Linking Citations to their Bibliographic references

Huy Do Hoang Nhat
Web IR / NLP Group (WING)
National University of Singapore
huydo@comp.nus.edu.sg

Praveen Bysani
Web IR / NLP Group (WING)
National University of Singapore
bpraveen@comp.nus.edu.sg

Abstract

In this paper we describe our participation in the contributed task at ACL Special workshop 2012. We contribute to the goal of enriching the textual content of ACL Anthology by identifying the citation contexts in a paper and linking them to their corresponding references in the bibliography section. We use Parscit, to process the Bibliography of each paper. Pattern matching heuristics are then used to connect the citations with their references. Furthermore, we prepared a small evaluation dataset, to test the efficiency of our method. We achieved 95% precision and 80% recall on this dataset.

1 Introduction

ACL Anthology represents the enduring effort to digitally archive all the publications related to CL and NLP, over the years. Recent work by (Bird et al., 2008) to standardize the corpus in ACL Anthology, makes it more than just a digital repository of research results. The corpus has metadata information such as 'title', 'author (s)', 'publication venue' and 'year' about each paper along with their extracted text content. However it lacks vital information about a scientific article such as position of footnote (s), table (s) and figure captions, bibliographic references, italics/emphasized text portions, non-latin scripts, etc.

We would like to acknowledge funding support in part by the Global Asia Institute under grant no. GAI-CP/20091116 and from the National Research Foundations grant no. R-252-000-325-279.

The special workshop at ACL 2012, celebrates 50 years of ACL legacy by gathering contributions about the history, evolution and future of computational linguistics. Apart from the technical programme, the workshop also hosts a contributed task to enrich the current state of Anthology corpus. A rich-text format of the corpus will serve as a source of study for research applications like citation analysis, summarization, argumentative zoning among many others.

We contribute to this effort of enriching the Anthology, by providing a means to link citations in an article to their corresponding bibliographic references. Robert Dale [1] defines *citation*, as a text string in the document body that points to a *reference* at the end of the document. Several citations may co-refer to a single reference string. As an example consider the following sentence,

> Few approaches to parsing have tried to handle disfluent utterances (notable exceptions are *Core & Schubert, 1999; Hindle, 1983; Nakatani & Hirschberg, 1994).*

The portion of texts in *italics* are the citations and we intend to annotate each citation with an unique identifier of their bibliographic reference.

```
<ref target="BI10">Hindle, 1983</ref>
```

Such annotations are useful for navigating between research articles and creating citation networks among them. These networks can be used to understand the bibliometric analysis of a corpus.

[1] http://web.science.mq.edu.au/ rdale/

Proceedings of the ACL-2012 Special Workshop on Rediscovering 50 Years of Discoveries, pages 110–113,
Jeju, Republic of Korea, 10 July 2012.

2 Design

The task organizers distribute the entire Anthology in two different XML formats, 'paperXML' that is obtained from Optical Character Recognition (OCR) software and 'TEI P5 XML' that is generated by PDFExtract (ϕyvind Raddum Berg, 2011). We chose to process the PDFExtract format as it has no character recognition errors. Since the expected output should also follow 'TEI P5' guidelines, the latter input simplifies the process of target XML generation. The task of linking citations to references primarily consists of three modules.

1. Processing the 'Bibliography' section of a paper using Parscit.
2. Formatting the Parscit output to TEI P5 guidelines and merging it with the input XML.
3. Generating an identifier and citation marker for each reference and annotating the text.

Figure 1 illustrates the overall design of our work. Below we describe in detail about the modules used to accomplish this task.

