Corpus and Method for Identifying Citations in Non-Academic Text

Yifan He   **Adam Meyers**

Computer Science Department
New York University
{yhe, meyers}@cs.nyu.edu

Abstract

We attempt to identify citations in non-academic text such as patents. Unlike academic articles which often provide bibliographies and follow consistent citation styles, non-academic text cites scientific research in a more ad-hoc manner. We manually annotate citations in 50 patents, train a CRF classifier to find new citations, and apply a reranker to incorporate non-local information. Our best system achieves 0.83 F-score on 5-fold cross validation.

Keywords: Citation identification, Corpus annotation, Sequence labeling

1. Introduction

Identification of citations in text is an important first step in citation analysis. Existing citation parsing systems (e.g. ParsCit (Councill et al., 2008)) mainly focus on academic papers, where inline citations can typically be recognized in a two-step procedure: 1) collecting citations from the bibliography section; and 2) mapping collected citations to their mentions in text.

Beyond academic papers, other genres of text such as patents, weblogs, and newswire articles also cite scientific research. Extracting and analyzing citations in non-academic text can improve citation analysis by identifying technical trends earlier (through citations in weblogs) or by assessing the industrial impact of a research area more comprehensively (through citations in patents).

Recognizing citations in non-academic text is more difficult than citation extraction in academic articles for the lack of several important resources:

- **Bibliography**. Non-academic text such as patents and weblogs do not always have a bibliography section, so we do not have a repository to track inline citation mentions;

- **Consistent style**. In non-academic text, citation segments can appear in any order or be missing: e.g. many citations in patents do not have the name of the journal, venue, or even the title of the article explicitly mentioned (cf. Section 2.);

- **Annotated data**. While citation segmentation and citation mapping (from bibliography to reference) in academic text is well-researched and annotated data is readily available (see e.g. (Anzaroot and McCallum, 2013)), annotated data that could support citation identification in non-academic text does not exist to the best of our knowledge.

We try to address these limitations by creating a new annotated dataset that is tailored specifically for this task and train CRF-based taggers to extract citation instances. The rest of the paper is organized as follows: we first propose guidelines and accordingly create an annotated corpus of 50 patents on speech processing in Section 2., train sequence labeling models to extract citations in Section 3., refine system output with a simple reranker in Section 4., and report experimental results in Section 5. We review related work in Section 6. and conclude in Section 7.

2. Annotating Citations in Patents

In this paper, we focus on identifying citations in patents. Patents are of particular interest to us because they can reflect industry's acceptance to a technology and they are sorted with a classification number. We can easily focus on a small domain in this pilot study with the classification number.

2.1. Annotation Guidelines

We randomly collect 50 US patents from the speech processing domain for annotation, taken from a corpus of LexisNexis patents made available by the U.S. government.[1] We apply the following guidelines while annotating the corpus:

- **Definition of a citation**. We define a citation as a reference to another scholarly work. We limit ourselves to only annotating citations to scientific articles, as patent citations can usually be recognized with regular expressions and can have very different textual features.

- **Extent of a citation**. A citation is the longest, non-redundant, consecutive sequence of constituents (words, phrases, numbers or punctuation), such that each constituent either: (a) fills a field from the set: author, title of article, title of journal/collection, page range, year, volume number and issue number; or (b) is a one to three word-long string of words splitting the fields into two groups (there can only be one such string, so a citation can only be divided one time in this way).

---

[1]We are investigating ways of releasing these data.

2.2. Examples

We illustrate annotation guidelines with the following examples (citations are underlined):

EX1: *Such techniques are described, for example, by Robert Roth et al., Dragon Systems 1994 Large Vocabulary Continuous Speech Recognizer, Proceedings of the Spoken Language Systems Technology Workshop, Jan. 22-25, 1995, pages 116-120*

EX2: *e.g., the aforementioned U.S. Patent No. 5,414,796, Rabiner & Schafer, supra, and Rabiner & Juang, supra, at 69-140.*

EX3: *... as described in the article Cochlear Modeling by J. B. Allen appearing in the IEEE ASSP Magazine, January 1985, page 3.*

EX1 is comparable to a bibliographic entry in a scientific paper, but other citations in patents, such as **EX2** and **EX3**, can sometimes be informal. There are two short citations in **EX2** that only include author names and page numbers. We annotate both cases, because although they lack some critical information, it is still possible to map them to the complete form of a citation that appears elsewhere in the patent. Besides author names and page numbers, we also allow article citations that are limited to other subsets of the appropriate fields, e.g., title, year, and publisher. However, the reference to another patent is not part of the annotation; see the unannotated "U.S. Patent No. 5,414,796" in **EX2**.

