Mining Service Contributions from Web-based Journal Listings

Xinze Li
National University of Singapore
Singapore, Singapore
lixinze888@gmail.com

Abhinav Ramesh Kashyap
National University of Singapore
Singapore, Singapore
abhinav@comp.nus.edu.sg

Min-Yen Kan
National University of Singapore
Singapore, Singapore
kanmy@comp.nus.edu.sg

ABSTRACT

The diversity of person names and their roles presents challenges to mine them. We refine the modern neural named entity recognition (NER) approach to extract person names and their roles by leveraging the relationship between them. By using high-quality embeddings extracted from cleaner datasets, we improve BiLSTM–CRF extraction performance in lower-quality datasets. Our method also addresses name data sparsity problems through data augmentation and refinement to improve the recognition of underrepresented name ethnicities. We employ our method to extract service contributions — in the form of editorial board roles — from journal websites. We use our method to augment limited supervised data tuples of researcher's names and affiliations and their board roles. On a constructed, large dataset of approximately 300 journals over three major scientific publication houses (Springer, ACM, and IEEE), we demonstrate that these refinements significantly and consistently reduce the errors by over 30% made by the standard BiLSTM–CRF in identifying names and their affiliations.

CCS CONCEPTS

• **Information systems → Digital libraries and archives**; • **Computing methodologies → Information extraction**.

KEYWORDS

Web Crawling, Information Extraction, Name Entity Recognition

ACM Reference Format:
Xinze Li, Abhinav Ramesh Kashyap, and Min-Yen Kan. 2021. Mining Service Contributions from Web-based Journal Listings. In *Woodstock '18: ACM Symposium on Neural Gaze Detection, June 03–05, 2018, Woodstock, NY*. ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

Contributing one's own time to serve the academic community is an hallmark of the academic tradition. Such scholarly service contributions can be in the form of hosting academic conferences, carrying editorial work at journals, posting academic-related blogs, etc. Recently, Tian et al. [10] quantify the service contributions from the computer science conferences by extracting information from Call For Papers (CFP). Unfortunately, mining such contributions from scientific journals has not received sufficient attention in our view. Considering journals — an important medium of communication for scholars — makes the quantification of their service contributions more holistic, reducing problems with coverage. However, mining names of researchers and their roles from academic journals is fraught with many problems.

```
<p> <strong> Editorial Board </strong> </p> <p> Xiaochuan
Cai , University of Colorado Boulder , USA </p>
```

Figure 1: Example input webpage excerpt (Source: CCF Transactions on High Performance Computing).

In particular, researchers' names come in a wide variety of forms. As such, the current standard approach of employing pre-trained neural network models may not perform well on them. In this work, we aim to improve entity extraction accuracy on journal editorial boards. We use named entity recognition (NER) techniques to extract relationship between researcher names and their editorial role relationships. We focus on the following aspects: (i) to establish an end–to–end journal editorial board information extraction procedure; (ii) to improve the performance for the sequence labelling to achieve higher accuracy for entity extraction; and (iii) to investigate the factors that affect performance on a dataset for this task. We handle this task via a BiLSTM–CRF network, and use the $BERT_{base}$ network as a competitive, modern baseline comparison. We make the following contributions:

- We propose methods to improve the performance of NER extraction when the data are noisy and limited. We show that our approaches that include data cleaning and synthesis bring an error reduction of at least 30%, and an increase of F_1 of at least 2.5 for all three datasets.
- We enhance word embedding models to markedly improve the extraction performance on a low-quality dataset. With our method, a BiLSTM–CRF model can reach a comparable performance to a $BERT_{base}$ state-of-the-art model for all datasets.

2 RELATED WORK

To the best of our knowledge, there is currently has been no mature system that extracts researcher information from journal editorial boards. Schneider's work [8] is an innovative attempt of extracting researcher information from *Call for papers* web pages using a linear-chain Conditional Random Field (CRF). It integrates information from both the textual content and the layout of the page with CRF. However, their system does not match names with the affiliations. More importantly, he relies on a manually-annotated training corpus. When such annotation corpora are small, models can be insufficient trained and fragile, hindering robust progress on this task. Further, since annotating training data manually is

Woodstock '18, June 03–05, 2018, Woodstock, NY

ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/10.1145/1122445.1122456

labour-intensive, it brings to the fore the necessity of methods to effectively increase dataset size with little human intervention. This is one of our key foci.

