Automatic annotation of incomplete and scattered bibliographical references in Digital Humanities papers

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

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

†LSIS, Aix-Marseille University / CNRS, 13397 Marseille
patrice.bellot@lsis.org

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

ABSTRACT. *In this paper, we deal with the problem of extracting and processing useful information from bibliographic references in Digital Humanities (DH) data. We present our ongoing project BILBO, supported by Google Grant for Digital Humanities that includes the constitution of proper reference corpora and construction of efficient annotation model using several appropriate machine learning techniques. Conditional Random Field is used as a basic approach to automatic annotation of reference fields and Support Vector Machine with a set of newly proposed features is applied for sequence classification. A number of experiments are conducted to find one of the best feature settings for CRF model on these corpora.*

RÉSUMÉ. *L'extraction d'informations bibliographiques depuis un texte non structuré demeure un probléme ouvert que nous abordons, via des approches d'apprentissage automatique, dans le domaine des Humanités Numériques. Nous présentons dans cet article le projet BILBO, soutenu par un Google Digital Humanities Award avec le soutien du projet ANR CAAS : constitution de 3 corpus de référence correspondant à trois localisations des références, élaboration d'un modéle d'annotation puis évaluation. Les champs aléatoires conditionnels (CRFs) sont utilisés pour l'annotation des références bibliographiques et des machines à vecteurs supports (SVMs) pour l'identification des références au sein du texte. De nombreuses expériences sont conduites afin de déterminer les meilleures propriétés devant être exploitées par les modèles numériques.*

KEYWORDS: *CRFs, Digital Humanities, Bibliographical Reference Annotation*

MOTS-CLÉS : *CRFs, Humanités Numériques, Annotation des Références Bibliographiques*

1. Introduction

This paper presents a series of machine learning techniques applied for information extraction on bibliographical references of DH documents. They are realized under a research project supported by Google digital humanities research awards in 2011. It is a R&D program for bibliographical references published on *OpenEdition*, an on-line platform of electronic resources in the humanities and social sciences. This aims to construct a software environment enabling the recognition and automatic structuring of references in digital documentation whatever their bibliographic styles.

The main interest of bibliographic reference research is to provide automatic links between related references in citations of scholarly articles. The automatic link creation essentially involves the automatic recognition of reference fields, which consist of author, title and date etc. A reference is considered as a sequence of these fields. Based on a set of correctly separated and annotated fields, different techniques can be applied for the creation of cross-links. Most of earlier studies on bibliographical reference recognition (or 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 *et al.*, 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 SVM classifiers (Joachims, 1999). Compared to HMM and SVM, CRF obtained better labeling performance.

Here we first choose CRFs as method to tackle the problem of automatic field annotation on DH reference data. It is a type of machine learning technique applied to the labeling of sequential data. The discriminative aspect of this model enables to overcome the restriction of previously developed HMM (Rabiner, 1989), then provides successful results on reference field extraction (Peng *et al.*, 2006). Most of publicly accessible on-line services are based on this technique (Councill *et al.*, 2008, Lopez, 2009). However, previous researches deal with relatively well structured data with simple format such as bibliography at the end of scientific articles. Besides, DH reference data generally includes a lot of less structured bibliographical parts and various different formats. Moreover, our target area is not only bibliography part but also footnotes of article. Footnote treatment involves a segmentation problem, which means extracting precisely bibliographical phrases. A primary issue concerns the selection of footnotes that contain some bibliographical information. To resolve this, another machine learning technique, SVM is applied for sequence classification of footnote strings. We propose a mixed use of three different types of features, input, local and

global features for SVM learning. Especially the last one is new type of feature describing global patterns of footnote string that has not been used for text data analysis so far to our knowledge.

In this paper, we present three major contributions. First part deals with bibliographical reference corpora from OpenEdition (Section 3) with manual annotation. We construct three types of bibliographical corpus : structured bibliography, less structured footnotes, and implicit references integrated in the body of text. Two formers are our main interest in this paper. Second part deals with defining effective labels and features for CRF learning with our new corpora (Section 4.1). This part enables us to establish standards concerning appropriate features for our data. The last one is sequence classification of footnotes (note hereafter), which constitute the second corpus (Section 4.2).

