Annotated Bibliographical Reference Corpora in Digital Humanities

Young-Min Kim **Patrice Bellot** **Elodie Faath, Marin Dacos**

LIA, University of Avignon
84911 Avignon France
young-min.kim@univ-avignon.fr

LSIS, Aix-Marseille University
13397 Marseille France
patrice.bellot@lsis.org

CLEO, Centre for Open Electronic Publishing
13331 Marseille, France
{elodie.faath, marin.dacos}@revues.org

Abstract

In this paper, we present new bibliographical reference corpora in digital humanities (DH) that have been developed under a research project, *Robust and Language Independent Machine Learning Approaches for Automatic Annotation of Bibliographical References in DH Books* supported by *Google Digital Humanities Research Awards.* The main target is the bibliographical references in the articles of Revues.org site, an oldest French online journal platform in DH field. Since the final object is to provide automatic links between related references and articles, the automatic recognition of reference fields like author and title is essential. These fields are therefore manually annotated using a set of carefully defined tags. After providing a full description of three corpora, which are separately constructed according to the difficulty level of annotation, we briefly introduce our experimental results on the first two corpora. A popular machine learning technique, Conditional Random Field (CRF) is used to build a model, which automatically annotates the fields of new references. In the experiments, we first establish a standard for defining features and labels adapted to our DH reference data. Then we show our new methodology against less structured references gives a meaningful result.

Keywords: Bibliographical reference, Automatic annotation, Digital Humanities, Bilbo, Conditional Random Field, TEI

1. Introduction

In this paper, we present new bibliographical reference corpora in digital humanities area. The corpora have been developed under a research project, *Robust and Language Independent Machine Learning Approaches for Automatic Annotation of Bibliographical References in DH(Digital Humanities) Books* supported by *Google Digital Humanities Research Awards.* It is a R&D program for in-text bibliographical references published on CLEO's OpenEdition platforms[1] for electronic articles, books, scholarly blogs and resources in the humanities and social sciences. The program aims to construct a software environment enabling the recognition and automatic structuring of references in academic digital documentation whatever their bibliographic styles (Kim et al., 2011).

Most of earlier studies on bibliographical reference annotation are intended for the bibliography part at the end of scientific articles that has a simple structure and relatively regular format for different fields. On the other side, some methods employ machine learning and numerical approaches, by opposite to symbolic ones that require a large set of rules that could be very hard to manage and that are not language independent. Day et al. (2005) cite the works of a) Giles et al. (1998) for the CiteSeer system on computer science literature that achieves a 80% accuracy for author detection and 40% accuracy for page numbers (1997-1999), b) Seymore et al. (1999) that employ Hidden Markov Models (HMMs) that learn generative models over input sequence and labeled sequence pairs to extract fields for the headers of computer science papers, c) Peng and McCallum (2006) that use Conditional Random Fields (CRFs) (Lafferty et al., 2001) for labeling and extracting fields from research paper headers and citations. Other approaches employ discriminatively trained classifiers such as Support Vector Machine (SVM) classifiers (Joachims, 1999). Compared to HMM and SVM, CRF obtained better labeling performance.

The main interest of our project is to provide automatic links between related references, articles and resources in OpenEdition site, which is composed of three different sub-platforms, Revues.org, Hypotheses.org and Calenda. The automatic link creation involves essentially automatic recognition of reference fields, which consist of author, title and date etc. Based on the correctly separated and recognized fields, different techniques can be applied for the creation of cross-links. The initial work of this project mainly consists of the corpora construction, especially the manual annotation of reference fields. This is concerned with a detailed analysis of target data in OpenEdition. We start with Revues.org journal platform because it has the most abundant resources in terms of bibliographic references. Faced with the great variety of bibliographical styles present on the three platforms and the dissemination of references within texts, we have implemented a series of stages corresponding to the various issues encountered on the platforms. In the paper, we first detail the nature of Revues.org data that justifies our methodology, then describe the corpora construction process and finally we discuss the experimental results.

In brief, we construct three different types of corpus with a detailed manual annotation using TEI guidelines. They will be a new valuable resource for research activities in natural language processing. There is no equivalent resource to date, neither in size nor in diversity.

2. Revues.org document properties

Revues.org is the oldest French platform of online academic journals. It now offers more than 300 journals available in all disciplines of the humanities and social sciences, with predominance of history, anthropology and sociol-

