Bilbo-Val: Automatic Identification of Bibliographical Zone in Papers

Amal Htait, Sebastien Fournier and Patrice Bellot
Aix Marseille University, CNRS, ENSAM, University of Toulon, LSIS UMR 7296,13397, Marseille, France
Aix-Marseille University, CNRS, CLEO OpenEdition UMS 3287, 13451, Marseille, France
{amal.htait, sebatien.fournier, patrice.bellot}@lsis.org

Abstract

In this paper, we present the automatic annotation of bibliographical references' zone in papers and articles of XML/TEI format. Our work is applied through two phases: first, we use machine learning technology to classify bibliographical and non-bibliographical paragraphs in papers, by means of a model that was initially created to differentiate between the footnotes containing or not containing bibliographical references. The previous description is one of BILBO's features, which is an open source software for automatic annotation of bibliographic reference. Also, we suggest some methods to minimize the margin of error. Second, we propose an algorithm to find the largest list of bibliographical references in the article. The improvement applied on our model results an increase in the model's efficiency with an $Accuracy$ equal to 85.89. And by testing our work, we are able to achieve 72.23% as an average for the percentage of success in detecting bibliographical references' zone.

Keywords: Bibliography, Automatic annotation, OpenEdition, Bilbo, SVM, TEI, PDF.

1. Introduction

In this paper, we present the automatic identification of bibliographical references' zone in papers, as far as we know an innovation in its domain. Our work is based on a research and development program presented at LREC (Kim et al., 2012a), it aimed to construct a software environment (BILBO[1]) enabling the recognition and the automatic structuring of references in scholarly digital documentation (papers, books, etc), independently from their bibliographic styles.

The final object is to provide automatic links between each reference and its article or book in OpenEdition site[2], which is composed of three different sub-platforms, Revues.org, Hypotheses.org and Calenda. Therefore the automatic recognition of references' zone is essential as a first step. Although our system works with semi-structured documents, since we only need to distinguish the paragraphs in the paper, but we served of the available corpora based on papers provided as structured files XML/Text Encoding Initiative[3] (TEI) by the OpenEdition's Revues.org platform[2].

As a first approach, we used an automate graph that can detect patterns of consecutive references and annotate them as the article's bibliography, and it is realised by the tool Unitex 3.0[4]. On the testing level, we are not capable of detecting long patterns such as bibliographical references' zones using Unitex 3.0. Therefore, we suggest the use of machine learning technique for the annotation of references, so we can treat each reference apart and not a large amount of data at once.

We present our contribution in two sub-tasks:

- First Sub-Task: Retrieving references using Support Vector Machines (SVM), due to a model initially created to differentiate between the footnotes containing or not containing bibliographical references.
- Second Sub-Task: Detecting references' zone of the document, if it exists, as the largest list of consecutive references detected on the first sub-task.

2. BILBO and Support Vector Machine

BILBO[1] is an open source software for automatic annotation of bibliographic reference. It labels the words according to their type (author, title, date, etc) as the example in Figure 1. Written in Python programming language, it is principally based on Conditional Random Fields (CRFs), machine learning technique to segment and label sequence data. As external software, Wapiti[5] is used for CRF learning and inference and SVMlight[6] is used for sequence classification.

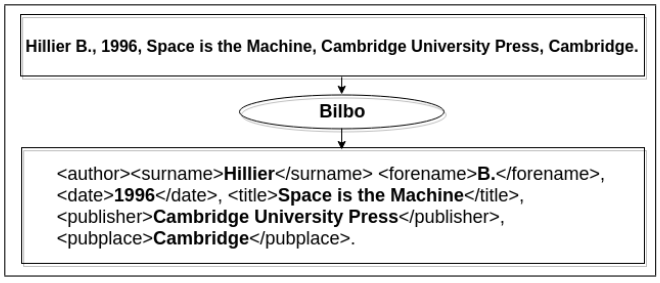

Figure 1: Example of reference annotation using BILBO.

BILBO's automatic annotation includes the bibliographical references in bibliographical zones, in footnotes and in text. To annotate bibliographical references in footnotes, we should first identify bibliographical parts, because the footnotes include both bibliographical and non-bibliographical

[1] bilbo.hypotheses.org
[2] www.revues.org
[3] www.tei-c.org/
[4] http://www-igm.uniov-mlv.fr/ unitex/
[5] https://wapiti.limsi.fr/
[6] http://svmlight.joachims.org/

information. We choose SVM for the classification between bibliographical and non-bibliographical information.

