Fine-Grained Named Entities for Corona News - Presentation

Efeoglu, Sefika | Paschke, Adrian

Version: Postprint (Verlagsversion)/Postprint (Publisher Version)
Typ/Type: Kongressschrift/Conference Proceeding
Jahr/year: 2023
Quelle/Source: https://repository.publisso.de/resource/frl:6440380
Schlagwörter/Keywords: named entity recognition, corona news, fine-grained entities, contextual embedding

Zitationsvorschlag/ Suggested Citation:
Efeoglu, Sefika; Paschke, Adrian (2023): Fine-Grained Named Entities for Corona News - Presentation. International SWAT4HCLS Conference 2023. DOI: 10.4126/FRL01-006440380

Fine-Grained Named Entities for Corona News

Sefika Efeoglu and Adrian Paschke

Freie Universität Berlin

February 15, 2023

Outline

1. Introduction
2. Methodology
3. Evaluation
4. Conclusion
5. Bibliography

Motivation

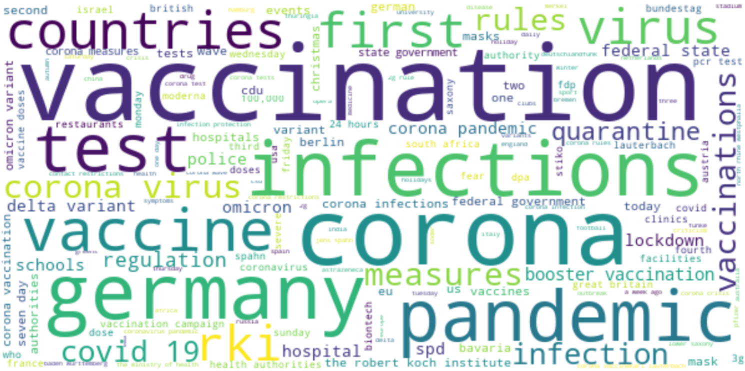


Figure 1: The word cloud of corona news corpus from tagesschau.

Introduction

Problem:

- Huge amount of unstructured text data in the corona domain since December 2019.
- Analyzing these unstructured texts is time-consuming.

Drawbacks of existing corpora:

The existing corpora, such as CORD-19 [1] and LitCovid [2]:

- fail to identify recent variants of the coronavirus and generic mentions.
- include earlier published scientific papers in this domain.

Approach:

This study aims to develop an annotation pipeline that generates annotated training data from newer corona news articles for named entity recognition (NER).

Introduction

Problem:

- Huge amount of unstructured text data in the corona domain since December 2019.
- Analyzing these unstructured texts is time-consuming.

Drawbacks of existing corpora:

The existing corpora, such as CORD-19 [1] and LitCovid [2]:

- fail to identify recent variants of the coronavirus and generic mentions.
- include earlier published scientific papers in this domain.

Approach:

This study aims to develop an annotation pipeline that generates annotated training data from newer corona news articles for named entity recognition (NER).

Introduction

Problem:

- Huge amount of unstructured text data in the corona domain since December 2019.
- Analyzing these unstructured texts is time-consuming.

Drawbacks of existing corpora:

The existing corpora, such as CORD-19 [1] and LitCovid [2]:

- fail to identify recent variants of the coronavirus and generic mentions.
- include earlier published scientific papers in this domain.

Approach:

This study aims to develop an annotation pipeline that generates annotated training data from newer corona news articles for named entity recognition (NER).

An Annotation Pipeline for Training Texts

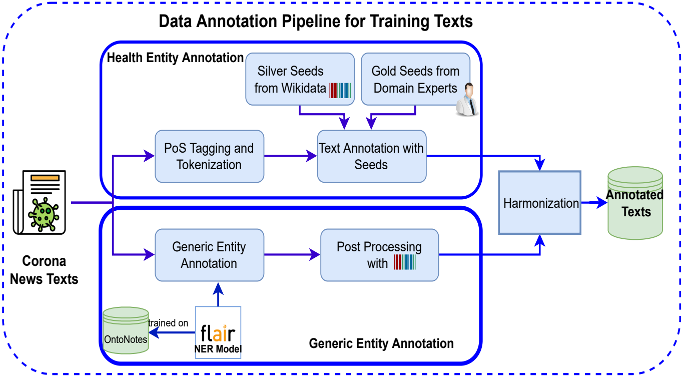
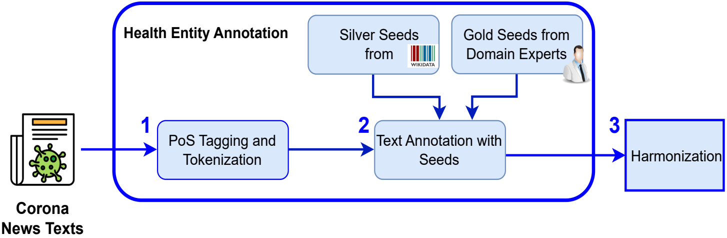

