Using Sentence-level Classification Helps Entity Extraction from Material Science Literature

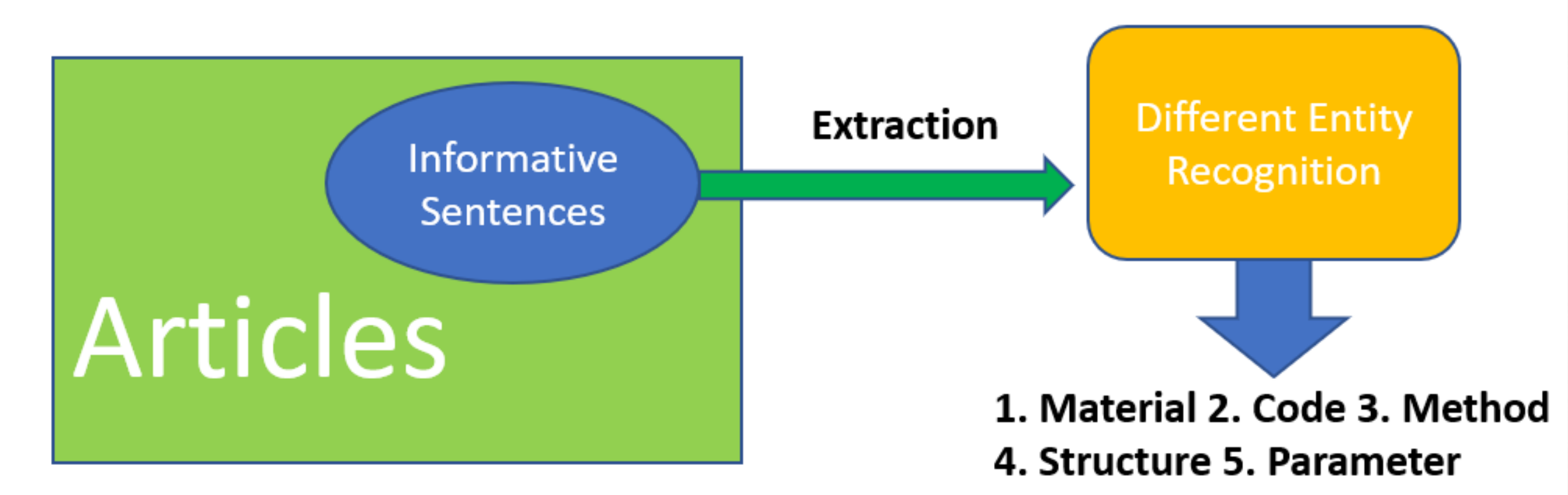
Ankan Mullick[1], Shubhraneel Pal[1], Tapas Nayak[2], Seong-Cheol Lee[3], Satadeep Bhattacharjee[3], Pawan Goyal[1]
[1]Department of Computer Science & Engineering, Indian Institute of Technology, Kharagpur, India
[2]TCS Research Lab, India, [3]Indo-Korea Science and Technology

LREC 2022
Marseille
13th edition
Le Palais du Pharo
June 20-25

Motivation

- Material Science research articles are a rich source of information about various entities related to material science such as names of the materials used for experiments, the computational software used along with its parameters, the method used in the experiments, etc.
- The distribution of these entities is not uniform across different sections of research articles
- Most of the sentences in the research articles do not contain any entity.

Figure 1


Articles
Informative Sentences
Extraction
Different Entity Recognition
1. Material 2. Code 3. Method
4. Structure 5. Parameter

Our Contributions

- In this work, we first use a sentence-level classifier to identify sentences containing at least one entity mention.
- Next, we apply the information extraction models only on the filtered sentences, to extract various entities of interest.
- Our experiments for NER in the material science research articles show that this additional sentence-level classification step helps to improve the F1 score by more than 4%.

Background

1. OS-CAR4 recognizer (Jessop et al., 2011) is an n-gram based Bayesian binary classifier that classifies tokens to 'chemical' or 'non-chemical' classes.