To build BILBO's SVM annotated corpora, we served of Revues.org articles references, in Figure 2 an example of these references (Kim et al., 2012b). That corpora contained 1147 annotated bibliographical footnotes references and 385 non-bibliographical footnotes that do not contain any reference.

1 Mathieu KALYNTSCHUK, *Le développement agricole et ses acteurs. L'exemple du département du Doubs (19e–milieu 20e 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.

Figure 2: Example of Footnotes from Revues.org papers as references and texts.

For testing purposes (Kim et al., 2012b), 1532 footnote instances were randomly divided into learning and test sets (70% and 30% respectively). It was tested for more than 20 different feature selection strategies. The best results, in Table 1, were achieved with the combination of the features, input words, punctuation marks and four different local features (posspage indicating page expressions such as 'p.', weblink, posseditor indicating editor expressions such as 'Ed.', and italic).

Accuracy	Prec_+	Rec_+	Prec_-	Rec_-
94.78%	95.77%	97.42%	91.43%	86.49%

Table 1: Previous results for identifying references in Footnotes (Kim et al., 2012b).

We should note that positive precision (Prec_+) and positive recall (Rec_+) measure the performance of the system to annotate correctly footnotes which contain references. And that negative precision (Prec_-) and negative recall (Rec_-) measure the performance of the system to annotate correctly footnotes which do not contain any references.
BILBO SVM model was basically oriented to work with footnotes, applying the knowledge gained on texts anywhere in the body of the article will be considered as Transfer Learning (Pan and Yang, 2009) technique. Although the high performance of BILBO in the bibliographical footnote field annotation, the transfer learning technique might decrease its performance. Therefore, while applying our sub-tasks, we modify the models results to increase its performance concerning the current task of identifying bibliographical zone.

As previously mentioned, we divide the work into 2 sub-tasks. For the first sub-task, we propose a strategy of 3 steps, as in Figure 3:

- The first step, we apply a possible filtering on paragraphs. We consider the length of a reference between 20 and 500 characters, based on an observation of 100 bibliographical references,
- The second step, we use BILBO SVM model to identify references,
- The third step, since our target is to detect bibliographical references' zone which is a list of consecutive references, we consider a non-bibliographical paragraph preceded and followed by references is most probably a reference. And the opposite is also true.

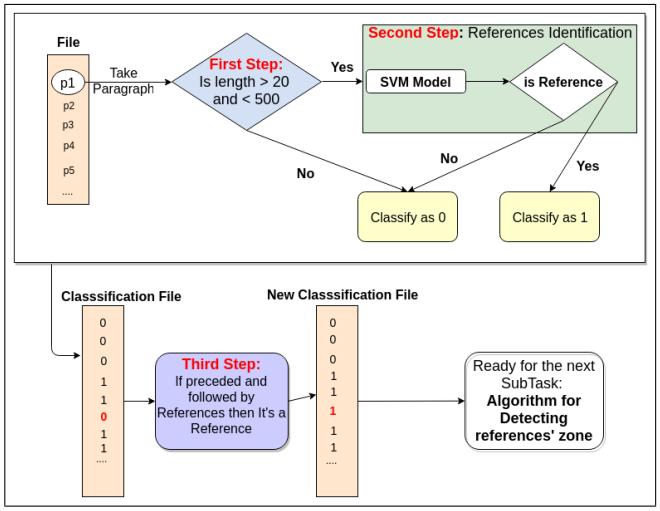

Figure 3: Subtask 1: The steps to find references in text.

For the second sub-task, we search for the largest list of consecutive references. Figure 4 explains the algorithm used to detect the bibliographical references' zone. The file is treated by paragraphs. Each paragraph is classified as reference or not reference by BILBO's SVM's model. Then the list of classified paragraphs is analysed: the first reference found is marked as the start of the zone, and with every new reference found we increment the size of the zone and mark it as the end of the zone. But once a non-bibliographical reference is found, in case of first appearance we ignore it and consider it an error by the SVM's model, but in case of second appearance, we reset our zone's variables (start, end and size) to zero, in the purpose of triggering a new search for another larger references' zone. And at the end of the list, we return the positions of the largest bibliographical references' zone found.