Data Annotation Pipeline for Training Texts
Health Entity Annotation
Silver Seeds from Wikidata
Gold Seeds from Domain Experts
PoS Tagging and Tokenization
Text Annotation with Seeds
Corona News Texts
Harmonization
Annotated Texts
Generic Entity Annotation
Post Processing with
OntoNotes
trained on
flair
NER Model
Generic Entity Annotation

Figure 2: Data Annotation Pipeline for Training Texts.

Health Entity Annotation

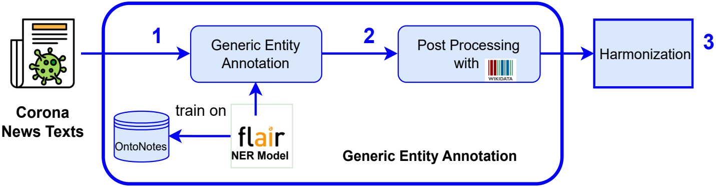

Health Entity Annotation
Silver Seeds from
WIKIDATA
Gold Seeds from Domain Experts
1
PoS Tagging and Tokenization
2
Text Annotation with Seeds
3
Harmonization
Corona News Texts

Input Sentence:
According to the Berlin virologist Christian Drosten, an unvaccinated person with an Omicron infection carries three quarters of the risk of being hospitalized for an unvaccinated person with the delta variant of Corona.

Output (Input to Step 3):
According to the Berlin virologist Christian Drosten, an unvaccinated person with an Omicron[CORONAVIRUS] infection[DISEASE_OR_SYNDROME] carries three quarters of the risk of being hospitalized for an unvaccinated person with the delta variant[CORONAVIRUS] of Corona[CORONAVIRUS].

Figure 3: Health Entity Annotation.

Generic Entity Annotation

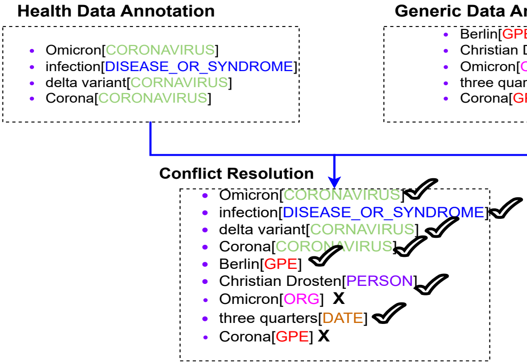


Input Sentence:
According to the Berlin virologist Christian Drosten, an unvaccinated person with an Omicron infection carries three quarters of the risk of being hospitalized for an unvaccinated person with the delta variant of Corona.

Output (Input to Step 3):
According to the Berlin[GPE] virologist Christian Drosten[PERSON], an unvaccinated person with an Omicron[ORG] infection carries three quarters[DATE] of the risk of being hospitalized for an unvaccinated person with the delta variant of Corona[GPE].

Figure 4: Generic Entity Annotation.

Harmonization



Health Data Annotation

- Omicron[CORONAVIRUS]
- infection[DISEASE_OR_SYNDROME]
- delta variant[CORNAVIRUS]
- Corona[CORONAVIRUS]

Generic Data Annotation

- Berlin[GPE]
- Christian Drosten[PERSON]
- Omicron[ORG]
- three quarters[DATE]
- Corona[GPE]

Conflict Resolution

- Omicron[CORONAVIRUS] ✓
- infection[DISEASE_OR_SYNDROME] ✓
- delta variant[CORNAVIRUS] ✓
- Corona[CORONAVIRUS] ✓
- Berlin[GPE] ✓
- Christian Drosten[PERSON] ✓
- Omicron[ORG] X
- three quarters[DATE] ✓
- Corona[GPE] X

The Annotated Sentence:
According to the Berlin[GPE] virologist Christian Drosten[PERSON],
an unvaccinated person with an Omicron[CORONAVIRUS] infection[DISEASE_OR_SYNDROME]
carries three quarters[DATE] of the risk of being hospitalized for an unvaccinated person
with the delta variant[CORONAVIRUS] of Corona[CORONAVIRUS].