2. Bibliographical reference annotation

The OpenEdition platform is composed of three sub-platforms, Revues.org, Hypotheses.org and Calenda that correspond to academic on-line journals, scholarly blogs, and event & news calendar respectively. We work with Revues.org platform which has the richest bibliographical information. As a primary work, we want to automatically label reference fields in articles of the platform.

2.1. *Conditional Random Fields*

Automatic annotation can be realized by building a CRF model that is a discriminative probabilistic model developed for labeling sequential data. We apply a linear-chain CRF to our reference labeling problem as in the recent studies. 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 *et al.*, 2011). This aspect derives a key characteristic of CRFs, the ability to include a lot of input features in modeling. It is essential for some specific sequence labeling problems such as ours, where input data has rich characteristics. The conditional distribution of a linear-chain CRF for a set of label **y** given an input **x** is written as follows :

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

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 $\mathbf{x}$ 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, $D = \{\mathbf{x}^{(i)}, \mathbf{y}^{(i)}\}_{i=1}^{N}$.

2.2. *Sequence classification*

Apart from the main sequence annotation task, our work brings the sequence classification issue to segment exactly bibliographical parts because the less structured note data naturally includes both bibliographical and non-bibliographical information. By selecting just bibliographical notes, then learning a CRF model on these notes, we can expect an improvement in terms of annotation accuracy.

Sequence classification simply means that classification target is sequence data as ours. A recent work of (Xing *et al.,* 2010) provides a good summary of existing techniques in this domain by dividing them into three different categories: feature based classification, sequence distance based classification, and model based classification. All techniques commonly aim to reflect the specific sequence structure of input data into classification. Our approach is a kind of feature based classification because our main concept is mixing local and global features, which depict the local and global properties of notes. Global features describing the characteristics of a whole instance are often applied for the analysis of visual data such as face recognition and handwritten character recognition but rarely used in text classification. After extracting suitable local and global features, a SVM is applied for note classification but any other techniques could substitute for it.

3. Corpus constitution

Faced with the great variety of bibliographical styles in OpenEdition platform, we construct three different corpora according to their difficulty levels for manual annotation. The target source is Revues.org site, the oldest French online academic platform that offers more than 300 journals in the humanities and social sciences. Considering the reuse of corpora, we try to include maximal information in them instead of making annotation fit perfectly for the reference field recognition. From Revues.org articles, we prudently select and several references to build a corpus. To keep the diversity of bibliographic reference format over various journals on Revues.org, we select only one article for a specific journal. Since this paper concerns the first and second one, we focus on these two constructed according to the following difficulty levels whose examples are shown in Figure 1.

Figure 1. *Different styles of bibliographic references according to the difficulty level*

– Corpus Level 1: references are at the end of the article in a heading "Bibliography". Manual identification and annotation are relatively simple.

– Corpus Level 2: references are in footnotes and are less formulaic compared to level 1.

We started from the first level of corpus, which is relatively simple than the other, however needs the most careful annotation because it offers the standard for the construction of second one. Considering the diversity of bibliographical format, 32 journals are randomly selected and 38 sample articles are taken. Total 715 bibliographic references have been identified and annotated using TEI guidelines. A remarkable point in our manual annotation is that we separately recognize authors and even their surname and forename. In traditional approaches including recent ones, different authors in a reference are annotated as a single field (Peng *et al.*, 2006, Councill *et al.*, 2008). Another detailed annotation strategy is about treatment of punctuation. The human annotator reveals some punctuation marks, which play a role for the separation of reference fields.

In the second level of corpus, references are in the notes of articles. An important particularity of the corpus level 2 is that it contains links between references. That is, several references are shorten including just essential parts such author name, and sometimes are linked to previous references having more detailed information on the shorten ones. This case often occurs when a bibliographic document is referred more than once. The links are established through several specific terms as supra, infra, ibid, op. cit., etc. We select 41 journals from a stratified selection and we extract 42 articles after analysis of document. Note that the selected articles in the second corpus reflect the proportion of two different note types where one includes bibliographic information while the other does not. Since the final objective is a totally automated annotation system, it is necessary to let the system filter notes to obtain only bibliographical ones. For this purpose, the corpus 2 consists of two sub-sets: manually annotated reference notes as in the corpus 1 and the non-bibliographical notes. Consequently, we have 1147 annotated bibliographical references and 385 non-bibliographical notes in the second corpus.