[1] http://www.openedition.org

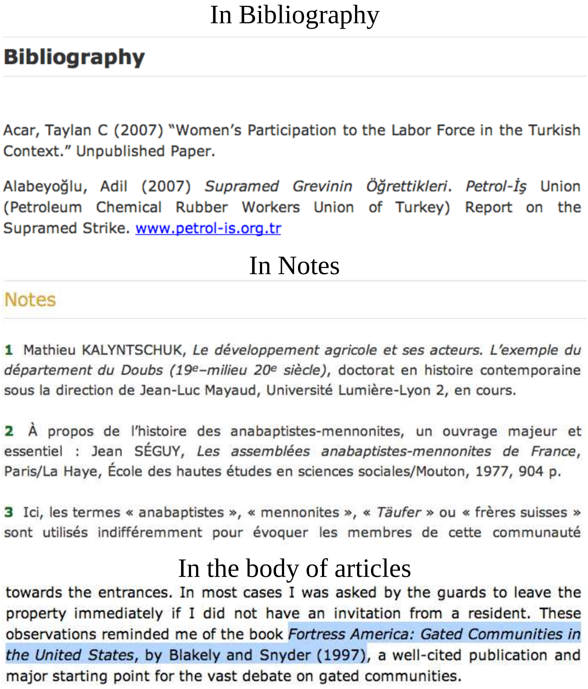
ogy, geography and archaeology. The original language is French but it has been designed for the electronic publishing on an international scale. About 10% of articles are in a different language besides French. Beyond the commitment in favor of open access (more than 40,000 articles in open access), the platform is based on a model of appropriation of the electronic publishing process by publishers and producers of content. The online publication is made through the conversion of articles into XML TEI format and then into XHTML format and allows the viewing of the full text in web browsers. The specific technical quality needed for the publishing of scientific texts is provided by many functions: metadata management, multiple indexes, management of endnotes, automatic table of contents, numbering of paragraphs and attribution of DOI.
Apart from the well organized technical functions, the bibliographical reference parts of the articles on Revues.org are rather diverse and complicated compared to that of scientific research papers. One main reason of this complexity is the diversity of source disciplines that makes various styles in reference formatting. Moreover, even on a same discipline or journal, we can easily find quite different reference forms caused by the absence of a strict format recommended. Another important difficulty is also from the irregularity of reference part that sometimes arises in footnote or body of articles. Especially, reference in the latter case usually has no particular form but is just integrated in a sentence such that even segmentation of bibliographic part is not easy.
Considering these difficulties that occur depending mainly on the physical position of reference, we divide bibliographical references into three different types as in Figure 1. The first type of references are located at the end of article under a heading “Bibliography”, “Citation’, etc. They are traditional targets of bibliographical reference treatment (Giles et al., 1998; Peng and McCallum, 2006; Councill et al., 2008). The second type of references are found in footnotes (henceforth called notes) of article and are less formulaic compared to the first type. A typical particularity in this type is that note can have non-bibliographical text such as adjective phrases before a citation. The third type includes partial bibliographical references found in the body of articles. It is the most difficult type for both manual and automatic annotations. Even finding the beginning and end of a suitable bibliographical reference is difficult.

3. Manual annotation of Corpora

In this section, we detail the manual annotation process of our corpora. Against the difficulties introduced above, we first well define TEI XML tags to tag the bibliographic parts, then construct three different corpora according to the type of reference. We try to take into account the specificity of target data as well as the generality of the digital humanities area.

3.1. TEI and tags for manual annotation

TEI is a consortium that develops, defines and maintains a markup language for describing structural, renditional and conceptual features of texts. And in our case, TEI guidelines are used for describing the fields of bibliographic references. There are three possible levels of description for this kind of information :



In Bibliography

Bibliography

Acar, Taylan C (2007) “Women’s Participation to the Labor Force in the Turkish Context.” Unpublished Paper.

Alabeyoğlu, Adil (2007) *Supramed Grevinin Öğrettikleri*. *Petrol-İş* Union (Petroleum Chemical Rubber Workers Union of Turkey) Report on the Supramed Strike. www.petrol-is.org.tr

In Notes

Notes

1 Mathieu KALYNTSCHUK, *Le développement agricole et ses acteurs. L’exemple du département du Doubs (19^{e}–milieu 20^{e} siècle)*, doctorat en histoire contemporaine sous la direction de Jean-Luc Mayaud, Université Lumière-Lyon 2, en cours.

2 À propos de l’histoire des anabaptistes-mennonites, un ouvrage majeur et essentiel : Jean SÉGUY, *Les assemblées anabaptistes-mennonites de France*, Paris/La Haye, École des hautes études en sciences sociales/Mouton, 1977, 904 p.

3 Ici, les termes « anabaptistes », « mennonites », « *Täufer* » ou « frères suisses » sont utilisés indifféremment pour évoquer les membres de cette communauté

In the body of articles

towards the entrances. In most cases I was asked by the guards to leave the property immediately if I did not have an invitation from a resident. These observations reminded me of the book *Fortress America: Gated Communities in the United States*, by Blakely and Snyder (1997), a well-cited publication and major starting point for the vast debate on gated communities.

Figure 1: Different types of bibliographic references

- <bibl> : for all bibliographical elements.
- <biblStruct> : it structures the reference with predefined elements and it can be found on other electronic archives such HAL and TEL.
- <biblFull> : it uses only elements allowed under <fileDesc>.

In our corpus, we use the standard description <bibl> to freely tag references. Indeed, OpenEdition presents a variety of bibliographic styles that <biblStruct> or <biblFull> can not describe. Another reason is that this standard description can be adapted for special references such as the case of inclusion or to indicate a working paper or published in a forthcoming scientific event.
Table 1 lists the defined tags for the manual annotation of our reference corpora. We try to encode as much information as possible in case of the reuse of the corpora for other objectives. Let us elaborate the ‘Author or Editor’ row, which shows an important particularity of our way. We tag author name with <surname>, <forename> and <author> tags that the first two are always wrapped by the last one. The editor name is tagged in the same manner. There are two main distinctions between our annotation system and the traditional ones. First, we separate different authors and even author’s surname and forename. In traditional approaches, different authors in a reference are tagged as a field and there are no separation of surname and forename of course. Meanwhile, our detailed separation can facilitate the automatic extraction of each author name that is essential to make useful cross-links. Moreover,

