UNER: Universal Named-Entity Recognition Framework

Diego Alves[1], Tin Kuculo[2], Gabriel Amaral[3], Gaurish Thakkar[1], and Marko Tadić[1]

[1] Faculty of Humanities and Social Sciences, University of Zagreb, Zagreb 10000, Croatia {dfvalio,marko.tadic}@ffzg.hr, gthakkar@m.ffzg.hr
[2] L3S Research Center, Leibniz University Hannover, Hannover, Germany kuculo@l3s.de
[3] King's College London, London, United Kingdom gabriel.amaral@kcl.ac.uk

Abstract. We introduce the Universal Named-Entity Recognition (UNER) framework, a 4-level classification hierarchy, and the methodology that is being adopted to create the first multilingual UNER corpus: the SETimes parallel corpus annotated for named-entities. First, the English SETimes corpus will be annotated using existing tools and knowledge bases. After evaluating the resulting annotations through crowdsourcing campaigns, they will be propagated automatically to other languages within the SETimes corpora. Finally, as an extrinsic evaluation, the UNER multilingual dataset will be used to train and test available NER tools. As part of future research directions, we aim to increase the number of languages in the UNER corpus and to investigate possible ways of integrating UNER with available knowledge graphs to improve named-entity recognition.

Keywords: named-entity · universality · low-resourced languages

1 Introduction

In the span of little more than a year, with pre-trained language models becoming ubiquitous in the field of Natural Language Processing (NLP), a wide range of tasks have received renewed interest; not the least of which is information extraction. Information extraction (IE) is the task of automatically extracting structured information from natural language texts. A common sub-task of information extraction includes Named Entity Recognition and Classification (NERC) [1].

NERC corpora usually respond to the specific needs of local projects in terms of the complexity of the annotation hierarchy and its format. Table 1 shows some of the various NERC annotation schemes that have been proposed in previous research efforts.

Table 1. Description of existing NE annotation schemes in terms of hierarchy complexity.

Source	Number of Levels	Nodes per level
MUC-7[5] (English)	2	3/8
CoNLL 2003 [18] (English and German)	1	4
spaCy [9] (based on OntoNote5, English)	1	18
Czech Named Entity Corpus 2.0 [20] (Czech)	2	8/46
NKJP corpus[17] (Polish)	2	6/8
Second Harem [7] (Portuguese)	3	10/36/21
Sekine [19] (English v.7.1.0)	4	3/28/87/125

With the aim of creating a universal multilingual annotation NE scheme, we follow previous work (Table 1) to propose the Universal Named Entity Recognition (UNER) framework, a four-level NE classification hierarchy.

We intend to use the proposed framework to annotate the multilingual SETimes parallel corpora[2] (described in [23]) and create a multilingual named entity recognition dataset.

This paper is organised as follows: In Section 2 we describe our approach to defining the UNER hierarchy; Section 3 describes our strategy for the creation of the first multilingual open-source UNER corpora; and in Section 4 we provide conclusions and possible future directions for research, including applications of UNER corpora and increasing UNER multilingualism.

2 UNER Tagging Framework Definition

The UNER hierarchy is built upon the NER hierarchy proposed by Sekine [19], which presents the highest conceptual complexity between the compared NER schemes (Table 1). Maintaining its main structure, UNER is organised through one root node ("TOP"), and three first-level leaf nodes ("Name", "Timex TOP" and "Numex") which correspond to MUC-7[5] main categories. Analysing each of the child nodes, we make the following changes to compose the UNER hierarchy:

- In Sekine's hierarchy, "Person" is a child node of "Name" and has no ramifications. In UNER, "Name Other" and "God" have become child nodes of "Person", their names changed to "Other" and "Entity". In addition to that, "Person" gets new child nodes: "Name", "Profession" and "Fictional".
- Concerning the leaf node "Location" (inside "Name"), we have introduced a new child node "Fictional" and have removed "Phone Number" which was added as a child node of "Numex".
- Inside the "Product" node's ramifications, we have moved "ID Number" to the "Numex" node, suppressed "Character" and "Title" as these nodes were replaced by the "Profession" and "Fictional" nodes respectively, inside the parent node "Person". We have also added a child leaf node "Brand" inside

