for Detecting Allusive Bibliographical References in Scholarly Publications

Anaïs Ollagnier
Laboratoire des sciences de l'information et des systèmes
Marseille, France
anais.ollagnier@lsis.org

Sébastien Fournier
Laboratoire des sciences de l'information et des systèmes
Marseille, France
sebastien.fournier@lsis.org

Patrice Bellot
Laboratoire des sciences de l'information et des systèmes
Marseille, France
patrice.bellot@lsis.org

ABSTRACT

Exploiting the links between content is crucial in recommendation approaches. In the case of a scientific article library, bibliographic references serve as a major link source. Among them, some are explicit references as we can find at the end of articles or books, while other references are scattered in the text or in the footnotes, according to a more or less strong implicit degree. We propose to focus on the detection of this type of references that we call allusive, in scientific articles from the field of Human and Social Sciences. To overcome the inherent difficulties raised by such reference detection, we present a method which aims at (i) identifying paragraphs that contain references via a classification process and (ii) at applying CCRFs (Cascaded Conditional Random Field) in order to detect more accurately the bibliographic entries and consequently annotate their contents.

CCS Concepts

•**Information systems** → **Recommender systems; Bibliographical references detection;** *Bibliographical references analysis;*

Keywords

bibliographical references detection; supervised classification; cascaded conditional random field.

1. INTRODUCTION

The use of bibliographical references is a crucial element in the creation and dissemination of information [4]. In recent years, bibliographical reference detection has become an important task especially in scientific context. The volume of scientific research has increased significantly and it is so complex and specialized that personal knowledge and experience are no longer sufficient tools for understanding trends or for making decisions. In our work, we decide to use bibliographical reference detection in part of the implementation framework of a reading recommendation system.

WIMS'16 June 13-15, 2016, Nîmes, France

ISBN 978-1-4503-4056-4/16/06…$15.00

DOI: http://dx.doi.org/10.1145/2912845.2912883

Usually, in reading recommendation systems, the document content characterization use solely information that can be derived from the respective theme. Document selection is based on a comparison between the topics covered in the documents related to the topics of interest for the user. In the case of a scientific article library, we find many bibliographic references that can be used as a basis for linking the documents directly related to targeted content. The work presented in this article focuses on the detection of references and, particularly, of the allusive references scattered through the text, that are incomplete and non formatted. These aspects are influenced by how implicit or elusive the reference are. Indeed, allusive references can be more or less explicit, meaning that we can find them scattered throughout the text or in the form of quotations that are conventionally standardized through with respect to the publication type. These considerations make the identification difficult when conventional reference analysis methods are employed.

Our contribution fits around implementing a methodology dedicated to detect, extract and annotate these allusive references. This methodology consists, firstly, to identify paragraphs that contain references via a classification process and, secondly, in the application of CCRFs to more accurately detect the bibliographic entries and annotate their content. The originality of our contribution is to exploit allusive references to establish links between documents as part of the realization of a recommendation reading system. Moreover, in our work, we deal with Social and Human Sciences data (SHS), which is a challenge because, in addition to treating a diverse multidisciplinary field, the presence of many structural variations for references makes the homogenization treatment difficult.

Our paper is structured as follows: we presents the related work in reference parsing as well as a brief introduction of freely accessible online reference parsing tools in Section 2. Section 3 describes our approach. The next section focuses on data set presentation employed to validate our approach. Section 5 presents the experiments and the results.

2. RELATED WORK

In an automated reference parsing system, a reference is considered as a sequence string composed by different fields to be recognized like author, title, date, etc. Many works have addressed this task, we can classify these works into three main approaches: regular expressions based on heuristics [5], learning algorithms [1] and knowledge-based systems [8]. With widely used techniques for reference analysis, many tools have been developed. Most of these tools are

based on machine learning algorithms, especially on CRFs. ParsCit[1] [3], based on CRFs, provides a toolkit that allows general reference annotation in bibliographic areas. Biblio Citation Parser[2] was developed simultaneously with ParsCit and it is also designed for references at the end of articles. Freecite[3] is also inspired by ParsCit, and uses CRFs with implementation of CRF++ libraries. Grobid[4] [7] allows the extraction and analysis of CRFs-based references. An interesting point is that this tool can enhance the annotation through external data. Bibpro[5] [2] captures the structural properties and transforms properties in a sequence model. Its model is based on the sequence alignment, learning algorithms and knowledge-based systems. Bilbo[6] [6] allows the annotation of references that are present in bibliographic areas and in the footnotes. The models are based both on CRFs and on Support Vector Machine (SVM), for footnote classification. Bilbo is deployed on the collections of OpenEdition.org, a digital library dedicated to Humanities and social sciences. Other tools like, Pdf-extract[7] identifying and extracting semantically significant regions of a scholarly journal article (or conference proceeding) PDF. The pdf-extract tool uses a similar "visual" technique to identify semantically important areas of a PDF. After identifying semantically significant regions of text, it uses a set of heuristics to analyze certain "traits" in each region which help the tool understand what that region is doing. We find, after studying these different tools, that none of them allows the annotation of the allusive references scattered in the body text and missing in the global final list of references.

