Automatically Identify and Label Sections in Scientific Journals Using Conditional Random Fields

2

Sree Harsha Ramesh, Arnab Dhar, Raveena R Kumar, Anjaly V, Sarath K S, Jason Pearce, and Krishna R Sundaresan

OVERVIEW

We have developed a system that parses journal content in PDF format to extract article metadata and to identify major structural divisions and funding and supplementary information. The system is an adaptation for Task 2 of the ESWC Semantic Publishing Challenge 2016 of a larger machine learning system to convert unstructured Word manuscripts into full-text XML for journal frontlist production.

The system uses predominantly conditional random fields (CRF) for information extraction. CRF belongs to a class of probabilistic graphic models and is especially popular in sequence labelling because of the context-aware predictions it can be trained to make, unlike standard classifiers.

Feature Extraction. We used Apache PDFBox to extract typographical and positional information from the PDF, such as font weight, font style, and line length. NLTK was used for part-of-speech tagging and named entity recognition.

The Level 1 CRF predicts the main structural classes of the article: front, body, back, and floats-group.

The Level 2 CRF predicts the article title, the contributor group, the abstract, sections, acknowledgments, and table and figure captions.

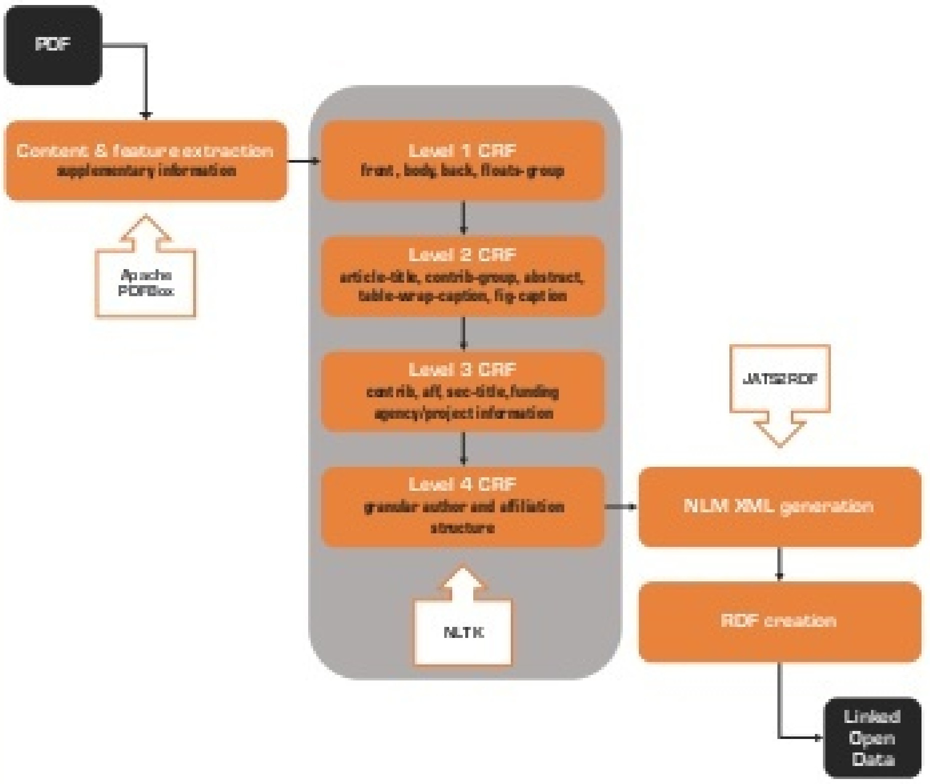
The Level 3 CRF separates the author and affiliation information within the contributor group; it also identifies funding agency/project information and section titles.

The Author Name CRF was trained on typographical features such as characterCase, tokenLength, and isSingleCapitalLetter and a keyword feature (tokenAsFeature) to predict the given-name and surname in lines predicted as contrib-group-contrib-name by the Level 3 CRF.

The Affiliation CRF predicts department, institution, street address, city, state, postal code, and country.

NLM XML and RDF Creation. NLM JATS XML is created from the output of the CRF models. To generate RDF, we used the JATS2RDF XSL transform with the SPAR (Semantic Publishing and Referencing) ontologies.

PIPELINE


PDF
Content & feature extraction
supplementary information
Apache PDFBox
Level 1 CRF
front, body, back, floats-group
Level 2 CRF
article-title, contrib-group, abstract, table-wrap-caption, fig-caption
Level 3 CRF
contrib, aff, sec-title, funding agency/project information
Level 4 CRF
granular author and affiliation structure
NLTK
JATS2RDF
NLM XML generation
RDF creation
Linked Open Data

EVALUATION: RESULTS OF SPARQL QUERIES

	Results for training set		Results for evaluation set		Performance of Cermine on evaluation set	
Query	Precision	Recall	Precision	Recall	Precision	Recall
Q1 Identify the affiliations of the authors	0.84	0.65	0.55	0.48	0.51	0.51
Q2 Identify the countries of the affiliations of the authors	0.87	0.75	0.74	0.71	0.85	0.84
Q3 Identify the supplementary material(s) for the paper	w.i.p.	w.i.p.	0.69	0.67	n.a.	n.a.
Q4 Identify the first-level sections of the paper	0.66	0.53	0.53	0.64	0.43	0.49
Q5 Identify the captions of the tables in the paper	0.58	0.25	0.75	0.89	n.a.	n.a.
Q6 Identify the captions of the figures in the papers	0.53	0.30	0.76	0.66	n.a.	n.a.
Q7 Identify the funding agencies that supported the research presented in the paper	0.90	0.40	0.48	0.48	n.a.	n.a.
Q8 Identify the EU project(s) that supported the research presented in the paper	0.70	0.60	0.70	0.70	n.a.	n.a.

NewGen
KnowledgeWorks