We also consider **EX3** as a citation, although it does not conform to typical citation styles in academic writing. We include the words "appearing" and "in", as these link the title of the article with other fields of the citation.

Following the above guideline, we annotate 390 citations in 50 patents, averaging 7.8 citations per patent.

3. Citation Identification

In this section, we treat citation identification as a sequence labeling problem. We describe our baseline that uses a linear chain CRF model and report the features used in our models.

3.1. Citation Identification as Sequence Labeling

Following common approaches in natural language chunking, we model citation identification as a sequence labeling problem. We use a linear chain conditional random fields (CRF: (Lafferty et al., 2001)) model, which predicts a sequence of citation labels $y_{1...T}$, given a token sequence $\mathbf{x}_{1...T}$. We use CRF as it is shown to perform better than other sequence labeling algorithms (e.g. HMM) in closely related tasks such as citation segmentation (Peng and McCallum, 2006).

We define citation labels ($y_{1...T}$) under the BIO paradigm, in which the first token in a citation has the label "B", the other tokens in a citation have the label "I", and non-citation tokens have the label "O".

In most NLP tasks, the input token sequence $\mathbf{x}_{1...T}$ represents a sentence, but as citations often consist of abbreviated journal names, author names, and irregular use of punctuation marks, existing sentence splitters developed mainly for news text do no perform reliably. Therefore, we conduct training and testing on paragraphs instead of sentences. Paragraph boundaries are determined by XML metadata and visual clues (consecutive line breaks).

3.2. Sequence Labeling Features

We use both surface level text features and gazetteer features of journal, conference, and author names.

- **Text features** are inspired by word and word shape features defined in (Collins, 2002), originally designed for named entity extraction. These include: the word; the shape (as in (Collins, 2002)) of the word; the type (letter, digit, or other) of the first and the last character. We extract features on a five word window (i.e. two words before the current word, two word after the current word, and the current word itself).

- **Gazetteer features** are collected from the internet. Authors include researchers in the NLP and speech field who are cited more than 500 times according to Microsoft Academic Search[2]. Journals include ACM[3] and IEEE[4] published journals, plus Computational Linguistics, IBM Journal of Research and Development, and Bell System Technical Journal. Conference names are collected from Microsoft academic search[5] as well. If a phrase matches an item of type t in the gazetteer, the first token in the phrase will fire the feature "Gazetteer-B-t", and the other tokens in the phrase will fire "Gazetteer-I-t". For example, in "Computational Linguistics", "Computational" will fire the binary gazetteer feature "Gazetteer-B-journal" and "Linguistics" will fire "Gazetteer-I-journal".

4. Incorporating Non-local Information

Citations in non-academic articles have some non-local features that are hard to encode in a ±2 word context window used by sequential CRF taggers. For instance, in a typical citation, quotation marks and brackets are paired and most of the words start with a alpha-numerical character. To incorporate these features into a CRF model will be expensive, but they are helpful for disambiguating some of the cases that the CRF model could not effectively recognize. In this section, we first show examples where non-local information could help improve the recognition of citations, then we present a simple reranking scheme which utilizes such information.

4.1. Non-local Information in Citation Extraction

Consider the following outputs from the CRF model using features described in Section 3.2..

EX4: *include a list of phrases such as ...* <u>*"AM", "PM",*</u> <u>*"Buzzer", ... "Volume Up", "Volume Down"... etc..*</u>

EX5: *...* *filter-bank based recognizers (*<u>*Rabiner, L. R. and*</u> <u>*Juang, B. H., Fundamentals of Speech Recognition, Pren-*</u> <u>*tice Hall, 1993).*</u>

The predictions of the CRF model are underlined. The predicted "citation" in **EX4** is in fact a list of commands acceptable by a dialog system. The annotated sequence contains a lot of capitalized characters and the word "Volume", which confuses the tagger. However, this sequence consists of many non-alphanumerical tokens, which suggests that it is not a legitimate citation. In **EX5**, the tagger makes a boundary error, erroneously including the right round bracket.