3. METHODOLOGY

Firstly, this section presents, the classification model put in place to detect areas potentially containing references and secondly, the various CRFs established for the detection of bibliographic entries and annotation of their contents.

3.1 Supervised classification of paragraphs containing bibliographical references

In view of the large amount of paragraphs having no bibliographical references we decide to perform a pre-filtering through the use of a supervised classification. One for each category in order to establish two SVM models specific to each. We define for, each corpus, two classes: "bibliographic field" versus "no bibliographic field". The "Short references" category contains 20.8% paragraphs with bibliographic entries from a total of 725 paragraphs. "Nested references" category counts 23.3% paragraphs with bibliographic entries from a total of 1342 paragraphs. For the SVM implementation, we use the tool *SVMLight*[8]. Regarding the settings, we make a list of the most characteristic words of each class we use as attributes. This list is performed by the algorithm *InfoGainAttribute* (IGA) which reduces a bias of multi-valued attributes. After several tests, we decided to use 1 as a minimum occurrence frequency of terms combined with a list from which we have removed the words with score "*Recursive Feature Elimination*" (RFE) is equal to 0. For each subcategory, we conduct 10-fold cross-validations to assess how the results generalize to a set of data. The table 1 presents the results obtained from the classification for each of the subcategories. Regarding the classification performance, for the "Short references" category, we obtain an "accuracy" of 74.8% on the test set, which corresponds to 89 references classified correctly and 30 references classified as incorrect. For the "Nested references", we get a "accuracy" of 79.0% on the test set, which corresponds to 156 references classified correctly and 44 references classified as incorrect.

Corpus name	Accuracy	Precision	Recall
Short references	74.8%	71.0%	83.1%
Nested references	79.0%	88.6%	71.2%

Table 1: Results for supervised classification

3.2 Annotating bibliographical references with CCRFs

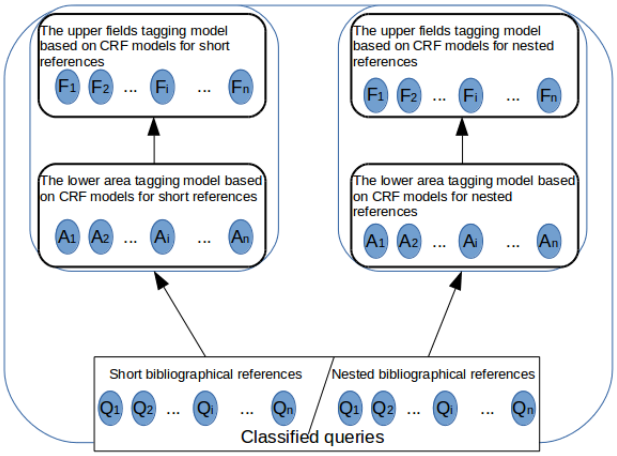

Figure 1: Annotating algorithm based on CCRFs

As shown in Figure 1, we use the result of the classification which provides a list of paragraphs containing potentially bibliographical references. Once queries sorted in predefined classes, we use an algorithm based on CCRFs. For each class, lowers CRFs perform a first filtering in order to identify the zone in which the different bibliographical fields are located. This first CRF reduces the execution field of the second CRF via the affixing of the <bibl> around the sequence containing bibliographic fields. Then, bibliographical references areas results are transferred to the upper CRF models. The upper CRF models annotate fields within bibliographical reference. Each CRF models are learned on specific corpus which are constructed on each class.

4. TEST COLLECTION

In our work, we use data provided by the *Centre pour L'édition Électronique Ouverte* (CLEO) extracted from OpenEdition Portal[9]. OpenEdition is an online platform for elec-