Table 1: Defined tags for manual annotation

Type	sub-type	tag name	@attribute value
Reference	Reference	<bibl>	
	Included Ref.	<relatedItem>	
Author or Editor	Author	<author>	
	Editor	<editor>	@role editor
			translator
	Surname	<surname>	
	Forename	<forename>	
	Gen. name	<genName>	
	Name link	<nameLink>	
	Org. name	<orgName>	
Title	Title	<title>	@level a (article)
			j (journal)
			m (monograph)
			u (unpublished work)
			s (series)
	Confernce or Symposium	<meeting>	
Publication Mark	Date	<date>	
	Place	<pubPlace>	
		<settlement>	
		<country>	
		<region>	
	Publisher	<publisher>	
		<distributor>	
		<sponsor>	
	Edition	<edition>	
	Page extent	<extent>	
	Edition detail	<biblScope>	@type vol (volume)
			pp (pages)
			issue (journal num.)
			issn
			part
Punctuation	Punctuation	<c>	@type point
			comma
			(other punctuation marks)
Pre-citation	Special term	<w>	
	Link between references	<link>	
Etc.	Abbreviation	<abbr>	@type contraction
			acronym
	Web page	<ref>	

when the occurring position of author or editor field is flexible as in note data, this separation would be more useful. We expect that by identifying person name with surname and forename instead of author and editor, the specificity of each field will be strengthened in learning process then the automatic annotation will become easier. Second distinction is that a name token is attached by two different but hierarchical tags. We allow this kind of multi-tagging in our manual annotation that enables rich information encoding. But in our current automatic annotation system, we estimate just a single label to a token.

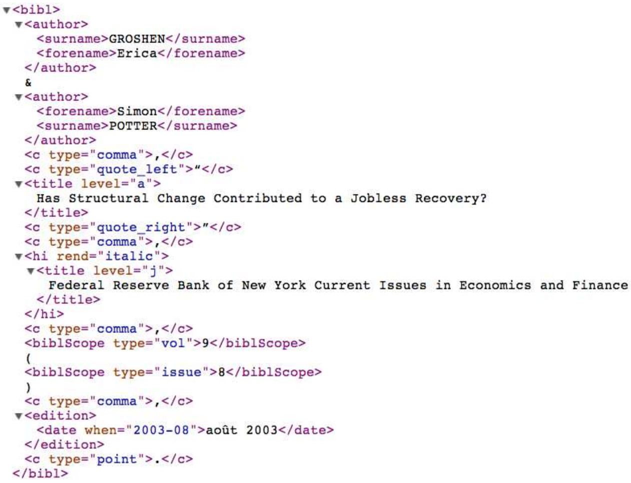
There are many other distinguishable aspects compared to the traditional methods. Titles are classified into five different categories : article, journal, monograph, unpublished work, and series. We also detailed place with four different tags, pubPlace, settlement, country and region. Publisher is tagged with publisher, distributor, and sponsor. Another distinguishable strategy is concerned with the treatment of punctuation. The annotator, a specialist in a humanities-related field, have annotated the punctuation marks, which play a role for the separation of reference fields, with the tag <c> . Finally, note that we introduce an important tag <w>, which signifies that the tagged word is a special term or expression indicating a previously cited reference. The tag is exclusive to note data, and the most frequent terms wrapped by this tag are 'Ibid.', 'op. cit.', 'ouv. cité', 'supra', etc.

3.2. Three corpora with different difficulty levels

Corpus construction starts with selecting some representative references from the Revues.org site. To keep diversity of bibliographic reference formats of various journals published on Revues.org, we try to select only one article for a specific journal. As pointed out in Section 2., three corpora have been constructed according to the difficulty level of annotation identified by type of reference as illustrated in Figure 1.

Corpus level 1

We first construct the corpus level 1, which is relatively simple than the others, however needs the most prudent annotation because it offers the standard for the construction of next corpora. Considering the diversity, 32 journals are randomly selected and 38 sample articles are taken. Total 715 bibliographic references are identified and recognized using TEI guidelines. Figure 2 shows an example of the corpus level 1. All present tags in this figure are explained in Table 1 except the <hi> tag with an attribute value of 'italic'. It is a default element provided by TEI guidelines that is used to mark words with emphasis. In the example, the journal name is highlighted by italic characters. There are other tags used for emphasis but only italic characters are considered as valid ones for the reference annotation.

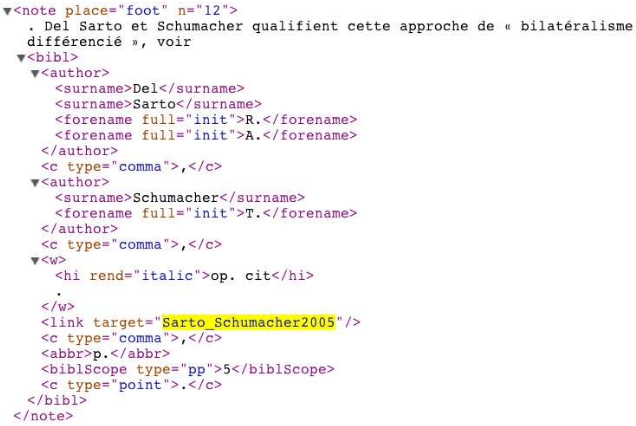


```
▼<bibl>
  ▼<author>
     <surname>GROSHEN</surname>
     <forename>Erica</forename>
   </author>
   &
  ▼<author>
     <forename>Simon</forename>
     <surname>POTTER</surname>
   </author>
   <c type="comma">,</c>
   <c type="quote_left">"</c>
  ▼<title level="a">
     Has Structural Change Contributed to a Jobless Recovery?
   </title>
   <c type="quote_right">"</c>
   <c type="comma">,</c>
  ▼<hi rend="italic">
    ▼<title level="j">
       Federal Reserve Bank of New York Current Issues in Economics and Finance
     </title>
   </hi>
   <c type="comma">,</c>
   <biblScope type="vol">9</biblScope>
   (
   <biblScope type="issue">8</biblScope>
   )
   <c type="comma">,</c>
  ▼<edition>
     <date when="2003-08">août 2003</date>
   </edition>
   <c type="point">.</c>
 </bibl>
```