2. ChemSpot (Rockt ̈aschel et al., 2012), tmChem (Leaman et al., 2015), and ChemDataExtractor (Swain and Cole, 2016) are machine learning-based tools that can extract chemical entities from the chemistry literature.

3. (Hakimi et al., 2020) use machine learning-based NLP models for biomaterial text mining.

4. (Kim et al., 2017a; Kim et al., 2017b) apply information extraction and machine learning algorithms to extract the parameters of synthesis procedures from material science articles.

5. (Court and Cole, 2020) explore machine learning to extract transition temperatures and phase diagrams of magnetic materials and superconducting materials from text.

6. (Goldsmith et al., 2018) show how machine learning can be useful for aiding heterogeneous catalyst understanding, design and discovery (Mysore et al., 2017) extract graph structures from material science literature using neural network approaches.

7. (Guha et al., 2021) develop tool to generate database for material science literature.

8. Correa-Baena et al., 2018) study machine learning and natural language processing to accelerate the research of novel materials development.

Dataset

❑ We collect material science articles from (Guha et al., 2021) where total 10,500 articles.
❑ Articles are crawled from 'cond-mat.mtrl-sci' category with at least one code listed on https://psi-k.net/software
❑ We use spacy for tokenization and extract tokens with their labels.
❑ Out of 10,500 articles, 214 randomly selected articles are annotated using material science domain experts using Pdfanno.
❑ Two annotators annotate independently and Inter-annotator agreement (Cohen κ) is 0.81. Any conflict is resolved by the third annotator. Total annotation time is three weeks.
❑ Five informative entity types are labeled by annotators- a) material b) method c) code d) parameter e) structure.
❑ We label a sentence as "informative" if it contains an entity from any of the five class; otherwise the sentence is labeled as "uninformative". This dataset contains a total 15,699 (~ 31.64%) informative sentences among a total of 49,610 sentences.

Table 1

Entity type	Example	Count w.r.t. total entities	Percentage
CODE	BOLTZTRAP	304	1.75%
MATERIAL	EuCd2As2	9,161	52.74%
METHOD	DFT (Density Functional Theory	4,602	26.49%
PARAMETER	4*4*4 K-Point	1,387	7.98%
STRUCTURE	Hexagonal	1,918	11.04%
Total		17,372	100%

Table 2

Section	Inf	Code	Mat	Meth	Param	Struct
abstract	40.45	1.77	62.32	38.63	5.85	4.78
introduction	31.93	0.97	61.76	45.30	0.80	9.03
experiment	39.95	2.81	58.04	41.62	3.82	8.90
conclusion	28.46	0.57	61.60	39.38	6.34	7.97
other	31.51	3.65	48.30	40.87	10.82	10.95

Distribution of sentences in different sections of the articles to five types of entities in the annotated dataset.

Metrics

1. Precision, Recall, F1 score, and accuracy for our sentence identification models.
2. Accuracy is measured for informative and uninformative sentences together.

Experimental Results

- Precision (P), Recall (R), F1 score for informative sentences and overall Accuracy (A) [in %] with respective standard deviations (SD) of the models on the binary informative sentence identification task from entire articles.

Model	P, SD(P)	R, SD(R)	F1, SD(F1)	A, SD(A)
NB	74.08, 0.92	78.1, 0.92	76.03, 0.91	85.11, 0.42
SVM	61.17, 0.8	55.91, 0.56	58.42, 0.56	75.93, 0.39
LR	91.19, 0.78	77.61, 1.39	83.85, 0.89	91.16, 0.45
RF	92.75, 0.35	77.71, 0.92	84.56, 0.66	91.42, 0.33
Bg	92.06, 0.88	69.04, 1.9	78.91, 1.02	88.83, 0.39
BiLSTM (M2V)	85.61, 0.33	86.4, 0.35	86.01, 0.18	87.8, 0.06
CNN (Sci)	91.27, 1.56	92.76, 0.64	92.01, 0.84	92.10, 1.01
BERT	**98.69**, 0.24	**97.67**, 0.22	**98.18**, 0.13	**98.84**, 0.08
SciBERT	90.16, 2.72	91.45, 0.29	90.81, 1.43	92.03, 1.16
DistilBERT	98.54, 0.17	97.29, 0.33	97.92, 0.11	98.75, 0.06