These examples show that non-local information such as the ratio of alpha-numerical tokens and bracket pairing can potentially help improve citation extraction accuracy. To avoid the cost of incorporating such information into CRF directly, we use the n-best output of the CRF tagger to approximate its search space and try to find the best prediction by reranking the n-best list.

4.2. Incorporating Non-local Information via Reranking

We utilize a reranker that follows a very simple deterministic rule: the reranker should traverse the n-best list from the top, and return the first citation in which all round/square/curly brackets match and the ratio of alpha-numeric tokens is greater than 0.25. If there is no item in the n-best list that could satisfy this constraint, the top item will be returned.

We use very simple rules to show the effectiveness of non-local information and the necessity of exploiting the search space of the decoder. However, statistical reranking might work much better when we have abundant training data[6].

5. Experiments

We perform 5-fold cross-validation on 50 annotated patents. We report precision, recall, and F-scores on chunk level. CRF training and decoding is performed with the CRF++ package[7] using its default setting.

5.1. Feature Experiments

In Table 1, we experiment with different feature combinations. Many of the errors occur on the boundary of citations, such as stopping before the name of the publisher or allowing unpaired brackets. We try to alleviate some of them with reranking (cf. Section 5.2.).

---

[6]We also experimented with statistical ranking algorithms, such as the ranking SVM (Joachims, 2002), but the result was unstable, because currently we can only obtain a very small number of training examples for statistical rerankers.

[7]http://crfpp.sourceforge.net

	Precision	Recall	F-score
TEXT	0.7997	0.7805	0.7900
+ AUTHORS	0.8062	0.7832	0.7945
+ CONFERENCES	0.8127	0.7959	0.8042
+ JOURNALS	0.8035	0.7878	0.7955
+ ALL GAZETTEER	0.8010	0.7854	0.7931

Table 1: Experimental results using different feature sets

Looking at the impact of features, we notice that the conference gazetteer (TEXT + CONFERENCE) helps performance the most. We expect this, because there are several very influential conferences on speech processing (such as ICASSP). They appear very often in citations and are usually unambiguous. Journal names help as well, but not as much as conference names. We suspect that this is domain-dependent: if we analyze patents on biology, where journals are more prominent than conferences, journal names could be more informative.

The gazetteer of author names (TEXT + AUTHORS) is not as helpful as conference and journal gazetteers. There could be two reasons: 1) author names are ambiguous, in the sense that identifying a person name in gazetteer does not always mean the existence of a citation. It could simply be a name mention or the author of a patent (patent citations should not be extracted according to our annotation guideline); and 2) our gazetteer only covers the most cited authors. Although the best papers that could lead to patents are most likely published on top journals and conferences, the same does not apply to authors.

5.2. Reranking

	Precision	Recall	F-score
TEXT + ALL GAZETTEER	0.8010	0.7854	0.7931
RERANKED	0.8243	0.8082	0.8162
TEXT + CONFERENCES	0.8127	0.7959	0.8042
RERANKED	0.8363	0.8187	0.8274

Table 2: Experimental results using reranking

We experiment with the reranker and present results in Table 2. We use the probability of the top CRF prediction as a confidence measure and only rerank instances whose top-1 probability is less than 0.99.

We limit ourselves to the 50-best output. We examine one fold of our test set, where the CRF model makes 21 incorrect top-1 predictions. Among these 21 errors, 10 gold sequences can be recovered from the 10-best list, 13 from the 27-best, 14 from the 63-best, and the rest are not found in the 500-best list. We therefore determine that the 50-best list is large enough for our purpose.

We rerank the 50-best output of both the best performing tagger (TEXT + CONFERENCES) and the tagger with the richest feature set (TEXT + ALL GAZETTEER). Experiments show that for TEXT + CONFERENCES, reranking improves F-score from 0.80 to 0.83, while for TEXT + ALL GAZETTEER, F-score improves from 0.79 to 0.82. This confirms the necessity of utilizing non-local features in citation extraction. Our current reranking scheme is very simple due to the scarcity of data. We believe that statistical

rerankers, or a model that incorporate non-local features are promising on this task when more training data is available.