There is work that measures service contribution through crawling, extracting, and quantifying researcher roles in in the technical program committees (TPC) of Computer Science conferences [3]: the Gatekeeper system. Gatekeeper crawls records of information of conferences and TPC, using standard NER techniques to identify citation distributional information from the citations to works published in particular conference venues. Subsequent work [2] improves Gatekeeper's extracting process through improving the quality and robustness of their data.

The work that is the most closely related to us is ServiceMarq [10]. It mines service contribution from CFPs by using a word embedding based BiLSTM line classification, followed by entity extraction and name disambiguation. However, journal editorial boards have limited data and varied page formats compared to CFPs. We make modifications to the entity extraction methods used by ServiceMarq to cater to specific features of editorial board listings. In addition, we introspect the word embedding and BiLSTM network model components with an aim to improve task performance.

3 IMPROVING EDITORIAL BOARD NAMED ENTITY RECOGNITION

Our approach consists of five components. **Initial Data Selection** (§3.1) first crawls web-based journal listings; second, **Base NER Tagging** (§3.2) then applies a pretrained NER library to the listings; **Seed Data Generation** (§3.3) then generates a large dataset suitable for model traning, **Data Refinement** (§3.4) refines the output data for editorial board tags and problems with the data distribution; and finally, we train a second, final **Sequence Labeling** model and word embeddings (§3.5) to apply to our task .

We select three representative publishers, largely representative of the domains of computer science and engineering: **Springer**, the Association for Computing Machinery (**ACM**), and the Institute of Eletrical and Electronics Engineers (**IEEE**). Springer is an industrial for-profit publisher, while the ACM and IEEE represent their academic and industrial disciplinary constituents.

3.1 Initial Data Selection

Publishers often include journal listing pages on their website, which disseminates key details of each of the journals from the publishing house. These can include listing the editorial board members, a key piece of metadata that can serve as a quality marker for readers, prospective authors and stakeholders involved quality assessment (e.g., faculty promotion committees). Our system initially takes such journal listing pages as input. From the selected publishers' journal listing pages, we extract journal names, homepage links and editorial board links. Next, our crawler extracts the editorial board information from the individual journal webpages and parses them into a sequence of lines.

3.2 Base NER Tagging

Traditional NER sequence labelling tasks use labelled data instances, derived from individual sentences.

For example, "Xiaochuan [PER] Cai [PER] , University [ORG] of [ORG] Colorado [ORG] Boulder [ORG] , USA [LOC]" is an annotated sample following the standard CONLL format, where annotations of interest are given in square brackets after each token. In the example, the contiguous "University of Colorado" is a tagged as a ORGANIZATION entity, despite also being an example of a LOCATION.

Our scenario is similar to general use cases where PERSON, ORGANIZATION, and LOCATION are useful tags to recognize as evidence for editorial board roles. For this reason, we first apply the popular pre-trained named entity (NE) recognizer from the FLAIR library to the extracted lines. As the FLAIR NLP library[1] comes with a pre-trained model trained on a large document set, it is a robust solution for most natural language data — inclusive of ours — that manifest such entity types. The output lines, which now include automatically-tagged, possibly noisy PERSON, ORGANIZATION and LOCATION entities, is passed to the data generation stage.

3.3 Data Generation

There is a caveat: our data is similar but not identical to that of FLAIR. Like most supervised systems, the FLAIR NER tags standard NE types well, but cannot tag unseen entity types such as editorial board roles. Standard pre-trained models alleviate the need for manual collection of training data, as long as the data to be tagged are distributionally similar and the tagset is identical to the original task of the model. We need to identify tokens that indicate editorial board ROLES as well (such as "editor-in-chief"). This violates the assumptions of the original model application and as such, such pre-trained model fails to work well without suitable algorithmic adaptation. To tag such entities well, we must adapt our recognition pipeline. To design a new NER, traditionally, human experts are recruited to annotate, an expensive and time-consuming process.