Figure 2: An example of reference in corpus level 1

Corpus level 2

In our second level of corpus, the target references are located in notes. We annotate references using the same tags to the first corpus except <w>. Recall that notes contain some special terms marked as <w> indicating a previously cited reference. Some references including <w> tag are shorten including just essential parts such author name, but sometimes are linked to another reference, which has more detailed information about the shorten ones. This case often occurs when a bibliographic document is referred more than once. To make a link between two references citing identical source, we first add an identifier to the original one in its <bibl> tag, then add this identifier to the recited reference using a <link> tag as in Figure 3. This figure shows



```
▼<note place="foot" n="12">
   . Del Sarto et Schumacher qualifient cette approche de « bilatéralisme
   différencié », voir
  ▼<bibl>
    ▼<author>
       <surname>Del</surname>
       <surname>Sarto</surname>
       <forename full="init">R.</forename>
       <forename full="init">A.</forename>
     </author>
     <c type="comma">,</c>
    ▼<author>
       <surname>Schumacher</surname>
       <forename full="init">T.</forename>
     </author>
     <c type="comma">,</c>
    ▼<w>
       <hi rend="italic">op. cit</hi>
       .
     </w>
     <link target="Sarto_Schumacher2005"/>
     <c type="comma">,</c>
     <abbr>p.</abbr>
     <biblScope type="pp">5</biblScope>
     <c type="point">.</c>
   </bibl>
 </note>
```

Figure 3: An example of reference in corpus level 2

a recited note example whose original reference identifier is ‘Sarto_Schumacher2005’.

Besides the above property, the corpus level 2 naturally has a segmentation issue for the extraction of exact bibliographical parts. Notes are originally intended for describing any supplementary information of a part of text. Therefore there can exist obviously non-bibliographical notes. Moreover, even a note having citation information is more freely written than formal references in corpus level 1, then it can probably have non-bibliographical phrases. We call this kind of problem a segmentation problem. Table 2 depicts several examples in this issue. The examples are extracted from same article[2] in a political sociology journal. Note no. 26 is an example without any citation in it whereas note no. 27 has two different references (in grey) separated by a short phrase. Note no. 31 has a non-bibliographic phrase at the beginning of note.

Table 2: Segmentation problem in the notes

	Examples
Non bibl note	26. La nature euro-centrée du projet était encore plus apparente dans la version originale du texte, qui, comme nous l’avons déjà mentionné, était appelé “ Europe élargie ”.
Multi bibl note	27. “ Une Europe sûre dans un monde meilleur. Stratégie européenne de sécurité ”, Bruxelles, 12 décembre 2003. Pour un commentaire critique, voir Toje A., “ The 2003 European security strategy:A critical appraisal ”, *European Foreign Affairs* Review, vol. 10, n^{o}1, 2005, pp. 117-134.
Part. info.	31. Voir par exemple la communication de la Commission relative au “ Renforcement de la politique européenne de voisinage ”, COM (2006) 726 final.

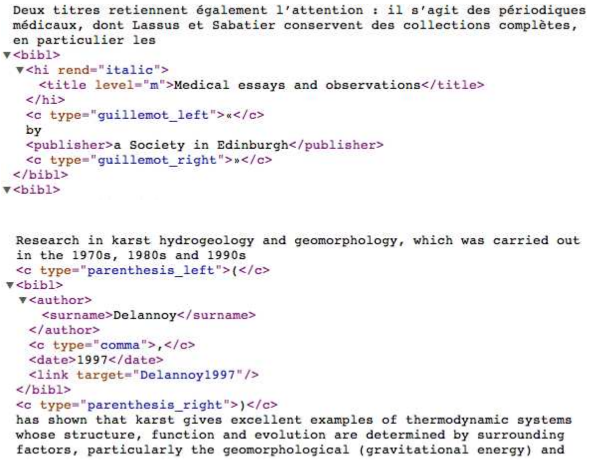
We select 41 journals from a stratified selection then choose 42 articles. Note that the selected articles reflect the proportion of two different note types where one includes bibliographic information while the other does not. Since the objective of the initiated project is totally automated annotation of bibliographical references, the detection of bibliographic note should be preceded before annotating notes. For this purpose, we design the corpus level 2 to be composed roughly by two groups: manually annotated reference notes similar to the corpus level 1 and non-bibliographical notes, which do not need any manual annotation. Figure 3 is an example of the first group. The beginning phrase, which is not part of reference, is just excluded from manual annotation. Consequently, we have 1147 bibliographical notes and 385 non-bibliographical notes.

Corpus level 3

Since the target of corpus level 3 is diffused throughout the body of article, manual annotation of this corpus is much more difficult than the previous corpora. Even if the basic tags have been already defined through the construction of corpus level 1 and 2, we need another standard to well annotate the third one. The most urgent problem is that we have no guidelines for accepting a phrase as a bibliographical reference. Therefore, we first observe in detail the nature of phrases, which seem to be an implicit reference. The same specialist who constructed the previous corpora again analyzes the implicit citations for this third level of corpus. We decide to examine not only the body of article but also the notes, because some notes having scattered reference fields had been ignored in the construction of corpus level 2. That is, any annotated reference in corpus 2 does not have an interrupted part, that is, non-bibliographical part.