Figure 5: Harmonization

Experimental Setup

- **Dataset:** corona-related news articles from a German news-channel “Tagesschau” between December 2020 and June 2022.
- **Fleiss Kappa**: 0.98 (test) (calculated for event, product, immune_response, coronavirus, disease_or_sydrome, sign_or_symptom, ‘empty’).
- **NER models**: base (Glove) [3], advanced (Flair+Glove) [4] and SciBERT [5].

Corpus	# of sentences
Training	89986
Dev	4999
Test	1000

Table 1: The entities in the data set have been categorized with 23 entity types.

Results

- Fine-tuned SciBERT [5] model's micro F1-score is 0.7765.
- Its entity-specific F1-scores are 0.81 (coronavirus), 0.84 (sign_or_symptom), 0.79 (disease_or_syndrome), 0.8 (immune_response), and 0.85 (group).

Embedding	Model	Model Std	Coronavirus	Disease or Syndrome	Group	Immune Response	Sign or Symptom
Glove	0.71084	0.003414	0.76522	0.84152	0.80078	0.96364	0.81922
Glove+Flair	0.77162	0.002322	0.78614	0.81214	0.85016	0.83264	0.86562

Table 2: This table shows the statistical details about mean micro-F1 scores of the NER models (implemented by using Flair framework [6]), which were trained and evaluated five times. Besides, the table gives the mean micro-F1 scores of new entity types on the models trained with our corona news corpus.

F1 Scores of Specific Entities

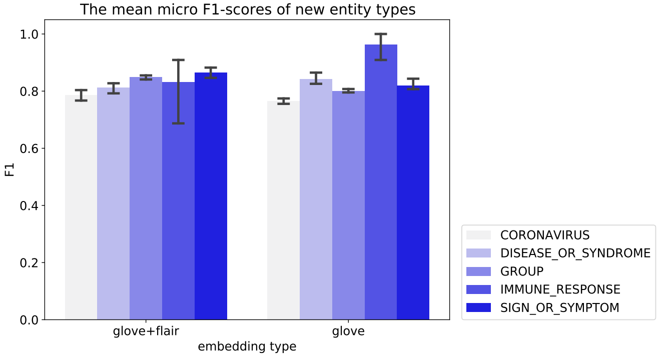


Figure 6: F1 scores of the new entities.

Conclusion

- Contributions:
  1. An annotation pipeline to create **annotated texts** from the corona news articles for NER.
  2. A **new up-to-date annotated corpus** in the corona domain to identify corona-related mentions on the corona news articles via the NER models.
- The models utilizing contextual embedding surpass the model using an only word embedding in terms of micro-F1 score.
- Besides, the fine-tuned SciBERT model has performed well in the domain-specific entity types.

Acknowledgements

The research presented in this presentation was supported in part by the German Federal Ministry of Education and Research (BMBF) project "PANQURA" under grant 03COV03F, in part by the European Union project "FAST-LISA" under grant 101049342.

Thank you!

sefika.efeoglu@fu-berlin.de

References I

- X. Wang, X. Song, B. Li, Y. Guan, and J. Han, "Comprehensive named entity recognition on cord-19 with distant or weak supervision," 2020.
- Q. Chen, A. Allot, and Z. Lu, "LitCovid: an open database of COVID-19 literature," *Nucleic Acids Research*, vol. 49, pp. D1534–D1540, 11 2020.
- J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1532–1543, 2014.
- A. Akbik, D. Blythe, and R. Vollgraf, "Contextual string embeddings for sequence labeling," in *COLING 2018, 27th International Conference on Computational Linguistics*, pp. 1638–1649, 2018.
- I. Beltagy, K. Lo, and A. Cohan, "Scibert: Pretrained language model for scientific text," in *EMNLP*, 2019.
- A. Akbik, T. Bergmann, D. Blythe, K. Rasul, S. Schweter, and R. Vollgraf, "FLAIR: An easy-to-use framework for state-of-the-art NLP," in *NAACL 2019, 2019 Annual Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pp. 54–59, 2019.