The key aim of our work is to address this bottleneck. Recognizing the benefit of using a robust NER tagger for general NER, rather starting from scratch to create a new tagger, we elect to layer an additional classifier which includes editorial board ROLE labels, on top of the FLAIR output. Note this approach does not require access to the original training data for FLAIR, and also allows us to fix distributional data shifts: FLAIR is trained on standard newswire data (i.e., Reuters Newswire data appearing in the CoNLL dataset [7]), whereas our deployment is over publisher webpages.

Rather than tag a large set of clean data, we proceed in a semi-automated, *generate-then-correct* manner. We collect a comprehensive gazetteer from more than 150 Computer Science journals. For instance, the words "associate" and "deputy" are respectively likely to indicate the beginning of the roles of "Associate Editor" and "Deputy Editor". We then use this to training a classifier which also classifies roles. We randomly assign all sentences into train data, test data, and validation data with a ratio of 8:1:1.

With all the journals from the selected publishers processed using this method, we generate a large-scale (potentially noisy) dataset that can be used for supervised training.

3.4 Data Refinement

The above step has two limitations. First, the data generated might be noisy due to error cascades from the pre-trained FLAIR tagger.

[1] https://github.com/flairNLP/flair

Second, the size of the dataset is constrained by the number of journals under each publisher. Therefore, we also contribute a two-step data refinement process as follows.

1. Correcting Name Distributional Shifts: Machine learned models (inclusive of sequence labelers such as NER) often fail when there is a distribution shift [6]. In our error analysis of the output from the general NER tagger, we see many errors on editorial board member names. We surmise this to mean that the PERSON entities that feature in the underlying FLAIR data (newswire) differ from that in journal service contributions (academic and industrial research). Indeed, many errors are traceable to low recall on ethnically Asian surnames (especially Chinese ones, which are usually romanized using *pinyin* resulting in short tokens that are less descriptive and easily misclassified), which feature as a larger proportion of scholar names compared against person mentions in newswire [4]. To address this aspect, we take a deterministic approach to correct the FLAIR tagged data. We correct mentions of 100 common Asian surnames following a manually compiled listing. As surnames are only part of names, we extend our PERSON labels to preceding and subsequent word tokens, until the extension process encounters punctuation or hypertext markup. This first correction step is context-free, changing all tokens that the 100 common surnames unconditionally to PERSON tags. This garners high recall but is inherently noisy: e.g., it mislabels the "Hong Kong University" as PERSON, since "Kong" is one such surname. To recover precision, we add one exception: if the extension process encounters words that signify an ORGANIZATION, such as "university" and "institution", the located name entity is relabeled as ORGANIZATION.

2. Data Augmentation: Current state-of-the-art neural network models for natural language tagging hinges on access to sufficient training data to set the large number of parameter values effectively. Our refined, supervised dataset is insufficient to meet this need. We must augment our training data with additional synthetic data. We take a simple-yet-effective approach: for each sentence (which may have 0, 1 or more labelled entity instances), we add a permuted copy where all PERSON and ORGANISATION entities are individually replaced by entities randomly picked from a respective entity lexicon. This results in an augmented corpus (Table 1) that doubles the amount of training data. We form our replacement PERSON lexicon by including the full names of 500 scientists representing a diverse set of ethnicities. We form our replacement ORGANISATION lexicon by including the names selected from the top 500 QS Ranking 2020 Universities, since such institutions frequently appear in editor affiliations.

3.5 Sequence Labelling

The data passes through both the two steps of rule-based data refinement and data augmentation before being deployed to train our final sequence labelling model. For our input of journal listings, ORGANIZATIONS and LOCATIONS map to the scholar's AFFILIATION; hence the final training data has four classes: PERSON, AFFILIATION, editorial board ROLE and institutional LOCATION, as shown in Fig. 1.

We train a standard bi-directional long short term memory (Bi-LSTM) network with a final CRF layer for the sequence labelling task. The first layer of such a BiLSTM–CRF network normally employs word embeddings, such as GloVe [5], as its input token representation. However, as our data exhibits a diversity of PERSONS and AFFILIATIONS, many such entities will not be previously seen during testing (zero-shot, out of vocabulary). To mitigate these difficulties, we employ Byte-Pair Embedding (BPE) [9] which represents the tokens at a sub-word level, such that sub-word letter patterns can indicate the entity class of interest.

