TNNT: The Named Entity Recognition Toolkit

Sandaru Seneviratne[1,4][0000−0001−8320−5084], Sergio J. Rodríguez Méndez[1,2][0000−0001−7203−8399], Xuecheng Zhang[1,4], Pouya G. Omran[1,3][0000−0002−4473−3877], Kerry Taylor[1,4][0000−0003−2447−1088], and Armin Haller[1,4][0000−0003−3425−0780]

[1] Australian National University, Canberra ACT 2601, AU
[2] `Sergio.RodriguezMendez@anu.edu.au`
[3] `P.G.Omran@anu.edu.au`
[4] {`firstname.lastname`}`@anu.edu.au`
`https://cs.cecs.anu.edu.au/`

Abstract. Extraction of categorised named entities from text is a complex task given the availability of a variety of Named Entity Recognition (NER) models and the unstructured information encoded in different source document formats. Processing the documents to extract text, identifying suitable NER models for a task, and obtaining statistical information is important in data analysis to make informed decisions. This paper presents[5] *TNNT*, a toolkit that automates the extraction of categorised named entities from unstructured information encoded in source documents, using diverse state-of-the-art Natural Language Processing (NLP) tools and NER models. *TNNT* integrates 21 different NER models as part of a Knowledge Graph Construction Pipeline (KGCP) that takes a document set as input and processes it based on the defined settings, applying the selected blocks of NER models to output the results. The toolkit generates all results with an integrated summary of the extracted entities, enabling enhanced data analysis to support the KGCP, and also, to aid further NLP tasks.

Keywords: Information Extraction · Named Entity Recognition · Natural Language Processing · Knowledge Graph Construction Pipeline

1 Introduction

NER is a major component in NLP systems to extract information from unstructured text. Recent advances in Deep Learning and NLP have resulted in the availability of a large number of NER tools and models for use which have enabled NER of different categories from text. However, given the existence of a wide range of document formats, extracting information is difficult considering the pre-processing required prior to using NER tools and the challenge of

[5] The manuscript follows guidelines to showcase a demonstration that introduces an overview of how the toolkit works: input document set, initial settings, processing, and output set. The input document set is artificial in order to show various toolkit capabilities.

Table 1. Tools and models integrated in *TNNT*

#	Tool	Number of Models
1	NLTK [6]	1
2	spaCy[7]	3 (en_core_web_sm, en_core_web_md, en_core_web_lg)
3	Stanford NER [7]	3 (3-class model, 4-class model, 7-class model)
4	Stanza [8]	1
5	Flair [1]	5 (ner, ner-fast, ner-pooled, ner-ontonotes, ner-ontonotes-fast)
6	Allen NLP [5]	2 (Elmo-based NER, Fine-grained NER)
7	Polyglot [2]	1
8	Deeppavlov [3]	4 (ner_conll2003, ner_ontonotes, ner_conll2003_bert, ner_ontonotes_bert)
9	NER based on BERT [4]	1

identifying which models to use. Having a system which provides easy processing of different document formats, easy selection of different models or tools, an integrated summary of the entities identified by the models and an API which enables basic functionalities to access the results of the models can enhance data analysis, accurate decisions and provide a thorough overview of the data used.

This paper introduces *TNNT*[6]. Its main goal is to automate the extraction of categorised named entities from the unstructured information encoded in the source documents, using recent state-of-the-art NLP-NER tools and models. *TNNT* is integrated with the "Metadata Extractor & Loader" (*MEL*) [9] which implements a set of methods to extract metadata (and content-based information) from various file formats.

2 Core Features

TNNT integrates 21 different NER models from 9 state-of-the-art NLP tools (Table 1). These 21 models can identify up to 18 categories (Table 3) of named entities in text. The system is capable of processing different models sequentially based on the input settings (processing blocks) defined by the user. All textual content extracted by *MEL* is processable for *TNNT* with a hybrid processing data flow, either from/to a document store[8] or via direct processing from files.

For data analysis tasks, TNNT keeps general statistics of the models and generates an integrated summary of all the identified entities. The results are JSON[9] files (one for each processed source document) with the list of models, categories, and identified entities. For each recognised entity, the toolkit retrieves its context information and the start/end index in the document text[10]. Table 2 gives an overview of the results obtained using some of the models for two publicly available datasets: CONLL 2003[11] and NIST IE-ER[12].

Additionally, a built-in RESTful API provides basic functions to browse the results and to complement them by performing other NLP tasks, such as part-of-

[6] The project's URI is `https://w3id.org/kgcp/MEL-TNNT`. All resources along with demo videos are available at this address.

[8] Currently, *TNNT* only supports CouchDB (`https://couchdb.apache.org/`)

[9] `https://www.json.org/`

[10] Sample results can be found at the project's `w3id` URI.