4. Model construction

In this section, we detail our methodology towards an efficient CRF modeling, especially the definition of appropriate features and labels. Recall that it is difficult to find an earlier study on reference annotation of humanities data. Therefore we cannot be certain that the same fields as the previous works will be adapted to our corpora. The diversity in formats in our references also increases this incertitude. We therefore concentrate on the demonstration of the usefulness of CRF that involves finding the most effective set of features on the first corpus at first. The more original part of this paper is intended for the processing of corpus 2 that consists of applying new features for sequence classification. The sequence classification have been chosen against the major difficulty of corpus 2 that is the exact segmentation of bibliographical notes.

4.1. *Basic strategies on feature definition and label selection for CRF learning*

Since the included information in our corpus exceeds the necessary data of usual CRF learning for reference annotation, we first well define output labels before applying a CRF. And as the crucial part is to encode the useful formal characteristics into features, we focus on extracting the most effective features. The simplest way is using all tags in manual annotation. In this case, we have two major problems: a token often has multi-tags, not supported by a simple CRF, and original text often includes meaningless tags from labeling perspective, such as page break tags. So for the learning data, we give importance to cleaning up the useless tags and choosing the closest tag given a token. Then we empirically test different label selection strategies to obtain a set of optimized rules. Note that the labels are carefully selected in consideration of the main objective of the system that constructs a link structure between articles based on the extracted fields. So there is anyway the difference in the importance of fields. For example, surname, forename and title are the most important for the construction of useful links but the type of publication where the article is published is not a critical factor. We do not detail several editorial information such as issue numbers but they are integrated into the "biblscope" label. The determined labels after a number of experiments are described in Table 1.

Table 1. *Selected labels for learning data*

Labels	Description
surname	surname
forename	forename
title	title of the referred article
booktitle	book or journal etc. where the article is
date	date, mostly years
publisher	publisher, distributor
c	punctuation
place	place : city or country etc.
biblscope	information about pages, volume etc.
abbr	abbreviation

Labels	Description
orgname	organization name
nolabel	tokens having no label
bookindicator	the word "in", "dans" or "en" when a related reference is followed
extent	total number of page
edition	information about edition
name	editor name
pages	pages, in this version we don't use it
OTHERS	rare labels such as genname, ref, namelink, author, region

The feature manipulation is essential for an efficient CRF model. To avoid confusion, we call the features used in CRF that describe the formal characteristics of each token, "local features". We distinguish this kind of features from that in the general sense, and from global features introduced in the next section for sequence classification. As in the label selection, we explore the effects of local features via a lot of experiments. Too detailed features sometimes decrease the performance, so we need a prudent selection process. Currently defined features are presented in Table 2.

Table 2. *Defined local features for learning data and their descriptions*

Feature name	Description	Feature name	Description
ALLCAPS	All characters are capital letters	DASH	One or more dashes are included in numbers
FIRSTCAP	First character is capital letter		
ALLSAMLL	All characters are lower cased	INITIAL	Initialized expression
NONIMPCAP	Capital letters are mixed	WEBLINK	Regular expression for web pages
ALLNUMBERS	All characters are numbers	ITALIC	Italic characters
NUMBERS	One or more characters are numbers	POSSEDITOR	Abbreviation of editor

Recall that we manually annotate several punctuation marks, which play an important role for the separation of fields. We counted this aspect in case of developing a new technique, which can make use of them. Unfortunately the standard CRF is unable to reflect these specific marks in modeling because we do not have same information for a new reference to be estimated in practice. Therefore we decide to ignore them and this decision accompanies tokenization problem. In other words, we should also make a choice of tokenization particularly for the handling of punctuation.

Punctuation marks can be treated as either individual tokens or attached features to the previous or next word. In both cases, we need to make a supplementary decision on label and feature. For instance, when we decide to separate all the punctuation marks as tokens, we should also decide which label they will have. Our final criteria is to tokenize all the punctuation marks and label them with an identical label.