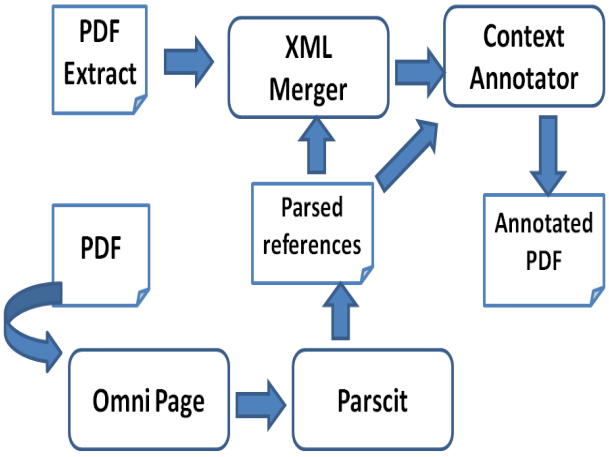

Figure 1: Overall design for linking citation text to references

Bibliography Parser: Parscit (Councill et al., 2008) is a freely available, open-source implementation of a reference string parsing package. It formulates this task as a sequence labelling problem that is common to a large set of NLP tasks including POS tagging and chunking. Parscit uses a conditional random field formalism to learn a supervised model and apply it on unseen data. During training, each reference is represented using different classes of features such as n-gram, token identity, punctuation and other numeric properties. Parscit can label each reference with 13 classes that correspond to common fields used in bibliographic reference management software. Unlike heuristic methods, Parscit's supervised learning model can handle different standards followed by different communities and inadvertent manual errors in the Bibliography. Prior to processing, Parscit segments the Bibliography section from the rest of the paper using SectLabel (Luong et al., 2010), its extension for logical document structure discovery.

Parscit works either with plain text or the Omnipage output of a paper. Omnipage [2] is a state of the art OCR engine that provides detailed information about the layout of a document. Omnipage also handles older, scanned papers. It gives the logical index of every line in terms of page, column, paragraph, and line number. The layout information is used by Parscit to remove noise such as page numbers and footnotes between references and properly divide them. Following is the Omnipage output for the word "Rahul Agarwal" in the original pdf,

```
<ln l="558" t="266" r="695" b="284"
  bold=true  superscript="none"
  fontSize="1250" fontFamily="roman">
 <wd l="558" t="266" r="609" b="284">
   Rahul </wd> <space/>
 <wd l="619" t="266" r="695" b="283">
   Agarwal </wd>
</ln>
```

The 'l' (left), 'r' (right), 't' (top), 'b' (bottom) attributes gives the exact location of an element in a page. Further, features such as 'bold', 'underlined', 'superscript/ subscript' and 'fontFamily' contribute towards an accurate identification and parsing of references. For example, the change from one font family to another usually serves as a separator between two different fields like 'author' and 'title' of the paper. As PDFExtract currently does not provide such information, we processed the original 'pdf' file using Omnipage and then finally parsed it using Parscit. Below is the XML output from Parscit for a single reference,

[2] www.nuance.com/omnipage/

```
<citation valid="true">
<authors>
 <author>R Agarwal</author>
 <author>L Boggess</author>
</authors>
<title>A simple but useful approach
to conjunct identification.</title>
<date>1992</date>
<marker>Agarwal, Boggess, 1992
</marker>  </citation>
```

We used Parscit to segment the Bibliography section into individual references. Additionally we use the author, title, publication year information together with the original marker of each reference to generate citation markers that are used to find the context of each reference (explained later). During this process, we generated the Omnipage output for the present Anthology that consists of 21,107 publications. As the ACL ARC has Omnipage outputs only till 2007, our contribution will help to update the corpus.

XML Merger: The original XML output from Parscit doesn't conform with the TEI P5 guidelines. The 'XML Merger' module formats the Parscit output into a 'listBibl' element and merges it with the PDFExtract. The 'listBibl' element contains a list of 'biblStruct' elements, in which bibliographic sub-elements of each reference appear in a specified order. Each reference is also assigned a 'unique id' within the paper to link them with their citation texts. The Bibliography section in the PDFExtract is replaced with the TEI compatible Parscit output such as below,

```
<listBibl>
 <bibl xml:id="BI2">
  <monogr>
  <author>R Agarwal</author>
  <author>L Boggess</author>
  <title>A simple but useful approach
   to conjunct identification.</title>
  <imprint>
   <date>1992</date>
  </imprint>
  </monogr>
 </bibl>
```

To ensure a proper insertion, we search for labels such as "References", "Bibliography", "References and Notes", or common variations of those strings. In the case of having more than one match, the context of first reference is used to resolve the ambiguity. The match is considered as the starting point of the Bibliography section, and the terminal reference string from the Parscit output is used to mark the end of it. After validating the matched portion based on the position of its starting and ending markers, it is replaced with the formatted 'listBibl' element.