- For traditional classifiers, following Four categories of features are used –

(i) Parts of speech (POS) tag-based features: We use Stanford POS tagger (Manning et al., 2014) to find the number of nouns, verbs, adjectives, presence of adverbs, etc.
(i) Tf-Idf based features: n-gram (one, two etc.) based features.
(ii) Dependency parse based features (using Stanford Dependency parser (De Marneffe and Manning, 2008)): dobj (direct object), amod (adjective modifier) etc.
(iii) Others: no. of characters, presence of wh-words, numbers, strong, weak adjectives etc.

- Precision, Recall, F1-score [in %] of Different NERs for entity Extraction from all sentences and only informative Sentences in the articles.

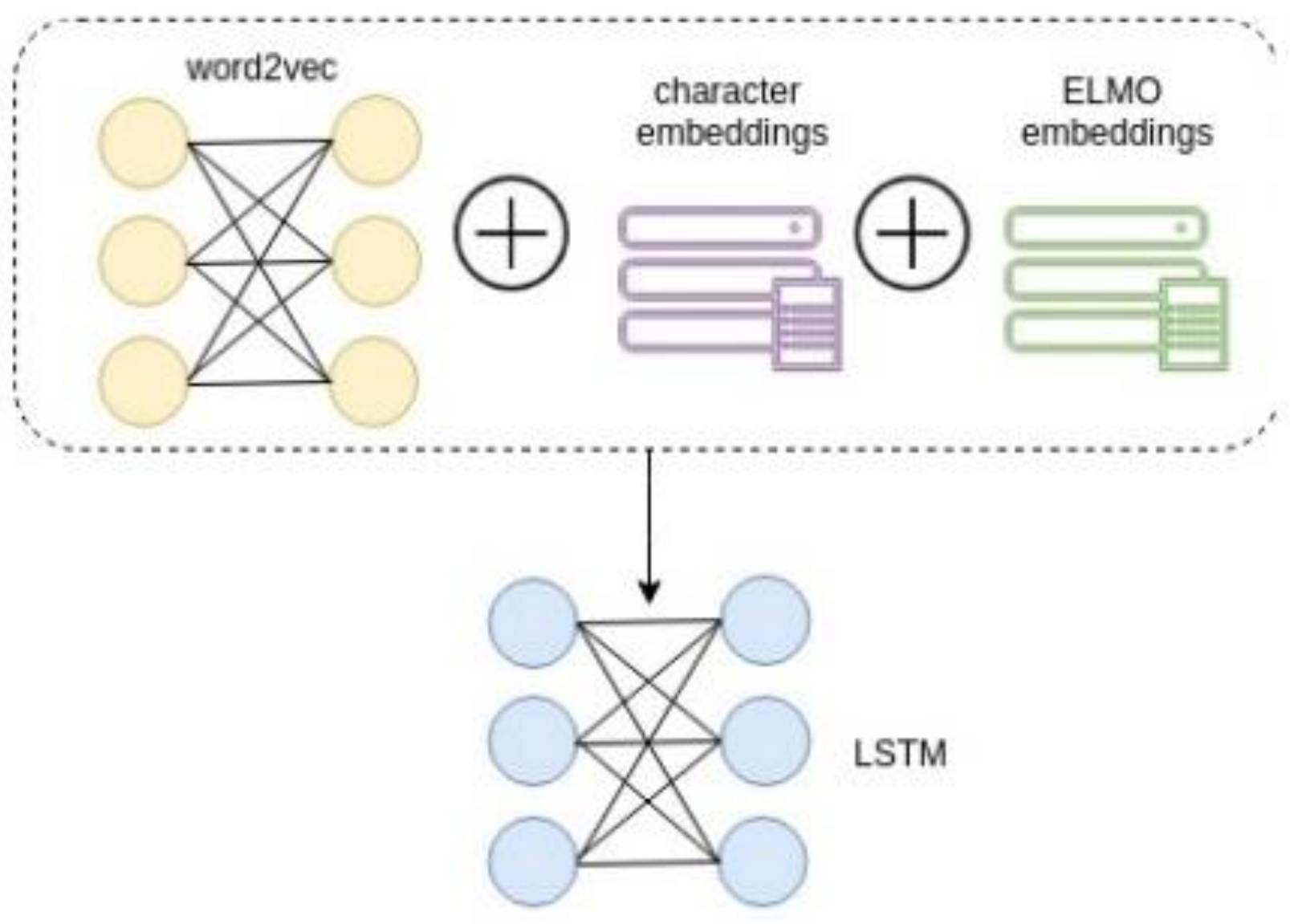
Table 3

	Precision		Recall		F1	
Method	All	Inf	All	Inf	All	Inf
Sci-BERT (Beltagy et al., 2019)	77.32	81.93	70.24	71.21	73.61	76.19
BERT (Devlin et al., 2018)	74.08	76.45	70.72	72.86	72.36	74.61
DistilBERT (Sanh et al., 2019)	71.87	72.69	69.48	70.59	70.66	71.62
DCNN (Mysore et al., 2017)	79.43	83.02	78.89	79.53	79.15	81.24
BiLSTM-CRF (Mysore et al., 2017)	81.56	84.29	79.37	80.52	80.45	82.36
BiLSTM (Weston et al., 2019)	79.61	82.53	73.82	75.98	76.60	79.12
MGM (Da San Martino et al., 2019)	74.33	-	70.91	-	72.58	-
SC-NER (Wang et al., 2019)	75.64	-	79.12	-	77.34	-
Mimicking (Guha et al., 2021)	82.08	85.93	83.72	86.08	82.89	86.01