[2] http://nlp.ffzg.hr/resources/corpora/setimes/

the nodes "Clothing", "Drugs", "Food", "Vehicles" and "Weapon". Brands that do not correspond to these categories will be annotated as "Product Other", a child node of "Product".
- Inside "Event", a child node of "Name", we have introduced a new child node "Personal" concerning personal facts such as births and weddings. We have also renamed the node "Incident" as "Historical Event" and have added a child node named "Other" inside it. In Sekine's hierarchy, historical events (e.g. French Revolution) were classified as "Event Other" inside "Event".
- Concerning the leaf node "Timex TOP" which corresponds to time expressions, we have added "Holiday" as a child node inside "Timex". We have also introduced a child node named "Timex Relative" with the same ramifications as the node "Timex". The idea is to differentiate between absolute time expressions such as "April 1st, 2003" and relative expressions as in "last August".

We have focused only on the nodes inside Sekine's hierarchy; the attributes proposed by the authors to increase the knowledge (for example: "Place of Origin" as an attribute of the node "River") were not taken into account in this version of the UNER framework.

As previously explained, UNER follows Sekine's NER hierarchy. It is composed of a root node "TOP" and three child nodes: "Name", "Timex TOP" and "Numex", each of which contain several ramifications. The number of nodes of each UNER level is described in Table 2.

Due to the size of the UNER hierarchy scheme, we have decided to present it fully in digital form[3].

Table 2. Description of the number of nodes per level inside UNER hierarchy.

Level	**Number of nodes**
0 (root)	1
1	3
2	29
3	95
4	129

2.1 Format

Available open-source NER tools can be broadly classified into 3 main types based on the format they support. These are BIO (Begin-Inside-Outside), inline XML-tags (<PERSON> George Clooney </PERSON>), or Index-based spans (text:'George Clooney', start:0, end:14, label:'Person'). Since we want the dataset to support existing training frameworks, we performed a study on existing available NER tools and their corresponding support for various input

[3] https://tinyurl.com/sb3u9ve

formats. This is summarised in Table 3. The input column is the support of the tool for taking in data for training a new model and output denotes the final output format when new text is passed through the tagging process. Cells marked as "-" refer to tools that do not support training and only have a prediction functionality. In the deployment phase of UNER, we intend to use BIO and Index-based span format to represent the NEs, but all formats are convertible into each other with simple scripts, which is important to guarantee its universality.

Table 3. Existing NER tools and their support for various formats.

Tool	BIO		XML		Index-based	
	Input	Output	Input	Output	Input	Output
Polyglot [4]	-	Yes	-	-	-	-
NER BERT [12]	Yes	Yes	No	No	No	No
Stanford NER [13]	Yes	Yes	No	Yes	No	Yes
ESNLTK [15]	No	Yes	No	No	Yes	Yes
Finnish-tagtools [8]	-	No	-	Yes	-	No
Spacy [9]	Yes	Yes	No	No	Yes	Yes
Poldeepner [14]	Yes	No	No	No	No	Yes
BILSTM-CRF-CHAR [2]	Yes	Yes	No	No	No	No
Stagger [16]	-	No	No	Yes	-	No
Swener [11]	-	No	-	Yes	-	No
Flair [1]	Yes	Yes	No	No	Yes	Yes

3 Application methodology

The methodology that will be adopted to create the 10 multilingual UNER annotated corpora is schematized in Figure 1.

Fig. 1. Steps at the application of the UNER framework methodology to a multilingual parallel corpus.

Using the defined UNER hierarchy, we will tag the SETimes English corpus by combining automatic annotation methods. Following this, we will use crowdsourcing campaigns to evaluate the quality of the annotated data, and to provide insight on possible improvements to our annotation methods. These annotations

will then be propagated to the other languages in the SETimes corpora, and the generated data will be used to train new NER models that will be evaluated in terms of precision, recall, and F-measure.

The SETimes parallel corpus (CC-BY-SA license) is based on the contents published on the SETimes.com news portal, which published "news and views from Southeast Europe" and it covers ten languages: Bulgarian, Bosnian, Greek, English, Croatian, Macedonian, Romanian, Albanian, Serbian and Turkish. In this way we can quickly build UNER systems for many languages and check the universality of the proposed scheme. The English corpus is composed of 4,248,417 words and 205,910 sentences[4].