3. Evaluation

3.1. Testing of reference identification

For testing purposes, we built an annotated artificial document of 1411 paragraphs, of which 275 are bibliographical references and 1136 are not bibliographical references, extracted from 10 papers of the OpenEdition's Revues.org platform (5 French papers, 3 English papers, 1 Italian paper and 1 Spanish paper). An extract of the file is in Figure 5.
The prediction of SVM's model, as shown in the first line of Table 2, results an $Accuracy$ equals to 80.51, $PrecisionPositive$ equals to 59.64, recall positive equals to 50, $PrecisionNegative$ equals to 85.56,

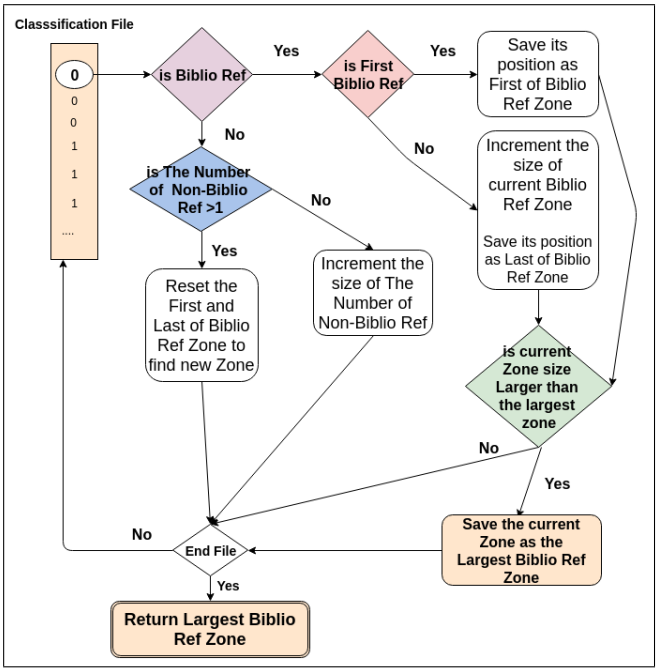

Figure 4: Subtask 2: Algorithm to detect the bibliographical references' zone.

```
<p>The challenge is therefore to conserve a unique patrimony through
     history that has been present in this territory throughout the c
<p>Figure 14: Cugnano: vestiges d'une structure métallurgique. </p>
<p> <hi rendition="#italic">Figure 14 : Cugnano : remains of a metall
<p> <hi rendition="#bold">Bibliography </hi> </p>
<p><bibl> <hi rendition="#bold">Arangure, B., Bagnol, P., Dalla, L.,
          Archeologia Medievale</hi>, XXXIV, 79-113. </bibl></p>
<p><bibl> <hi rendition="#bold">Belli, M., De Luca, D. and Grassi, F.
        P. (dir.), <hi rendition="#italic">III Congresso Nazionale d
```

Figure 5: Example of the testing set for reference identification.

$RecallNegative$ equals to 89.75, $f\_measure$ positive[7] (Sasaki, 2007) equals to 54.4 and $f\_measure$ negative equals to 87.6.

By adding step 1 from Figure 5, the results, as shown in the second line of table 2, reflect an improvement of 2.76 points in the $Accuracy$ and 2.7 points in the $f\_measure$ positive. The most important improvement shown in our results is in the value of $recall\_positive$, and that can be be explained by the following: our method excludes the ambiguous non-bibliographical paragraphs from being mistaken for a bibliographical and by that we are increasing the number of the true positives (TP) in the Equation 2 of $recall\_positive$, where TP are examples correctly labeled as positives and false negatives (FN) refer to positive examples incorrectly labeled as negative (Davis and Goadrich, 2006).

$$recall\_positive = \frac{TP}{TP + FN} \quad (1)$$

An example of similar mistakes is "$< p > Figure13 :< /p >$". First, during the conversion from PDF to XML and since the concept of paragraph is based on space between lines, the label of the image (here the example of "Figure 13") can be considered as a paragraph. Then, since this label contains: a word that starts with a capital letter, a number and a punctuation, this label may be detected as a part of a reference. This can be explained by the fact that scholarly papers used for learning include a lot of bibliographic references that are very short and incomplete.

And by adding step 3 from Figure 5, we can detect, as in the third line of table 2, an improvement on all the levels of measurement, since we seek for the consecutive bibliographical references, and that method serves greatly our purpose.

Using step 1 and step 3, as in fourth line of table 2, leads to an improvement of accuracy and $f\_measure$ positive and negative by almost 1 point, but a decrease in precision positive by 7 points. Although this decrease, we decided to use both methods due to their positive effect on accuracy and $f\_measure$.