BiLSTM-CRF Elmo	87.35	**91.71**	82.19	**86.09**	84.57	**88.76**

- **Baselines:**
  1. Scientific text embedding SciBERT (Beltagy et al., 2019).
  2. Fine tuned version of BERT (Devlin et al., 2018) (uncased with linear model and base)
  3. Distil-BERT (Sanh et al., 2019) for classification (using CLS embedding).
  4. DCNN (Diluted CNN) and Bi-LSTM-CRF model by (Mysore et al., 2017).
  5. BiLSTM NER for specific material science articles by (Weston et al., 2019).
  6. Multi-Granularity model (MGM) (Da San Martino et al., 2019) – joint entity extraction
  7. SC-NER (Wang et al., 2019) joint entity extraction
  8. Bi-LSTM CRF with noise (Mimicking Model) by (Guha et al., 2021).

- **Our Approach:** Figure 2

We use pre-trained Elmo (Peters et al., 2018) embedding for material science articles. The input to the Bi-LSTM-CRF model is thus a concatenation of pre-trained Word2Vec embedding (Mikolov et al., 2013),character embedding, and pre-trained ELMO embedding (Peters et al., 2018) along with the IOB tags as the target of each word. We fine-tune the model with Adam optimizer, dropout of 0.5, hidden dimension of 200, number of epochs at 120, and a batch size of 8 to get the overall optimum (for all entities together) F1 score.



- Entity-wise Precision, Recall, F1 score [in %] of BiLSTM-CRF ELMO NER for entity extraction from all sentences and only informative sentences).