After a careful analysis, we categorize implicit references into the following three sub-groups:

- The first group includes the implicit references located in the body of article. They can be also found in the notes of the article at the same time.
- The second group includes the implicit references composed by only the author name and date. They can be found either in the body of the article or in the notes of the article. The difference with similar references in corpus level 2 is that their original references are located in the bibliography part of article.
- The third group includes the implicit references located in the notes. The difference with the similar references in corpus level 2 is that they have an interrupted part annotated as non-bibliographical part.



```
 Deux titres retiennent également l’attention : il s’agit des périodiques
 médicaux, dont Lassus et Sabatier conservent des collections complètes,
 en particulier les
▼<bibl>
 ▼<hi rend="italic">
    <title level="m">Medical essays and observations</title>
  </hi>
  <c type="guillemot_left">«</c>
  by
  <publisher>a Society in Edinburgh</publisher>
  <c type="guillemot_right">»</c>
 </bibl>
▼<bibl>


 Research in karst hydrogeology and geomorphology, which was carried out
 in the 1970s, 1980s and 1990s
 <c type="parenthesis_left">(</c>
▼<bibl>
 ▼<author>
    <surname>Delannoy</surname>
  </author>
  <c type="comma">,</c>
  <date>1997</date>
  <link target="Delannoy1997"/>
 </bibl>
 <c type="parenthesis_right">)</c>
 has shown that karst gives excellent examples of thermodynamic systems
 whose structure, function and evolution are determined by surrounding
 factors, particularly the geomorphological (gravitational energy) and
```

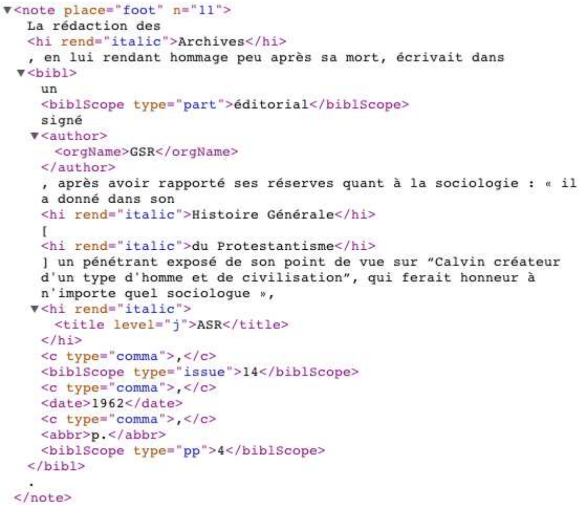
Figure 4: Implicit reference examples in the first (upper) and second (lower) groups of the corpus level 3

[2] http://conflits.revues.org/index2471.html

The upper example of Figure 4 is part of an article that contains an implicit reference of the first sub-group. It is well integrated in the content with the title and publisher information only. Sometimes, one or more fields of a bibliographical note are integrated in the body of the article. We also treat this case as an implicit reference in the first group. To make a link between the recognized implicit reference and its original note, we add an identifier as in the corpus level 2.

The lower example of the same figure contains an implicit reference composed by the author name and date. It is a typical form of a shorten reference in the second sub-group, which has its full description in another formal reference marked by an identifier, in this case, 'Delannoy1997' in the bibliography part. Sometimes this kind of short reference is only found at the notes part. We could have classified the latter case to the corpus level 2, but we decide to include it in the corpus level 3, because this reference can not be complete within the notes but needs the bibliography part in addition.

Figure 5 shows a note example, which includes an implicit reference of the third sub-group. The manually annotated bibliographical part begins right after the phrase 'écrivait dans' ('wrote in' in English). The reference fields are not continuous, instead interrupted by the words 'un' and 'signé', and by a phrase between <author> and <title>. It signifies that the segmentation problem becomes more complicated than the corpus level 2. We gather some expressions such as 'écrivait dans' that can be a sign of an implicit reference, expecting that we can find some useful patterns.



```
▼<note place="foot" n="11">
   La rédaction des
   <hi rend="italic">Archives</hi>
   , en lui rendant hommage peu après sa mort, écrivait dans
  ▼<bibl>
     un
     <biblScope type="part">éditorial</biblScope>
     signé
    ▼<author>
       <orgName>GSR</orgName>
     </author>
     , après avoir rapporté ses réserves quant à la sociologie : « il
     a donné dans son
     <hi rend="italic">Histoire Générale</hi>
     [
     <hi rend="italic">du Protestantisme</hi>
     ] un pénétrant exposé de son point de vue sur "Calvin créateur
     d'un type d'homme et de civilisation", qui ferait honneur à
     n'importe quel sociologue »,
    ▼<hi rend="italic">
       <title level="j">ASR</title>
     </hi>
     <c type="comma">,</c>
     <biblScope type="issue">14</biblScope>
     <c type="comma">,</c>
     <date>1962</date>
     <c type="comma">,</c>
     <abbr>p.</abbr>
     <biblScope type="pp">4</biblScope>
   </bibl>
   .
 </note>
```

Figure 5: Implicit reference example in the third group of the corpus level 3

There are many cases that can not be exactly classified to one of the sub-groups. And we are also faced with several practical problems such that different levels of references rise simultaneously in an article. The division between the corpus level 1 and level 2 is simple, whereas target area of the corpus level 2 and level 3 are somewhat overlapped.

Another important problem in corpus level 3 is to decide which elements are essential for an implicit reference. For that we suppose a situation that we should find a reference via a search engine. We can then select five different elements, 'title', 'author', 'date', 'place', and 'publisher' as necessary ones. A search with these five elements would give an accurate result. Besides, if we search a publication using author name without title, it would be difficult to obtain a desired result, even if other elements are also given. So, the main criteria to accept a phrase as an implicit reference is the possibility to find a result given information. That is why we accept a phrase including only title as an implicit reference.

For the corpus level 3, we select 34 articles considering the properties of implicit reference in the body of the articles and 8 articles having discontinuous reference notes. From these selected articles, we have 553 references of the first sub-group, 447 references of the second one, and 43 references of the third one.