3.2. Testing of reference's zone identification

For testing both sub-tasks, the detection of references and the detection of references' zone, we used 20 papers in XML/TEI format from the journals of OpenEdition.org. An extract of the expected result file is in Figure 6, with an annotation of the references by the tag $< bibl >$, and of the references' zone by the tags $< firstBibl >$ to show the beginning of the zone, and $< lastBibl >$ to show the end of the zone.

```
<p><hi rendition="#italic">Figure 14 : Cugnano : remains of a me
<p><hi rendition="#bold">Bibliography </hi></p>
<firstBibl><hi rendition="#bold">Belli, M., De Luca, D. and Gras
annocchieschi. <hi rendition="#italic">In</hi> Fiorillo, R. and
rcheologia Medievale</hi>, All'Insegna del Giglio, Firenze, 286
<bibl><hi rendition="#bold">Cascone, G. and Casini, A., 1997.</h
ampiglia M.ma. <hi rendition="#italic">In </hi>Zanini, A. (dir.
oscana centro-occidentale</hi>. Pacini, Pisa, 21-23. </bibl>
<bibl><hi rendition="#bold">Casin, A. and Zuccon, M. (dir.), 200
odo imprenditoriale e innovativo il patrimonio culturale e ambi
<bibl><hi rendition="#bold">Guideri, S., 1996.</hi> <hi renditio
erritorio a vocazione mineraria: le Colline Metallifere nella T
isa-Siena, Italie. </bibl>
<lastBibl><hi rendition="#bold">Insolera, I., 1990.</hi> <hi ren
iovanni Val d'Arno. </lastBibl>
```

Figure 6: Extract of a result file after bibliographical zone detection.

The below numbers show the results of our test, grouped by the level of correct bibliographical zone detection:

- 2 articles with a correct detection of the bibliographical zone, where the beginning and the end of the bibliography in the articles were marked correctly.
- 17 articles with a partially correct detection, where we have a detection of a major part of the bibliography, but not the complete zone is detected. An example is in Figure 7, the annotation skipped the first reference since our SVM's model considered it not a bibliographical reference paragraph.
- 1 article with wrong detection of bibliographical zone. An isolated reference in the middle of the article was

[7]The $f\_measure$ used is the harmonic mean of precision and recall.

	Accuracy	Precision_+	Recall_+	f_mesure_+	Precision_-	Recall_-	f_mesure_-
Initial (Step 2 alone)	80.51%	59.64%	50%	54.4%	85.56%	89.75%	87.6%
Applying Step1 (with Step 2)	83.27%	57.1%	57.1%	57.1%	89.61%	89.61%	89.61%
Applying Step3 (with Step 2)	84.47%	63.27%	59.59%	61.37%	89.6%	90.97%	90.28%
Applying Step1 and 3 (with Step 2)	85.89%	60.73%	64.73%	62.67%	91.98%	90.62%	91.29%

Table 2: Results of references' detection steps.

annotated as bibliographical zone, as shown in Figure 8. That's a result of not detecting any other reference in the bibliography of the article by the SVM's model.

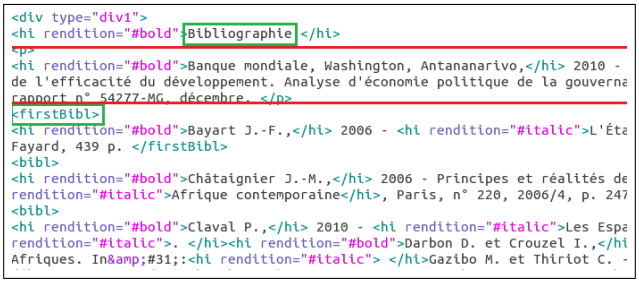

Figure 7: Extract of a partially correct zone detection.

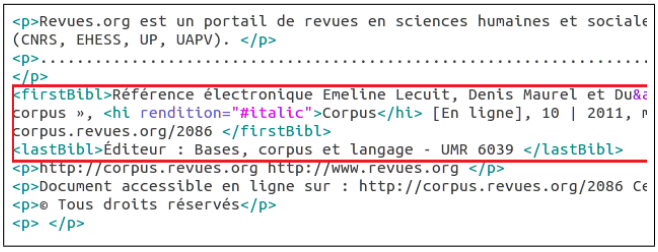

Figure 8: Extract of a wrong zone detection.