As GloVe is a generic embedding, we also train our own language models with our three separate datasets that contains editorial board tokens, as a replacement for GloVe. We can use these language models trained on our smaller, domain-specific corpora for stacking with BPE through embedding concatenation. Note that each stacked embedding is only applied the other two datasets which are not used to train itself. Such embeddings were trained on the original datasets, but in the case of the IEEE dataset, as it is much larger, we downsample it to 8K sentences of the other datasets, to ensure a fairer comparison. This practice shown to improve the performance on many previous tasks [1]. For comparison, we assess baseline representations from large-scale pretrained language models. We concatenate the contextual word representations from BERT along with traditional word representations.

Table 1: Dataset statistics. The 'Ratio' column divides the number of input sentences by the number of editors.

Dataset	# Journals	# Editors	# Sentences	Ratio
Springer	152	5,914	7,562	0.78
ACM	60	1,406	8,287	0.17
IEEE	75	4,233	79,870	0.05
Refined Springer	152	12,614	15,124	0.83
Refined ACM	60	3,322	16,574	0.20

4 EXPERIMENTS

We report our experimental results on our three datasets in Table 2. We use the pre-trained BiLSTM-CRF model from FLAIR [1] for our experiments and use the macro F_1 metric to evaluate performance. We report average scores from five independent run. With respect to hyperparameters, we set the learning rate to 0.1, maximum number of epochs to 100, and the dropout rate of 0.05.

For our baselines, we utilize (B1) GloVe embeddings[2], and a (B2) BERT transformer-based model, *base*. We test the effect of our data refinement step by varying the input data in R1. Model variations (§3.5) are given in the middle rows (M1–M4) which show the performance gains when using BPE alone or stacked with dataset-specific language models. We combine the refinement and representation modifications in the hybrid rows (MR1–2).

Examining these main results we make several observations. First, while BiLSTM–CRF garner good performance, a simple transformer BERT model improves performance markedly, reducing error rates significantly. Second, the data refinement consistently improves performance, especially on the IEEE dataset, where the dataset is large but entities of interest (editors) are more sparse (cf Table 1, "Ratio" column). Separately, adding subword embeddings via BPE yields consistent improvements, of the same order as data

[2]https://www.kaggle.com/watts2/glove6b50dtxt

Table 2: F_1 of baselines (B1–B2), with refined data (R1) and representation variants (M1–4), and combinations.

Model	Representation	Dataset	Springer/ACM/IEEE
B1. BiLSTM–CRF	GLoVE	Original	92.6 / 82.6 / 79.2
B2. $BERT_{base}$	token + position	Original	93.2 / 89.4 / 84.3
R1. BiLSTM–CRF	GLoVE	Refined	93.7 / 83.8 / 83.1
M1. BiLSTM–CRF	BPE	Original	**93.7** / 86.4 / 81.0
M2. BiLSTM–CRF	Springer+BPE	Original	— / **89.5** / **84.8**
M3. BiLSTM–CRF	ACM+BPE	Original	93.4 / — / 82.2
M4. BiLSTM–CRF	IEEE+BPE	Original	93.1 / 86.9 / —
RM1. BiLSTM–CRF	BPE	Refined	**95.1** / 87.2 / 86.2
RM2. BiLSTM–CRF	Springer+BPE	Refined	- / **90.1** / **87.5**

refinement. We also see marked improvements in the ACM and IEEE dataset when utilizing the language model derived from the Springer dataset. We now discuss a few angles in more depth.

On performance variation. The Springer, acm and ieee refined dataset reach their highest F_1 of 95.1, 90.1 and 87.5, respectively. Examining each source' sentences, we see that the ones from Springer mostly includes useful information like editors' names and organisations with appropriate punctuation and HTML tags, but that those from acm and ieee are irregularly formatted and noisy. This also bears out in the correlation between the ratio of retrieved sentences to approximate number of editors (cf Table 1; "Ratio"): The more well-organized and text-based the source is, the higher the editor-to-sentence ratio, and the better the performance.

Effect of our proposed data refining process. We notice that for all dataset and embedding combinations, the refined dataset yields consistent positive performance increases, of 0.6 to 5.2 in F_1. This corresponds to an error reduction of at least 15%. Both of the processes of noise removal and data augmentation result in positive improvements. We note that the extent of improvement for ieee is much greater, indicating that the proposed refining process is more effective on lower-quality data sources (as measured by the proxy of editor-to-sentence ratio).