[11] `https://www.clips.uantwerpen.be/conll2003/ner/`

[12] `https://github.com/juand-r/entity-recognition-datasets`

Table 2. *TNNT* results from some NER models for two public datasets

Model	Dataset	Exec. Time (seconds)	Number of Recognised Entities
Stanford-3 class model	CONLL 2003	17.16	location:2165, organisation:2586, person:2726 (Total = 7477)
	NIST IE-ER	7.55	location:403, organisation:431, person:831 (Total = 1665)
Spacy-encore_web_md	CONLL 2003	36.82	location:112, organisation:2047, person:2921, NORP:931, FAC:90, GPE:3015, product:62, event:221, work_of_art:43, law:11, language:21, date:2890, time:266, percent:138, money:129, quantity:141, ordinal:367, cardinal:3469 (Total = 16874)
	NIST IE-ER	14.55	location:102, organisation:1184, person:1675, NORP:380, FAC:57, GPE:707, product:41, event:37, work_of_art:53, law:10, language:7, date:771, time:112, percent:48, money:23, quantity:37, ordinal:118, cardinal:609 (Total = 5971)
BERT-based	CONLL 2003	1245.66	location:2312, organisation:2450, person:2723, miscellaneous:1381 (Total = 8866)
	NIST IE-ER	662.27	location:792, organisation:806, person:1269, miscellaneous:672 (Total = 3539)

Table 3. Categories identified by the models integrated in *TNNT*

Category	Description
PERSON	People, including fictional
NORP	Nationalities or religious or political groups
FAC	Buildings, airports, highways, bridges, etc.
ORG	Companies, agencies, institutions, etc.
GPE	Countries, cities, states
LOCATION	Non-GPE locations, mountain ranges, bodies of water
PRODUCT	Objects, vehicles, foods, etc. (Not services.)
EVENT	Named hurricanes, battles, wars, sports events, etc.
WORK_OF_ART	Titles of books, songs, etc.
LAW	Named documents made into laws
LANGUAGE	Any named language
DATE	Absolute or relative dates or periods
TIME	Times smaller than a day.
PERCENT	Percentage, including "%"
MONEY	Monetary values, including unit
QUANTITY	Measurements, as of weight or distance
ORDINAL	"first", "second", etc
CARDINAL	Numerals that do not fall under another type

speech tagging, dependency parsing, and co-reference resolution. These functionalities along with the comprehensive information provided by *TNNT*, facilitate the understanding of the models and data used for NLP and KGCP tasks.

3 Architecture

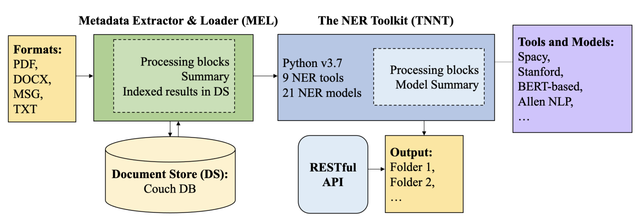

Fig. 1. *TNNT* Architecture

TNNT has been fully integrated with *MEL* (Figure 1). *MEL* settings establish the way *TNNT* will process some specific block sequence of NER models for the input dataset (either from content stored on a document store or from a direct document processing immediately after metadata extraction). More design details can be found at the project's `w3id` URI.

4 Conclusions and Future Work

TNNT provides a simple mechanism to extract categorised named entities from unstructured data using a diverse range of state-of-the-art NLP tools and NER models. The tool is still in its early stages of development. It has been tested using different document formats and datasets as part of the "Australian Government Records Interoperability Framework" (AGRIF) project. There are ongoing plans to integrate more NER tools and models into the architecture along with continuing improvements to the RESTful API with complementary NLP tasks to enrich the NER results, in order to support KGCP tasks. The major contributions of this tool are: (1) the ability to process different source document formats for NER; (2) the availability of 21 different state-of-the-art NER models integrated in one system, enabling easy selection of models for NER; (3) the provision of an integrated summary of the results from different models; and (4) a RESTful API that enables easy access to the NER results from the models.

References

1. Akbik, A., Bergmann, T., Blythe, D., Rasul, K., Schweter, S., Vollgraf, R.: FLAIR: An easy-to-use framework for state-of-the-art NLP. In: 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations). pp. 54–59 (2019)
2. Al-Rfou, R., Kulkarni, V., Perozzi, B., Skiena, S.: POLYGLOT-NER: Massive Multilingual Named Entity Recognition. In: Venkatasubramanian, S., Ye, J. (eds.) SDM. pp. 586–594. SIAM (2015), `http://dblp.uni-trier.de/db/conf/sdm/sdm2015.html#Al-RfouKPS15`
3. Burtsev, M., Seliverstov, A., Airapetyan, R., Arkhipov, M., Baymurzina, D., Bushkov, N., Gureenkova, O., Khakhulin, T., Kuratov, Y., Kuznetsov, D., et al.: Deeppavlov: Open-source library for dialogue systems. In: Proceedings of ACL 2018, System Demonstrations. pp. 122–127 (2018)
4. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv:1810.04805 (2018)
5. Gardner, M., Grus, J., Neumann, M., Tafjord, O., Dasigi, P., Liu, N., Peters, M., Schmitz, M., Zettlemoyer, L.: Allennlp: A deep semantic natural language processing platform. arXiv:1803.07640 (2018)
6. Loper, E., Bird, S.: NLTK: the Natural Language ToolKit. arXiv preprint cs/0205028 (2002)
7. Manning, C.D., Surdeanu, M., Bauer, J., Finkel, J.R., Bethard, S., McClosky, D.: The Stanford CoreNLP natural language processing toolkit. In: Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations. pp. 55–60 (2014)
8. Qi, P., Zhang, Y., Zhang, Y., Bolton, J., Manning, C.D.: Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. In: The 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations. pp. 101–108. Online (2020). https://doi.org/10.18653/v1/2020.acl-demos.14
9. Rodríguez Méndez, S.J., Omran, P.G., Haller, A., Taylor, K.: MEL: Metadata Extractor & Loader. In: ISWC 2021: P&D Track (2021), `http://ceur-ws.org/Vol-2980/paper399.pdf`