The process of annotating the English SETimes corpus consists of a hybrid computation workflow.

We employed state-of-the-art NERC tools to pre-annotate the corpus. We use spaCy [9] (3 models each with 18 tags) and Flair [1] (1 model with 4 classes and 2 models 18 classes) to tag the corpus. This step was performed by sorting all the previously mentioned models according to their reported recall, establishing a priority order. We then iterated through each model according to its priority and annotated the corpora. For each iteration, we checked all the tokens that were left untagged by the previous models and used the current one to complete the annotations. The aim of this step is to maximise the overall number of annotated entities, thus, increasing the overall recall. Using this method we have identified 631,068 entity occurrences inside the English SETimes corpus which correspond to 89,600 different character strings.

Then, using SPARQWrapper, we will correct the tags by retrieving information from DBpedia [3], Yago [21] and Wikidata [24]. For that, we will define precise one-to-one correspondences between the relevant classes of these databases and UNER nodes.

The quality of the annotation will be evaluated through crowdsourcing tasks and, lastly, we will propagate the corrected annotations from the English corpus to those in other languages. This final process will be done by employing existing label propagation algorithms and models, such as graph propagation methods [22] and machine-translation models [10].

Once the annotations are propagated from the English corpus to the other 9 parallel SETimes datasets, we aim to conduct an extrinsic evaluation of the generated data by using it to train and test NERC deep learning models (for example Stanford NLP NERC module based on Conditional Random Fields).

For each language, a deep analysis of the entity distribution inside the annotated corpus will be conducted to determine possible biases that may influence evaluation, also taking into consideration missing examples for some categories.

[4] The corrected version of SETimes corpus where diacritics and encoding system has been corrected and is available from nlp.ffzg.hr. We will consider this version in our study.

4 Conclusions and Future Directions

We have presented the UNER framework, a universal multilingual hierarchy building on Sekine's NER hierarchy [19], to be used as a universal framework for NERC annotations. We have also designed a hybrid computation workflow for the creation, correction and possible evaluation of an annotated parallel multilingual corpora using UNER as an annotation framework and the SETimes corpus as the basis. The realization of this workflow and the delivery of the annotated corpora are intended as future work. Similarly, we intend to implement automatic methods for the extraction of events, according to the ACE 2005 event annotation guidelines [6].

We also intend to apply the UNER framework to new languages and evaluate its value as a training corpus for cross-lingual and multilingual named entity and event extraction. Finally, we aim to employ the resulting trained models using UNER corpora in combination with multilingual knowledge graphs, such as Wikidata [24], for the creation of NERC sub-models for multilingual and multicultural environments.

5 Acknowledgements

The work presented in this paper has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement no. 812997 and under the name CLEOPATRA (Cross-lingual Event-centric Open Analytics Research Academy).

References

1. Akbik, A., Blythe, D., Vollgraf, R.: Contextual String Embeddings for Sequence Labeling. In: COLING 2018, 27th International Conference on Computational Linguistics. pp. 1638–1649 (2018)
2. de Araujo, P.H.L., de Campos, T.E., de Oliveira, R.R., Stauffer, M., Couto, S., Bermejo, P.: Lener-br: A dataset for named entity recognition in brazilian legal text. In: International Conference on Computational Processing of the Portuguese Language. pp. 313–323. Springer (2018)
3. Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., Ives, Z.: Dbpedia: A nucleus for a web of open data. In: The semantic web, pp. 722–735. Springer (2007)
4. Chen, Y., Skiena, S.: Building sentiment lexicons for all major languages. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Short Papers). pp. 383–389 (2014)
5. Chinchor, N., Robinson, P.: Appendix E: MUC-7 Named Entity Task Definition (version 3.5). In: Seventh Message Understanding Conference (MUC-7): Proceedings of a Conference Held in Fairfax, Virginia, April 29 - May 1, 1998 (1998), https://www.aclweb.org/anthology/M98-1028
6. Consortium, L.D.: ACE Event Descriptions (v5.4.3). https://www.ldc.upenn.edu/sites/www.ldc.upenn.edu/files/english-events-guidelines-v5.4.3.pdf (07 2005), (Accessed on 03/04/2020)