4. Automatic annotation of reference

In this section we briefly introduce the main tool used for automatic annotation of reference fields. We use one of the most popular techniques in sequence annotation problem, Conditional Random Field (Lafferty et al., 2001; Peng and McCallum, 2006).

4.1. Conditional Random Fields

Automatic annotation can be realized by building a CRF model that is a discriminative probabilistic model developed for labeling sequential data. By definition, a discriminative model maximizes the conditional distribution of output given input features. So, any factors dependent only on input are not considered as modeling factors, instead they are treated as constant factors to output (Sutton and McCallum, 2011). This aspect derives a key characteristic of CRFs, the ability to include a lot of input features in modeling. The conditional distribution of a linear-chain CRF for a set of label **y** given an input **x** is written as follows :

$$p(\mathbf{y}|\mathbf{x}) = \frac{1}{Z(\mathbf{x})} \exp\{\sum_{k=1}^{K} \theta_k f_k(y_t, y_{t-1}, \mathbf{x}_t)\}, \qquad (1)$$

where $\mathbf{y} = y_1 ... y_T$ is a state sequence, interpreted as a label sequence, $\mathbf{x} = x_1 ... x_T$ is an input sequence, $\theta = \{\theta_k\} \in R^K$ is a parameter vector, $\{f_k(y_t, y_{t-1}, \mathbf{x}_t)\}_{k=1}^{K}$ is a set of real-valued feature functions, and $Z(\mathbf{x})$ is a normalization function. Instead of the word identity x_t, a vector $\mathbf{x}_t$, which contains all necessary components of **x** for computing features at time t, is substituted. A feature function often has a binary value, which is a sign of the existence of a specific feature. A function can measure a special character of input token x_t such as capitalized word. And it also measures the characteristics related with a state transition $y_{t-1} \rightarrow y_t$. Thus in a CRF model, all possible state transitions and input features including identity of word itself are encoded in feature functions. Inference is done by the Viterbi algorithm for computing the most probable labeling sequence, $\mathbf{y}^* = \arg\max_y p(\mathbf{y}|\mathbf{x})$ and the forward-backward algorithm for marginal distributions. It is used for the labeling of new input observations after constructing a model, and also applied to compute parameter values. Parameters are estimated by maximizing conditional log likelihood, $l(\theta) = \sum_{i=1}^{N} \log p(\mathbf{y}^{(i)}|\mathbf{x}^{(i)})$ for a given learning set of N samples, $D = \{\mathbf{x}^{(i)}, \mathbf{y}^{(i)}\}_{i=1}^{N}$.

4.2. Learning data

Recall that we try to encode as much information as possible during the manual annotation of corpora (see Table 1). However, the rich information is not always useful for automatic annotation. Unnecessarily too detailed output labels, which are the reference fields in our case, can complicate the learning process of a CRF model. Then it can produce a less exact annotation result than what we could obtain with a simple necessary labels. Therefore, choosing appropriate output labels is important before applying a CRF model. Meanwhile, the superiority of a CRF compared other sequence learning model comes from its capacity to encode properties of each input token through features. That is why the feature extraction is an essential process in CRF model. To avoid the confusion between input tokens and the features describing the characteristics of input, we call the latter local features.

Output labels and tokenization

We have 20 unique reference fields which are described in the uppermost table of Table 3. Instead of recognizing author and editor, we choose to tag them identically but with more detailed fields : surname and forename. Since author and editor are naturally separated by other fields such as title, we can easily divide them after an automatic annotation. Punctuation treatment is another important issue in the sequence labeling. In our approach, we detach all punctuation marks from words except several strongly attached marks such as hypen, and treat them as tokens.

Local features

The middle table of Table 3 shows the defined local features to express characteristics of each token. The first 11 features depict the external appearance of token, and the last three features encode the position of token in reference. Once we define the way of tokenization, output labels and local features, we can learn a CRF model with input tokens and these defined elements.

Global features

Global features are new type of features introduced for the processing of the corpus level 2. They describe a global pattern of input or local features of a reference string. For example, we can encode the pattern that the reference string starts with an initial expression to a global feature. The global features are invented to pick out non-bibliographical notes (e.g. note no. 26 in Table 2) from the corpus level 2. By eliminating them, we can learn a more accurate CRF model. In short, we classify the notes into bibliographical class and non-bibliographical class using a SVM classifier with note data represented by input, local, and global features. The global features in the final table of Table 3 well catch the distinguishable characteristics of two different classes.

Table 3: Labels and local features for learning data

Output field labels

Labels	Description
surname	surname
forename	forename
title	title of the referred article
booktitle	book or journal etc. where the article is published
publisher	publisher, distributor
biblscope	information about pages, volume, number etc.
date	date, mostly years
place	place : city, country, etc.
abbr	abbreviation
nolabel	tokens difficult to be labeled
edition	information about edition
bookindicator	the word 'in' or 'dans' when a related reference is followed
orgname	organization name
extent	total number of page
punc.	punctuation
w	terms indication previous citation (corpus level 2 only)
nonbibl	tokens of non-bibliographical part (corpus level 2 only)
OTHERS	rare labels such as genname, ref, namelink

Local features

Feature name	Description	Example
ALLCAPS	All characters are capital letters	RAYMOND
FIRSTCAP	First character is capital letter	Paris
ALLSAMLL	All characters are lower cased	pouvoirs
NONIMPCAP	Capital letters are mixed	dell'Ateneo
ALLNUMBERS	All characters are numbers	1984
NUMBERS	One or more characters are numbers	in-4
DASH	One or more dashes are included in numbers	665-680
INITIAL	Initialized expression	H.
WEBLINK	Regular expression for web pages	apcss.org
ITALIC	Italic characters	*Regional*
POSSEDITOR	Possible for the abbreviation of editor	ed.
BIBL_START	Position is in the first one-third of reference	-
BIBL_IN	Position is between the one-third and two-third	-
BIBL_END	Position is between the two-third and the end	-