4.2. *Mix of input, local and global features for sequence classification*

Sequence classification is concerned with bibliographical note selection of corpus 2. We expect that pre-selected notes reduce usefulness parts in learning data then enhance the annotation performance. A text document for classification is basically represented by a set of word count-based features such as word frequency or tf-idf weight. We can also apply a dimension reduction technique such as feature selection to extract a more effective representation. However, the fact always remains that the original expression of text document is definitely based on word count. In general, this is reasonable because text classification aims to divide documents mostly according to their contextual similarity and dissimilarity, and the features having the most contextual information are words themselves. On the other hand, our note data needs not only contextual features but also formal patterns such as frequency of punctuation marks for classification. Because of this specificity, we try to newly generate features, which can reflect the sequential form of note texts.

Feature selection and generation

In our approach, we divide features into three different groups: *input feature*, *local feature* and *global feature*. Input features indicate words or punctuation marks in note string. Local features are the characteristics of input features just as that of CRF presented in Table 2. Global features, which are newly introduced in this paper describe the distributional patterns of local features in a note string. For example, a global feature "nonumbers" expresses the property of a note having no numbers in the string. This kind of information appears to be moderate for note classification because the discriminative basis often depends on the wide view of whole instance. Five global features, which are finally selected, are presented in Table 3.

Table 3. *Global features for note sequence classification*

Feature name	Description
NOPUNC	No punctuation marks in the note
ONEPUNC	Just one punctuation mark in the note
NONUMBERS	No numbers in the note
NOINITIAL	No initial expression in the note
STARTINITIAL	the note starts with an initial expression

While the typical features in traditional text classification are just input features, our approach mixes the above three types of features. We actively test many possible combinations of features, which can influence on the note classification. We empirically testify a number of feature combinations to select a set of moderate local and global features. In brief, global feature and binary local feature significantly improve the classification accuracy.

5. Experiments

In this section, we present the experimental result of our approach. The first is to find the most effective set of output labels and local features with corpus 1. The second is to verify the effectiveness of mixing three different feature types for sequence classification of note data in corpus 2, then to show that a CRF model with classified notes outperforms the method without classification. For the evaluation of automatic annotation, we used the micro-averaged precision, which computes the general accuracy of estimated result, and also the F-measure of each label. We also evaluate the sequence classification result with similar measures.

5.1. *Result of automatic annotation on corpus 1*

The very first two or three experiments aimed at verifying the suitability of CRFs to our task and signposting the various directions for the preparation of an appropriate learning dataset. With this purposes, we first started with a simple learning dataset where the input sequences are automatically extracted from the corpus with its internal tokenization manually done. After confirming that this trial gives a reasonable result in terms of labeling accuracy, we gradually add the extraction rules for labels

Table 4. *Overall accuracies of the CRF models with different learning data*

Stage	Tokenizing	Labels	Local features	Accuracy	Remarks
1	Manual annotation	The most nearest tag	No features	85.24%	Impossible for new one
15	[ditto]	Elimination of some rare or inappropriate tags	comma, point	88.54%	same as above
21	Tokenize all punctuation marks	Punctuation marks are labeled as <c>	No features	89.56%	No separation between title and booktitle.
28	[ditto] + Initial expression as a token	Separation of <title> and <booktitle>	6 features	86.32%	Separation of title/ booktitle. #Tokens decreases
35	[**ditto**]	[**ditto**] **+ Unified similar tag**	**11 features**	**88.23%**	[**ditto**]

and features. 70% of corpus 1 (so 500 reference) are used as learning data, and the remaining 30% (215 references) are used as test data.

The overall annotation accuracies of five selected models are represented in Table 4. The result on the first stage model confirms that a learned CRF with our first trial version without any preprocessing gives a reasonable estimation accuracy (85.34% in general accuracy). It is encouraging for continuing to use CRFs for our task, because we already get this positive result without any local features. The 15th stage, which eliminates some useless tags, gives the most effective result when not considering automatic tokenization but using manual tokenization. Moreover, two simple features 'comma' and 'point' are introduced to describe the nature of punctuation. With this learning data, we obtain 88.54% in general accuracy.