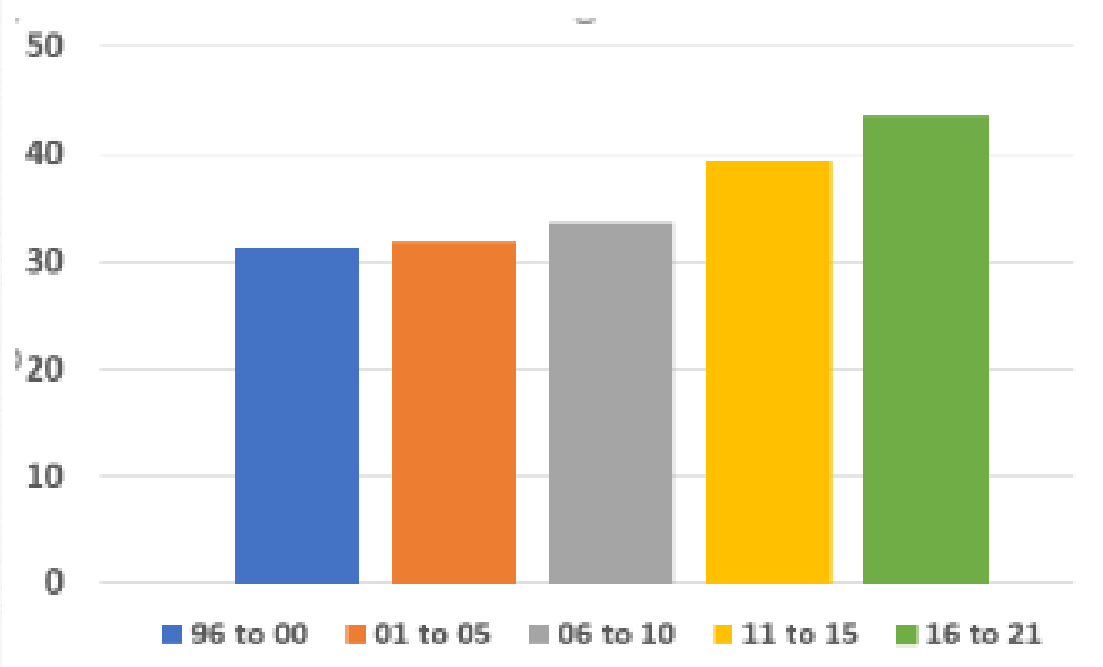
	Precision		Recall		F1	
Entity type	All	Inf	All	Inf	All	Inf
MATERIAL	85.49	91.66	88.44	90.08	86.94	90.86
METHOD	95.53	96.03	82.80	86.72	88.71	91.13
STRUCTURE	95.92	96.71	89.81	93.87	92.76	95.27
PARAMETER	73.27	81.76	62.51	70.31	67.46	75.60
CODE	86.54	92.39	87.38	89.47	86.96	90.91
Overall	87.35	**91.71**	82.19	**86.09**	84.57	**88.76**

Table 4

- Distribution of the sentences (in %) in different sections of the articles from the unannotated dataset from five types of entities predicted by 2-stage model. Randomly selected 7998 articles with 1.9 million sentences and 0.675 Million informative sentences.

Section	Inf	Code	Mat	Meth	Param	Struct
abstract	36.12	0.15	66.87	35.29	4.48	2.97
introduction	37.41	0.17	59.42	40.71	9.56	3.42
experiment	35.01	1.87	65.41	23.51	12.41	4.26
conclusion	34.75	0.56	63.97	38.65	7.60	3.11
other	30.99	1.18	65.75	28.09	12.61	3.22

Table 5



Yearly distribution [in %] of informative sentences as predicted by the BERT model on the unannotated dataset

Figure 3

Conclusion

- We propose deep neural network-based models to classify sentences into these two classes concerning five types of entities such as material, method, code, parameter, and structure.
- Our experiments show that the two-stage framework (identify informative sentences and then extract entity) leads to significant improvement in the performance of the end task of extracting five types of entities from the articles than direct extraction of entities from all sentences.

Email: ankanm@kgpian.iitkgp.ernet.in

Acknowledgement: We would like to thank the "**LREC'22**" organizers for accepting our work & PMRF Fellowship for awarding us with the student travel grant