Global features

Feature name	Description
NOPUNC	There are no punctuation marks in the reference string
ONEPUNC	There is just one punctuation mark in the reference string
NONUMBERS	There are no numbers in the reference string
NOINITIAL	The reference string includes no initial expressions
STARTINITIAL	The reference string starts with an initial expression

5. Experiments

The objective of our experiments is to establish a methodology to well estimate bibliographical reference fields. The primary experiments focus on finding an effective way of tokenization, a set of appropriate output labels, and useful local features for a CRF model. This work have been realized with corpus level 1. Once we set the standard on tokenization, output labels, and local features for CRF construction, we try other machine learning techniques to improve the automatic annotation result. The development of global features to eliminate non-bibliographical notes from the corpus level 2 is one of our unique attempts. After the elimination, we apply a CRF model with the remaining notes. In this section, we summarize the experimental results obtained until now (for a detailed result, see Kim et al. (2012))

For a CRF model construction, we used an existing language processing toolkit, MALLET software (McCallum, 2002). Elimination of non-bibliographical notes with SVM classifier is realized by a well-known implementation, SVM^{light} (Joachims, 1999). The automatic annotation result is evaluated with ground truth using precision and recall for each field. We count the number of well estimated tokens for each field to calculate them. Overall accuracy is computed by micro-averaged precision. We randomly split a corpus into learning and test data in the proportion of 7:3 for both CRF and SVM model respectively.

5.1. Primary evaluation of sequence annotation with corpus level 1

We have tested more than 40 different combinations of tokenization method, output labels, and local features. The labels and features in Table 3 are the finally selected ones. We detach all punctuation marks and special characters except hyphen. And finally we obtain about 90 % of overall accuracy on a test data as shown in Table 4. The most important fields are surname, forename and title in view of searching and making cross-links. The columns #true, #annot. and #exist. mean the total number of true, automatically annotated, and existing tokens for the corresponding field. Compared to the scientific research reference data used in the work of Peng and McCallum (2006), our corpus level 1 is much more diverse in terms of reference formats. However we have obtained a successful result in annotation accuracy, especially on surname, forename and title fields (92%, 90%, and 86% of precision respectively). They are somewhat less than the previous work of Peng (95% overall accuracy) but considering the difficulty of our corpus, the current result is quite encouraging.

Table 4: Bibliographical reference field annotation performance of a CRF model learned with corpus level 1

	PRECISION			RECALL		
Fields	#true	#annot.	prec.(%)	#true	#exist.	recall(%)
surname	305	331	**92.15**	305	341	**89.44**
forename	308	342	**90.06**	308	339	**90.86**
title	1911	2199	**86.90**	1911	2034	**93.95**
booktitle	252	352	71.59	252	469	70.41
publisher	316	387	81.65	316	373	84.72
biblscope	109	130	83.85	109	140	77.86
date	245	273	89.74	245	258	94.96
place	153	179	85.47	153	169	90.53
abbr	122	144	84.72	122	138	88.41
nolabel	71	106	66.98	71	100	71.0
edition	10	18	55.56	10	71	14.08
bookindicator	26	28	92.86	26	29	89.66
orgname	18	19	94.74	18	42	42.86
extent	29	29	100.0	29	31	93.55
punc.	2014	2027	99.36	2014	2024	99.51
OTHERS	5	5	100	5	11	45.45
Average	5894	6569	**89.72**	5894	6569	**89.72**

5.2. Sequence classification and annotation with corpus level 2

Experimental process on the corpus level 2 consists of two steps. First step is the classification of note data into bibliographical and non-bibliographical categories. Similar to the corpus level 1, we try a number of combinations of different input, local, and global features to obtain one of the most effective feature set for SVM classification. The finally chosen features are input words, punctuation marks, four different local features (posspage, weblink, posseditor, and italic), and five different global features in the final table of Table 3. Then in the second step, we learn a CRF model with the classified notes only into bibliographical category. In addition to the ourput labels used in the corpus level 1, 'w' and 'nonbibl' are introduced. But we do not use the position features this time because the scattered reference fields attenuate the effect of position then rather decrease the annotation accuracy.

Table 5: Bibliographical note field annotation performance of a CRF model learned with corpus level 2

	PRECISION			RECALL		
Fields	#true	#annot.	prec.(%)	#true	#exist.	recall(%)
surname	378	474	**81.22**	385	501	**76.82**
forename	360	440	**81.86**	360	460	**78.30**
title	2991	3634	**82.37**	2991	3465	**86.28**
booktitle	257	376	68.76	257	599	43.00
publisher	399	539	73.99	399	530	75.14
biblscope	416	471	88.38	416	481	86.65
date	391	432	90.52	391	433	90.21
place	204	231	88.81	204	228	89.67
abbr	419	454	92.31	419	444	94.36
w	215	222	97.06	215	233	92.38
nolabel	64	95	66.94	64	226	29.27
edition	11	25	44.38	11	59	21.74
bookindicator	40	42	94.05	40	53	74.90
orgname	15	21	72.22	15	35	44.03
extent	15	21	72.11	15	21	72.71
punc.	3111	3274	95.01	3111	3371	92.30
nonbibl	3133	4056	77.25	3133	3626	86.35
OTHERS	2	13	15.38	2	40	5.0
Average	12428	14820	**84.00**	12428	14820	**84.00**

Table 5 shows the final automatic annotation result. It is an averaged result of five CRF models learned with different splits of learning and test set for SVM and CRF learning. Overall accuracy is 84% and we obtain 81%, 82% and 82% of precision and 77%, 78% and 86% of recall for three most important fields. Of course, the annotation ability decreases compared to the corpus level 1 because of the segmentation problem (see Table 2) and the irregularity of reference form. However our approach significantly outperforms a baseline CRF model which is learned with all notes without classification.