[1]http://aye.comp.nus.edu.sg/parsCit/
[2]http://paracite.eprints.org/developers/
[3]http://freecite.library.brown.edu/welcome
[4]https://github.com/grobid/grobid
[5]https://github.com/ice91/BibPro
[6]https://github.com/OpenEdition/bilbo
[7]http://labs.crossref.org/pdfextract/
[8]http://svmlight.joachims.org/
[9]http://www.openedition.org/

tronic articles, books, science blogs in the field of SHS. The field of SHS is a multidisciplinary field that allows us to have a representation of the vast majority of types of bibliographic references formats. Indeed, the conventions used in this field are very diversified compared to those used in "hard" science. Although there are more widely used conventions, such as the *American Psychological Association*[10] or the *Modern Language Association*[11], many organizations establish their own convention to meet their needs.
The data we use comes from the Solr index OpenEdition format XML. As part of a R&D program of OpenEdition laboratory, three corpora were manually annotated following the TEI[12] formalism. Our corpora are built from the interrogation of the Solr index of OpenEdition. The first corpus is for the learning models for references present in bibliographic areas. The second corpus focuses solely on the footnotes. Finally, the third corpus consists of allusive references. All these corpora are established solely for articles taken from the Revues.org platform. To form the corpus allusive references, the selected journals were chosen randomly in the OpenEdition database regardless of the disciplinary affiliation of the journal or the style of the reference. // As part of our work, we are interested solely in the third corpus[13]. This corpus is divided into two categories, namely:

- **Short references**: composed of 449 references, this category refers to references composed of very few elements.

```
C'est ce que j'avais envie de faire tout le temps
<bibl>(<title>La Terrasse</title>,<date>octobre 2008</date>
, <abbr>n°</abbr> <biblScope>161</biblScope> :
<biblScope>54</biblScope>)</bibl>
```

Figure 2: Example of Short references

- **Nested references**: composed of 553 references, this category consists of references that are "implicit", that is to say, they are expressed informally.

```
Il suffirait de comparer, à la fin de l'édition en volume de
<bibl><title level="m">From Hell</title>, les longues et
méticuleuses notes d'<author><forename>Alan</forename>
<surname>Moore</surname></author></bibl> avec la drolatique
historiographie dessinée des spécialistes de Jack l'Éventreur.
```

Figure 3: Example of Nested references

5. EXPERIMENTATIONS AND RESULTS

In the next section, the different experiments conducted as well as the performance obtained when detecting bibliographical areas and annotation of their contents. To measure performance, we use micro-averaged f-measure (mf), micro-averaged precision (mp) and recall (mr).

5.1 Detection of bibliographical references areas

This section presents the results obtained from affixing the tag <bibl> around sequences containing bibliographic fields. To perform our experiments, we establish two corpus. The first corpus consists solely of paragraphs containing references (Gold Corpus), namely, this corpus is established independently of the previously presented classification. This gives us for the "Short references" 158 paragraphs containing references and for the "Nested references" 311 paragraphs containing references. The second corpus derives from the classification made in the previous experiment (Classify Corpus), namely, this corpus contains solely the paragraphs to which the classification model has awarded the bibliography field class. This gives us for the "Short references" 89 paragraphs containing references and for the "Nested references" 156 paragraphs containing references. All misclassified paragraphs provide a negative weight in the evaluation. The study of the results between these two corpus allows us to assess the impact of the classification performance and to assess the degree of difficulty in detecting bibliographic entries of each of our subcategories. We present experiments based on 10 cross validations composed of 70 % of the training corpus and 30 % of test corpus to avoid over-learning due to excessive amount of training data.

Corpus name	Gold Corpus		
	mp	mr	mf
Short references	94.1%	93.4%	93.8%
Nested references	70.2%	72.6%	71.7%

Table 2: Results for detection of references areas for Gold Corpus

Table 2 allows us to observe a significant difference in performance between our two subcategories. The best performance is obtained by the category "Short references" which can be explained on one hand, by a more formal structure at the onset of this type of references (often in brackets) and on the other hand, by an implicit degree much less important than the "Nested references" category that is often present in a larger structure in which its elements are scattered throughout the text.
Table 3 shows, as for Gold Corpus, a clear difference in performance between our two subcategories. The best performance is also noted for the "Short references". The phenomena we report previously to explain the substantial differences in performance between the two sub categories also apply in this case.

Corpus name	Classify Corpus		
	mp	mr	mf
Short references	89.6%	62.9%	73.6%
Nested references	52.3%	49.6%	50.9%

Table 3: Results for detection of references areas for Classify Corpus

Overall, we find that the best performance on Gold Corpus on our two subcategories. These results allow us to evaluate the significant impact of the classification of paragraphs on performance. We can also observe that the category "Nested references" gets weaker performance in our two corpus which can raise the difficulty to detect references whose structure and whose implicit degree is strong. In addition to the difficulties inherent to the characteristics of each of our subcategories, the concealment of certain paragraphs containing references has a strong impact on performance.

[10]http://www.apa.org/
[11]https://www.library.cornell.edu/research/citation/mla
[12]http://www.tei-c.org/index.xml
[13]http://lab.hypotheses.org/
Write to marin.dacos@openedition.org to access to collection.