At the remaining stages, we applied various tokenization techniques. As a result, separating all the punctuation marks as tokens works well, especially when the marks are all labeled with an identical one. In the $21th$ stage, the overall accuracy increased again up to 89.56% with this punctuation treatment. But here the <title> and <booktitle> are not yet distinguished because we detail the title type in tag attributes. In the $28th$ and $35th$ stages, we extracted the nature of title from the corpus. Because of diversity of title types, it is not always easy to divide titles to <title> and <booktitle>. Considering the attributes and the place of title, we successfully separated two labels. Of course the accuracy decreases compared to the $21th$ stage because the number of labels increases. However, by introducing appropriate features, we finally get 88.23% of overall accuracy on the test dataset. We obtained comparatively high performance on important labels, surname, forename and title with about 90% of precision and recall.

5.2. *Result of sequence classification on corpus 2*

Classification result

We randomly divide 1532 note instances into learning and test sets (70% and 30% respectively). We test more than 20 different feature selection strategies by replacing feature types and detailed selection criteria. For each strategy, a SVM classifier is

Table 5. *Note classification performance with different strategies*

Id	Strategy	Accuracy	Positive		Negative	
			Precision	Recall	Precision	Recall
S1	input words (baseline)	87.61%	88.42%	96.28%	83.75%	60.36%
S2	input words + punc. marks (input)	89.35%	90.76%	95.70%	83.70%	69.37%
S3	input + 12 local features	87.39%	89.01%	95.13%	80.46%	63.06%
S4	[ditto] + weighted global features	90.0%	92.20%	94.84%	82.18%	74.77%
S5	input + non-weighted global features	90.65%	92.5%	95.42%	84.0%	75.68%
S6	[ditto] + binary local 'posspage', 'weblink', 'posseditor'	91.30%	93.28%	95.42%	84.47%	78.38%
S7	input + non-weighted global features + binary local 'posspage', 'weblink', 'posseditor', 'italic'	**94.78**%	**95.77**%	**97.42**%	**91.43**%	**86.49**%
S8	input + binary local 'posspage', 'weblink', 'posseditor', 'italic'	93.91%	95.29%	96.90%	88.88%	83.80%
S9	input + binary local 'posspage', 'weblink', 'posseditor'	90.0%	92.33%	94.92%	81.05%	73.33%

learned with the selected or newly generated features. A baseline is just using input word counts, one of the traditional approaches for text classification. Tf-idf weight is also tried. [1] Then by adding gradually various features, we find the most effective feature combination. Table 6 shows the performance of several notable strategies.

The baseline's accuracy on test data is 87.61%. Compared to the positive category, the performance of negative one is not good. It means that the used features are not sufficient to well describe the characteristics of negative notes, which do not contain bibliographical informaion. With S2, by applying punctuation marks as input data, total accuracy increases about 2 point (89.35%). Especially a remarkable gain (9 point) is achieved on the recall of negative notes (69.37%). This result confirms that the punctuation marks are useful for the note classification task. However, when the local features are applied (S3), the result is different from what we were expecting. Note here that we add another local feature called 'posspages' indicating page expressions such as 'p.' compared to Table 2. Here we count the frequencies of local features as input features. In S4, by applying global features, not only total accuracy (90.0%) but also the other accuracies increase and especially the recall of negative notes that achieves about 5 point more (74.77% vs. 69.37%). Meanwhile, we obtain an interesting result that the classification performance when we eliminate the local features (S5) is not really different from the previous experiment. It means that when we use global features, the local features with their frequency do not influence the performance, while they were rather negative when being applied alone. In this circumstance we expect that a different representation of local features may give a different result. With this supposition, we expect that a binary expression of several selected local features may bring a positive effect.

Now, instead of counting the appearance of local feature in a note string, a binary value is used to mark the existence of each feature. In S6, we show a combination of three features, 'posspage', 'weblink' and 'posseditor' that achieves a small improvement but better than the other combinations except the following strategy S7, which gives the best result. In S7, we applied the 'Italic' binary feature by keeping other as-

1. As it has same result with baseline, we do not show it.

Table 6. *Bibliographical note field annotation performance of CRF models learned on NotesCL (our proposition) and NotesOR (baseline, without classification).*