In Table 3, based on the previous results, we are able to calculate the percentage of success in the detection of references' zone, Equation 2. For example, in the second line of the Table 3, paper_2 have a bibliographical zone formed of 8 references, 7 are detected as references' zone and 1 is not considered in the zone. That would result a percentage of success equals to 87.5%. As an average for the set of 20 papers tested, we achieved 72.23%.

$$Percentage\_of\_Success = \frac{Nb\_of\_Detected\_References}{Nb\_of\_Total\_References} \quad (2)$$

We notice that with 15 out of 20 papers we achieve a percentage of success higher than 70%, and for the rest of the papers the SVM had some limitation with the detection of references.

4. Conclusion

To automatically annotate bibliographical references' zones, we first serve of a BILBO SVM model, created to differentiate between bibliographical references and non-bibliographical references in footnotes, to identify bibliographical references in the text of the papers body. To improve the system performance, we take into consideration that the bibliographical references in papers have an average number of characters that we can limit into an interval of maximum and minimum. Additionally, we consider that bibliographical zones contain consecutive references, and therefore any non-bibliographical reference detected while surrounded by bibliographical reference is considered a bibliographical reference. We achieve a $f\_measure$ equals to 62.67%. Then, as a second step, we search for the largest list of bibliographical references, and with a test on 20 papers, we achieve an average for the percentage of success equals to 72.23%.

As a future goal, we aim to detect bibliographical reference zones in PDF files and not only in structured files (XML/TEI) or semi-structured files. Since our work will be introduced as a new feature for the open source software BILBO, using directly PDF files as an input would be practical by saving time and work on converting files, not to mention the coast of tools that convert files from PDF to XML/TEI. We can also use machine learning technique like Conditional Random Fields (CRFs) for labeling references' zones after the detection of references by the SVM's model. Due to CRF, we can reduce the SVM's model errors.

This work is available as open source with BILBO on github.com[8].

5. Bibliographical References

Benkoussas, C., Hamdan, H., Bellot, P., Béchet, F., and Faath, E. (2014). A Collection of Scholarly Book Reviews from the Platforms of electronic sources in Humanities and Social Sciences OpenEdition.org. *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 4172–4177.

Davis, J. and Goadrich, M. (2006). The Relationship Between Precision-Recall and {ROC} Curves. *International Conference on Machine Learning (ICML)*.

Kim, Y.-M., Bellot, P., Faath, E., and Dacos, M. (2012a). Annotated Bibliographical Reference Corpora in Digital Humanities. *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012)*, pages 494–501.

Kim, Y.-M., Bellot, P., Tavernier, J., Faath, E., and Dacos, M. (2012b). Evaluation of BILBO reference parsing in

[8]https://github.com/OpenEdition/bilbo

	Nb_of_Total_References	Nb_of_Skipped_References	Nb_of_Detected_References	Percentage_of_Success
Paper_1	16	0	16	100%
Paper_2	8	1	7	87.50%
Paper_3	12	1	11	91.67%
Paper_4	56	1	55	98.21%
Paper_5	34	1	33	97.06%
Paper_6	58	1	57	98.28%
Paper_7	24	1	23	95.83%
Paper_8	19	1	18	94.74%
Paper_9	14	1	13	92.86%
Paper_10	17	2	15	88.24%
Paper_11	14	9	5	35.71%
Paper_12	41	9	32	78.05%
Paper_13	17	17	0	0.00%
Paper_14	25	18	7	28.00%
Paper_15	34	22	12	35.29%
Paper_16	74	22	52	70.27%
Paper_17	15	1	17	88.23%
Paper_18	28	20	8	28.57%
Paper_19	11	1	10	90.9%
Paper_20	62	34	28	45.16%
Average				72.23%

Table 3: Results for the percentage of success on a set of 20 Articles.

digital humanities via a comparison of different tools. *Proceedings of the 2012 ACM symposium on Document engineering - DocEng '12*, pages 209–212.

Ollagnier, A., Fournier, S., Bellot, P., and Béchet, F. (2014). Impact de la nature et de la taille des corpus d'apprentissage sur les performances dans la détection automatique des entités nommées. *Traitement Automatique des Langues Naturelles - TALN'2014*, pages 7–9.

Pan, S. J. and Yang, Q. (2009). A survey on transfer learningNo Title. *IEEE Transactions on knowledge and Data Engineering*, pages 1–15.

Sasaki, Y. (2007). The truth of the F-measure. pages 1–5.