7. Freitas, C., Carvalho, P., Gonçalo Oliveira, H., Mota, C., Santos, D.: Second HAREM: advancing the state of the art of named entity recognition in Portuguese. In: quot; In Nicoletta Calzolari; Khalid Choukri; Bente Maegaard; Joseph Mariani; Jan Odijk; Stelios Piperidis; Mike Rosner; Daniel Tapias (ed) Proceedings of the International Conference on Language Resources and Evaluation (LREC 2010)(Valletta 17-23 May de 2010) European Language Resources Association. European Language Resources Association (2010)
8. Hardwick, S.: Finnish Tagtools. http://urn.fi/urn:nbn:fi:lb-201811141 (2018)
9. Honnibal, M., Montani, I.: spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing (2017), to appear
10. Jain, A., Paranjape, B., Lipton, Z.C.: Entity Projection via Machine-Translation for Cross-Lingual NER. arXiv preprint arXiv:1909.05356 (2019)
11. Kokkinakis, D., Niemi, J., Hardwick, S., Lindén, K., Borin, L.: HFST-SweNER—A New NER Resource for Swedish. In: LREC. pp. 2537–2543 (2014)
12. Larionov, D., Shelmanov, A., Chistova, E., Smirnov, I.: Semantic role labeling with pretrained language models for known and unknown predicates. Proceedings of Recent Advances of Natural Language Processing pp. 620–630 (2019)
13. Manning, C.D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S.J., McClosky, D.: The Stanford CoreNLP Natural Language Processing Toolkit. In: Association for Computational Linguistics (ACL) System Demonstrations. pp. 55–60 (2014), http://www.aclweb.org/anthology/P/P14/P14-5010
14. Marcińczuk, M.K., Gawor, M., Ogrodniczuk, M., et al.: Recognition of named entities for Polish–comparison of deep learning and conditional random fields approaches. In: Proceedings of the PolEval 2018 Workshop. pp. 77–92 (2018)
15. Orasmaa, S., Petmanson, T., Tkachenko, A., Laur, S., Kaalep, H.J.: EstNLTK - NLP Toolkit for Estonian. In: Calzolari, N., Choukri, K., Declerck, T., Grobelnik, M., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., Piperidis, S. (eds.) Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016). European Language Resources Association (ELRA), Paris, France (may 2016)
16. Östling, R.: Stagger: An open-source part of speech tagger for Swedish. Northern European Journal of Language Technology (NEJLT) **3**, 1–18 (2013)
17. Przepiórkowski, A., Bańko, M., Górski, R.L., Lewandowska-Tomaszczyk, B., Łaziński, M., Pęzik, P.: National corpus of polish. In: Proceedings of the 5th language & technology conference: Human language technologies as a challenge for computer science and linguistics. pp. 259–263 (2011)
18. Sang, E.F., De Meulder, F.: Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. arXiv preprint cs/0306050 (2003)
19. Sekine, S.: The Definition of Sekine's Extended Named Entities. https://nlp.cs.nyu.edu/ene/version7_1_0Beng.html (07 2007), (Accessed on 28/02/2020)
20. Ševčíková, M., Žabokrtskỳ, Z., Krůza, O.: Named entities in Czech: annotating data and developing NE tagger. In: International Conference on Text, Speech and Dialogue. pp. 188–195. Springer (2007)
21. Suchanek, F.M., Kasneci, G., Weikum, G.: Yago: a core of semantic knowledge. In: Proceedings of the 16th international conference on World Wide Web. pp. 697–706 (2007)
22. Tamura, A., Watanabe, T., Sumita, E.: Bilingual lexicon extraction from comparable corpora using label propagation. In: Proceedings of the 2012 Joint Conference

on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. pp. 24–36. Association for Computational Linguistics (2012)
23. Tyers, F.M., Alperen, M.S.: South-east european times: A parallel corpus of balkan languages. In: Proceedings of the LREC Workshop on Exploitation of Multilingual Resources and Tools for Central and (South-) Eastern European Languages. pp. 49–53 (2010)
24. Vrandečić, D., Krötzsch, M.: Wikidata: a free collaborative knowledgebase. Communications of the ACM **57**(10), 78–85 (2014)