F-MEASURE

	surname	forename	title	booktitle	publisher	date	place	biblscope	abbr	nolabel	Accuracy
NoteCL	**81.97**	**83.32**	**87.11**	**55.31**	**77.67**	**91.74**	**88.24**	**87.28**	**94.40**	**56.74**	**87.28**
NoteOR	77.77	80.13	81.23	45.44	73.62	85.71	85.34	86.94	93.42	51.49	85.16

pects on the previous strategy. This brings a large improvement, which gives 94.78% of accuracy, and especially we obtain a great increase in both precision and recall of negative note (91.43% and 86.49% respectively). On positive notes also, we obtains a significant improvement (95.77% and 97.42% respectively). This result is reasonable because the italic feature usually appears in the title of article that is one of the main contributions of bibliographic reference. However when 'italic' feature is used with frequency value, it was not effective or rather worse on the accuracy. And we suppose that if the influence of 'italic' feature is that much, we may get rid of the global features. S8 and S9 implement this idea by using only input and binary local features. The performance of the S8, which applies four local features verified before, is slightly below that of S7. Moreover, the elimination of 'italic' feature rapidly degrades all accuracies especially on negative category (S9).

Bibliographical field annotation result

Now we verify the usefulness of note classification on automatic annotation of bibliographical reference. For that we construct two different CRF models on both classified set and original set without classification. We decide to reuse the notes which had been used for SVM learning, because a CRF model is independently learned with the previous SVM construction. So at first, the SVM classifier with S7 strategy is applied on all notes to find notes having bibliographical information. The classified note set with our strategy S7 is called 'NotesCL' which consists of 1185 notes where randomly selected 70% are used for CRF learning and the rest 30% for test. Non-classified note set, 'NotesOR' just takes all 1532 notes and divided into 70% and 30% for learning and test likewise. Table 6 shows the automatic annotation result in terms of F-measure. Bold value means that the corresponding CRF model better estimates on the field than the other model. Our approach always outperforms the annotation without classification for different fields. Total annotation accuracy of our method is 87.28% and is better than annotation without classification (85.16%).

6. Conclusion

We have presented a series of experiments that records the progress of our project. We started from the constitution of bibliographical reference corpora with manual annotation. Then we tried to find the most effective setting in terms of labels, features and tokenization. We verified that CRFs are appropriate for our dataset, and we have

obtained about 90% of precision and recall on surname, forename and title. Then we moved to the next step, processing of more complicated note data. A mixing strategy of input, local and global features significantly enhanced the sequence classification using SVM. And we have verified that the note annotation on the pre-classified learning set outperforms same method on the non-classified dataset. After a detailed analysis on both corpora, we found several important directions for the improvement of current system. Incorporation of external resources such as proper noun lists can be realized by post-processing or modification of model.

Acknowledgements

We thank Google for the Digital Humanities Research Awards and ANR CAAS (ANR 2010 CORD 001 02) that support this research project.

7. References

Councill I. G., Giles C. L., yen Kan M., « ParsCit: An open-source CRF reference string parsing package », *LREC*, European Language Resources Association, 2008.

Day M.-Y., Tsai T.-H., Sung C.-L., Lee C.-W., Wu S.-H., Ong C.-S., Hsu W.-L., « A knowledge-based approach to citation extraction », *Proceedings of IRI -2005*, p. 50-55, 2005.

Giles C. L., Bollacker K. D., Lawrence S., « Citeseer: an automatic citation indexing system », *International Conference on Digital Libraries*, ACM Press, p. 89-98, 1998.

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

Lafferty J. D., McCallum A., Pereira F. C. N., « Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data », *Proceedings of ICML '01*, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, p. 282-289, 2001.

Lopez P., « GROBID: combining automatic bibliographic data recognition and term extraction for scholarship publications », ECDL'09, Springer-Verlag, p. 473-474, 2009.

Peng F., McCallum A., « Information extraction from research papers using conditional random fields », *Inf. Process. Manage.*, vol. 42, p. 963-979, July, 2006.

Rabiner L. R., « A tutorial on hidden markov models and selected applications in speech recognition », *Proceedings of the IEEE*, p. 257-286, 1989.

Seymore K., Mccallum A., Rosenfeld R., « Learning Hidden Markov Model Structure for Information Extraction », *In AAAI 99 Workshop on Machine Learning for Information Extraction*, p. 37-42, 1999.

Sutton C., McCallum A., « An Introduction to Conditional Random Fields », *Foundations and Trends in Machine Learning*, 2011. To appear.

Xing Z., Pei J., Keogh E., « A brief survey on sequence classification », *SIGKDD Explorations Newsletter*, vol. 12, p. 40-48, November, 2010.