5.2 Detection of fields within bibliographical references

In this section, we present the results obtained from the identification of various bibliographic fields. Just like the previous experiment we establish two corpus. The first corpus consists solely of paragraphs containing references (Gold Corpus). The second corpus derives from the classification by SVM (Classify Corpus). We emphasize that this experiment is done independently of the results obtained from the detection of bibliographical references areas. The corpus used do not stem directly from the results presented in the previous experiment. This experiment is carried out independently from the previous to test the quality of CRF and features that we define. We present experiments based on 10 cross validations composed of 70 % of the training corpus and 30 % of test corpus to avoid over-learning due to excessive amount of training data.

Table 4 present the results of Gold Corpus. We note that

Corpus name	Gold Corpus		
	mp	mr	mf
Short references	89.1%	88.8%	89.1%
Nested references	79.7%	65.6%	71.8%

Table 4: Results for detection of fields within references for Gold Corpus

the best performance on the category "Short references" with micro-average precision and micro-average recall substantially similar. We see weaker results for the "Nested references", this phenomenon is explained by the greater diversity of bibliographic fields that make up references but also by less standardized structures. We note in table 5, the same

Corpus name	Classify Corpus		
	mp	mr	mf
Short references	58.4%	55.2%	56.5%
Nested references	50.1%	37.1%	41.9%

Table 5: Results for detection of fields within references for Classify Corpus

behavior as Gold Corpus with significantly better performance for the "Short references". The assumptions to explain the performance gap Gold Corpus also apply, namely, that the differences in performance between the two subcategories are caused by different degrees of implicit but also by a composition of the field much less diversified to the "Short references" category.

Overall, the best performance are observed on Gold Corpus. We can note a marked deterioration in performance on our two subcategories between Gold Corpus and Classify Corpus. This experiment allows us to assess the impact of the classification performance. We can also note the significant difficulty in identifying different fields. This difficulty is particularly great in the category "Nested reference". We can explain this by the largest structural change in this category as well as a greater variation of the fields that compose it. For "Nested reference", we count an average of 3 fields in the structural composition and an average of 5 different fields which are the most frequent author and title. For "Short references" we count an average of 2 fields in the structural composition and an average of 3 different fields which are the most frequent author and date. In addition, less formal structures we meet in the category "Nested references" makes homogenization difficult processes. This experiment allows us to understand that more the degree of implicit is important, more the references are difficult to detect. It is also important to note that the use of micro average precision, recall and F-measure creates more severe results.

6. CONCLUSIONS

In conclusion, we observed the impact of the classification on performance with a marked deterioration on the two experiments. The results obtained during the classification quantify the impact of hidden references. Both experiments conducted, allow us to understand the difficulties in the detection of allusive references. Upon detection of bibliographical references areas, we noted the impact of the implicit degree of references in addition to the features inherent in each of our subcategories. We have also observed the same phenomenon when detection of fields within bibliographical references whose performance deteriorated substantially over the reference is implicit. In our future work, we intend to find another granularity that paragraphs in order to reduce the scope of CRFs. We plan to establish corpus in more representative way. We believe, both for each of the subcategories and for each task performed by the CRFs, select specific features to improve performance. Our goal is geared towards the realization of a complete processing line in which the we will integrate a multiclass SVM directly to guide the choice of the first CRF. We recall that the purpose of this work is the implementation of a reading recommendation system in which we intend to use the references to establish links between documents.

7. REFERENCES

[1] S. Anzaroot and A. McCallum. A new dataset for fine-grained citation field extraction. In *ICML Workshop*, 2013.

[2] C.-C. Chen, K.-H. Yang, H.-Y. Kao, and J.-M. Ho. Bibpro: A citation parser based on sequence alignment. *Knowledge and Data Engineering, IEEE Transactions on*, 24(2):236–250, 2012.

[3] I. G. Councill, L. C. Giles, and M.-Y. Kan. Parscit: an open-source crf reference string parsing package. LREC, 2008.

[4] B. Cronin. *The Citation Process: The Role and Significance of Citations in Scientific Communication*. Taylor Graham, London, 1984.

[5] I.-A. Huang, J.-M. Ho, H.-Y. Kao, and W.-C. Lin. Extracting citation metadata from online publication lists using blast. In *Advances in Knowledge Discovery and Data Mining*, pages 539–548. Springer, 2004.

[6] Y.-M. Kim, P. Bellot, E. Faath, and M. Dacos. Annotated bibliographical reference corpora in digital humanities. LREC, 2012.

[7] P. Lopez. Grobid: Combining automatic bibliographic data recognition and term extraction for scholarship publications. In *Research and Advanced Technology for Digital Libraries*, pages 473–474. Springer, 2009.

[8] B. A. Rezaei and A. H.-Y. M. Muntz. System and method for context-based knowledge search, tagging, collaboration, management, and advertisement, 2013.