Context Annotator: The final step is to bridge the links between references and citation strings in the merged XML. Several morphologically different markers are generated for each reference based on the 'author' and 'publication year' information provided by Parscit. These markers are used to find the corresponding citation string in the merged XML. The markers may vary depending upon the number of authors in a reference or the bibliography style of the paper. Sample markers for a reference with multiple authors are listed below,

```
Author1, Author2, Author3, Year
Author1 et.al, Year
Author1 and Author2, Year
```

Although Parscit provide the citation markers for each reference, the recall is very low. We extended these citation markers to make them more robust and thus improve the overall recall. Below are the extensions we made to the default markers.

1. Additional marker to allow square brackets and round brackets in the parentheses. Such markers help to identify citations such as (Author, Year), [Author, Year], (Author, [year])

2. Parscit markers only identify the citations with the 4-digit format of the year. We modified it to recognize both 4-digit and 2-digit format of the year. *e.g. Lin, 1996 and Lin, 96*

3. Parscit doesn't differentiate between identical reference strings with same author and year information. We resolved it by including the version number of the reference in the marker. *e.g. Lin, 2004a and Lin, 2004b*

4. Heuristics are added to accommodate the default citation texts as specified in the reference strings. For example in the reference string,

[Appelt85] Appelt, D. 1985 Planning English Referring Expressions. Artificial Intelligence 26: 1-33.

[Appelt85] is identified as the citation marker. Each marker is represented using a regular expression. These regular expressions are applied on the text from merged XML. The matches are annotated with the unique id of its corresponding reference such as '<ref target= BI10>'

3 Challenges

The accuracy of Parscit is a bottle-neck for the performance of this task. The false negatives produced by Parscit leads to erroneous linkage between citation texts and reference ids. In certain cases Parscit fails to identify portions of Bibliography section and skips them while processing. This results in an incorrect parsing and thus faulty linkage. Apart from Parscit, we faced problems due to the character mismatching between Omnipage and PDFExtract outputs of a paper. For example the string 'Pulman' is recognized as Pullan by Omnipage and as Pulman by PDFExtract. The citation markers generated from Parscit output in this case fails to identify the context in the PDFExtract.

4 Evaluation

As there is no dataset to test the efficiency of our method, we prepared a small dataset for evaluation purposes. We manually sampled 20 papers from the Anthology, making sure that all the publication venues are included. The citation strings in each paper are manually listed out along with the corresponding reference id. For citation styles where no Author and Year information is present, we used the contextual words to identify the citation text. The citation strings are listed in the same order as they appear in the paper. Below we provide an extract of the dataset, consisting of papers with three different citation styles,

```
P92-1006   proposed [13]      BI13
T87-1018   Mann&Thompson83    BI6
W00-0100   Krymolowski 1998   BI9
```

The first column is the Anthology id of the paper, second column is the citation string from the paper and third column is the unique id of the reference. We measure the performance in terms of precision and recall of the recognized citations. There are a total of 330 citation strings in the dataset. Our method identified 280 strings as citations, out of which 266 are correct. Hence the precision is 0.95 (266/280) and the recall is 0.801 (266/330). The low recall is due to the incorrect recognition of author and year strings by Parscit which lead to erroneous marker generation. The precision is affected due to the flaws in Parscit while differentiating citations with naked numbers.

In future we plan to devise more flexible markers which can handle spelling mistakes, using edit distance metric. Partial matches and sub-sequence matches need to be incorporated to support long distance citations. Parscit can further be improved to accurately parse and identify the reference strings.

Acknowledgements

We would like to thank Dr. Min-Yen Kan at National University of Singapore for his valuable support and guidance during this work.

References

Steven Bird, Robert Dale, Bonnie J. Dorr, Bryan Gibson, Mark T. Joseph, Min yen Kan, Dongwon Lee, Brett Powley, Dragomir R. Radev, and Yee Fan Tan. 2008. The acl anthology reference corpus: A reference dataset for bibliographic research in computational linguistics.

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

Minh-Thang Luong, Thuy-Dung Nguyen, and Min-Yen Kan. 2010. Logical structure recovery in scholarly articles with rich document features. *International Journal of Digital Library Systems*, 1(4):1–23.

ϕyvind Raddum Berg. 2011. High precision text extraction from PDF documents.