5.3. Size of Training Data

	Precision	Recall	F-score
25% training data	0.7300	0.6169	0.6687
50% training data	0.7347	0.7049	0.7196
75% training data	0.7894	0.7658	0.7775
100% training data	0.7997	0.7805	0.7900

Table 3: Experimental Results using different size of training set

To explore whether more annotation is necessary, we change the size of training data during cross-validation. As is demonstrated in Table 3, we consistently obtain around 5 points improvement on F-score when we expand the data set from 25% to 50% and from 50% to 75%. The final quarter of training data still improves F-score by 2 points. This result shows that the size of training data is not yet saturated. We expect to obtain more improvement simply by adding more training data. It is also worth noting that we currently focus on a small domain and building a general domain citation extractor will need more annotated data.

6. Related Work

Citation recognition and analysis (Garfield, 1972) has attracted much interest from the research community. However, most of the existing work on recognizing citations either focuses on citation segmentation in well-formed citations (Peng and McCallum, 2006), and/or performing coreference resolution among citation instances (Wellner et al., 2004).

The work presented in this paper is most closely related to the citation segmentation task. Early citation segmentation systems use either manual rules (Ding et al., 1999) or sequence labeling algorithms like HMM (Seymore et al., 1999). CRF is later applied to citation segmentation and achieves high accuracy (Peng and McCallum, 2006), but is shown to be sensitive to domain variations. (Anzaroot and McCallum, 2013) provides a new dataset that covers the computer science domain better and establishes baseline results on the dataset. Open source packages such as ParsCit (Councill et al., 2008) add to the popularity of this strand of research.

Although we are also trying to analyze citations in text, the problem we are handling is different from citation segmentation. In addition, citations in non-academic text are more informal and lack consistent style, which leads us to create new annotation and build our own system, instead of reusing published work.

7. Conclusions and Future Work

We developed resource to support system development on citation extraction in non-academic text. We presented CRF models to identify citations, which obtained 0.80 F-score on 5-fold cross validation. We further improved the performance of our system with reranking and achieved 0.83 F-score.

We plan to continue our annotation effort and experiment with other sequence labeling paradigms, such as the LD-CRF model (Morency et al., 2007) that can capture the internal structures of citations.

8. References

Sam Anzaroot and Andrew McCallum. 2013. A new dataset for fine-grained citation field extraction. In *ICML Workshop on Peer Reviewing and Publishing Models (PEER)*.

Michael Collins. 2002. Ranking algorithms for named-entity extraction: Boosting and the voted perceptron. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics (ACL 2002)*, pages 489–496.

Isaac G. Councill, C. Lee Giles, and MinYen Kan. 2008. Parscit: An open-source crf reference string parsing package. In *The 8th Language Resources and Evaluation Conference (LREC 2008)*.

Ying Ding, Gobinda Chowdhury, and Schubert Foo. 1999. Template mining for the extraction of citation from digital documents. In *Proceedings of the Asian Digital Library Conference*, pages 47–62.

Eugene Garfield. 1972. Citation analysis as a tool in journal evaluation. In *Essays of an Information Scienties*, volume 1, pages 527–544.

Thorsten Joachims. 2002. Optimizing search engines using clickthrough data. In *KDD '02: Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 133–142, New York, NY, USA.

John Lafferty, Andrew McCallum, and Fernando Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the 18th International Conference on Machine Learning (ICML 2001)*.

Louis-Philippe Morency, Ariadna Quattoni, and Trevor Darrell. 2007. Latent-dynamic discriminative models for continuous gesture recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2007)*, pages 1–8.

Fuchun Peng and Andrew McCallum. 2006. Information extraction from research papers using conditional random fields. *Information Processing & Management*, 42(4):963–979.

Kristie Seymore, Andrew McCallum, and Roni Rosenfeld. 1999. Learning hidden markov model structure for information extraction. In *AAAI-99 Workshop on Machine Learning for Information Extraction*.

Ben Wellner, Andrew McCallum, Fuchun Peng, and Michael Hay. 2004. An integrated, conditional model of information extraction and coreference with application to citation matching. In *Proceedings of the 20th conference on Uncertainty in artificial intelligence (UAI 2004)*, pages 593–601.