5.3. Discussion

While examining the applicability of CRFs into our bibliographical reference data in digital humanities field, we have discovered several interesting characteristics that would be useful in the treatment of other reference data in this domain and maybe in general cases. Recall that most of the existing works deal with the references of scientific research. Apart from their comparatively formulaic format, they have in many cases some specific words such as proceedings, conference, journal, etc. that make easier an accurate CRF prediction. However, in our DH references, these words have not been frequently found, and that is a reason why the accuracy of 'booktitle' field is not sufficiently high.

We also take notice of some phenomena, which are different from what was expected. First, as mentioned above, position features rather decrease accuracy when the target is note data. Second, the detailed features do not always helpful for annotation. For example, when we use a feature encoding the number of digits in a token, the accuracy decreases. Too detailed features might disturb well characterizing similar tokens having identical labels. Third, the model works better when the punctuation marks are identically labeled. We have tried various labeling strategies for punctuation marks such as taking the input token as output label, grouping them into several similar categories, or labeling only some important marks with input token. But these detailed treatments were always somewhat negative in terms of accuracy. Moreover, as we seek a simpler description of the features for a generalization, marking punctuation with an identical label seems reasonable.

For the scientific research purpose, the manually annotated three corpora will be distributed through a research blog[3], which records the progress of the project. The distribution modalities are now under discussion.

6. Conclusion

Three different levels of bibliographical reference corpora in digital humanities have been constructed. The target is the articles of Revues.org site, which is the oldest French online journal platform. The corpus construction involves a manual annotation of reference fields, that are then automatically estimated via machine learning techniques. According to the difficulty level of each corpus, we should employ an adapted methodology to well apply a CRF model. For the corpus level 1, we focus on finding the most effective set of tokenization basis, output levels and local features to establish a standard for the treatment of our DH reference data. We have obtained about 90% of overall accuracy. For the corpus level 2, we use another machine learning technique, SVM to select only the bibliographical notes, then we apply a CRF model to the selected ones. The accuracies have decreased compared to that of the previous corpus, but the model gives around 80% of accuracy for the three important fields. The construction of corpus level 3 is already finished, and it remains to develop a series of adapted methods to handle this corpus.

We are now testing several methods to improve the performance of CRF on corpus level 2. As the first step, we try to integrate proper noun lists into modeling to improve the author name and place fields. The most interesting part of the future work will be the treatment of corpus level 3. Topic models (Hofmann, 1999; Blei et al., 2003) will be appropriate tools to provide a semantic structure of the contents of the articles in corpus level 3 that can be useful for the extraction of implicit bibliographical part.

7. Acknowledgements

This work was supported by the French Agency for Scientific Research (Agence Nationale de la Recherche) under CAAS project (ANR 2010 CORD 001 02), and by Google with the Digital Humanities Research Awards that initiated the project.

8. References

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, March.

Isaac G. Councill, C. Lee Giles, and Min yen Kan. 2008. Parscit: An open-source crf reference string parsing package. In *LREC*. European Language Resources Association.

Min-Yuh Day, Tzong-Han Tsai, Cheng-Lung Sung, Cheng-Wei Lee, Shih-Hung Wu, Chorng-Shyong Ong, and Wen-Lian Hsu. 2005. A knowledge-based approach to citation extraction, information reuse and integration. In *Proceedings of IRI -2005 IEEE International Conference*, pages 50–55.

C. Lee Giles, Kurt D. Bollacker, and Steve Lawrence. 1998. Citeseer: an automatic citation indexing system. In *International Conference on Digital Libraries*, pages 89–98. ACM Press.

Thomas Hofmann. 1999. Probabilistic latent semantic indexing. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '99, pages 50–57. ACM.

Thorsten Joachims, 1999. *Making large-scale support vector machine learning practical*, pages 169–184. MIT Press, Cambridge, MA, USA.

Young-Min Kim, Patrice Bellot, Elodie Faath, and Marin Dacos. 2011. Automatic annotation of bibliographical reference in digital humanities books, articles and blogs. In *Proceedings of the CIKM 2011 BooksOnline11 Workshop*, pages 41–48.

Young-Min Kim, Patrice Bellot, Elodie Faath, and Marin Dacos. 2012. Automatic annotation of incomplete and scattered bibliographical references in digital humanities papers. In *Proceedings of the COnfrence en Recherche d'Information et Applications (CORIA 2012)*. To appear.

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, pages 282–289, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Andrew Kachites McCallum. 2002. Mallet: A machine learning for language toolkit. http://www.cs.umass.edu/ mccallum/mallet.

Fuchun Peng and Andrew McCallum. 2006. Information extraction from research papers using conditional random fields. *Inf. Process. Manage.*, 42:963–979, July.

Kristie Seymore, Andrew Mccallum, and Ronald Rosenfeld. 1999. Learning hidden markov model structure for information extraction. In *AAAI 99 Workshop on Machine Learning for Information Extraction*, pages 37–42.

Charles Sutton and Andrew McCallum. 2011. An introduction to conditional random fields. *Foundations and Trends in Machine Learning*. To appear.

[3] http://bilbo.hypotheses.org/