Effects of input representation. While both variations help to improve performance, applying subword embeddings via BPE consistently improves over vary word representations. BPE captures such internal features of words for testing on out-of-vocabulary words, especially names. These performance gains are aligned with better word representation, too; stacking BPE on a language model (LM) from a high-quality dataset significantly increases F_1. We note that LM derived from a low-quality dataset has a minimal or negative effect when applied to the higher-quality datasets.

Whole-editor tuple and per-class performance. Macro average F_1 does not give a complete picture. Correctly identifying all three (+ 1 optional) slots corresponding to a scholar's service role tuple is the real test: i.e., a contributing scholar's Name, Affiliation, Role (and Country). To assess this, we manually assessed a random validation set of 100 labeled editor instances from the Springer dataset. Our system correctly identifies 81 (*scholar*, *role*) tuples in full – showing that the performance loss of building correct tuples from correctly identified components is somewhat lower than an independent extrapolation ($95\%^3 \approx 85\% \geq 81\%$). While these results are only indicative, there is a non-trivial performance loss in composing whole service contribution records from individual components, and points to useful future work.

We also analyzed the component F_1 scores per slot class: 0.97, 0.951, 0.927 and 0.960 for the classes of Role, Person, Affiliation and Country, respectively. The above scores are over the best model RM1. We see that the classes with more variation — Affiliation and Person — are the performance bottlenecks. There is tally with our observation in the validation set, where Affiliation and Person are the lower bound for accuracy.

5 CONCLUSION

We improve person name recognition in mining editorial board roles from websites through cleaning and augmentation. These data refining methods address the challenge of limited data availability, and significantly improve dataset quality. We show editorial board information varies in quality and scale among three different publishers and also show that extraction performance correlates with the density of such labeled information with respect to input size.

We pair these data quality contributions with end-to-end training, varying the input representation through our enriched input representations (BPE subword and per-dataset embeddings). We prove that these refinements significantly and consistently reduce the recognition error by over 30% made by our baseline BiLSTM–CRF. These refinements increase the model performance to exceed a state-of-the-art $BERT_{base}$ baseline. In future work, we will apply such data refinement and augmentation to the transformer-based models. As our work refines the dataset and input representations, we believe our changes are model-agnostic and should apply well for these newer models.

REFERENCES

[1] Alan Akbik, Duncan Blythe, and Roland Vollgraf. 2018. Contextual string embeddings for sequence labeling. In *Proceedings of the 27th international conference on computational linguistics*. 1638–1649.

[2] Jingtao Han, Spyke Krepshaw, and Dongwon Lee. 2020. Gatekeeper: Analyzing G-Indexes and Improving Service Quantification. In *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020*. 17–26.

[3] Spyke Krepshaw and Dongwon Lee. 2019. Gatekeeper: Quantifying the Impacts of Service to the Scientific Community. In *International Conference on Theory and Practice of Digital Libraries*. Springer, 123–135.

[4] Hao Peng, Misha Teplitskiy, and David Jurgens. 2020. Author Mentions in Science News Reveal Wide-Spread Ethnic Bias. *arXiv preprint arXiv:2009.01896* (2020).

[5] Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on EMNLP*. 1532–1543.

[6] Abhinav Ramesh Kashyap, Devamanyu Hazarika, Min-Yen Kan, and Roger Zimmermann. 2021. Domain Divergences: A Survey and Empirical Analysis. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

[7] Erik F Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. *arXiv preprint cs/0306050* (2003).

[8] Karl-Michael Schneider. 2006. Information extraction from calls for papers with conditional random fields and layout features. *Artificial Intelligence Review* 25, 1-2 (2006), 67–77.

[9] Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015. Neural Machine Translation of Rare Words with Subword Units. *CoRR* abs/1508.07909 (2015). arXiv:1508.07909 http://arxiv.org/abs/1508.07909

[10] Shi Tian, Abhinav Ramesh Kashyap, and Min-Yen Kan. 2020. ServiceMarq: Extracting Service Contributions from Call for Papers. In *Proceedings of the ACM Symposium on Document Engineering 2020* (Virtual Event, CA, USA) *(DocEng '20)*. Article 20, 4 pages. https://doi.org/10.1145/3395027.3419596