Named Entity Recognition from Chernobyl Documentaries

Daniil Tikhomirov[a], Nikita Nikitinsky[b] and Ilya Makarov[a]

[a]National Research University, Higher School of Economics, Moscow, Russia
[b]National University of Science and Technology MISIS, Moscow, Russia

Abstract

The paper describes a system that extracts facts and opinions from documentary texts to create a domain ontology of a controversial topic for Chernobyl disaster. The pipeline of the system is based on RNN-based NER module, which was tested on an annotated text corpora.

Keywords

information extraction, NER, opinion mining, domain ontology, Chernobyl disaster

1. Introduction

In this paper, we propose a project of system that extracts named entities and events from documentary corpus dealing with a controversial topic of Chernobyl disaster. Below we will describe its architecture and initial results achieved by its first implementation.

The task of Named Entity Recognition (NER) is a crucial component in development of any NLP system that requires a certain level of general or domain-specific knowledge, such as systems for question answering. The present task is no exception: the step of retrieving relevant propositions about certain events pertaining to the Chernobyl disaster and the individuals' or objects' involvement in it should be prefaced with an extraction of such objects. This step will allow to both separate the relevant passages of our corpus from irrelevant ones, and to classify the extracted facts and opinions by their subjects.

The domain that the NER methods will be applied to in our work is fairly limited. The number of actors and events connected to the Chernobyl disaster is rather small and predictable, when compared to general domain NER tasks, or certain domain-specific, yet more broad, tasks, such as concept detection in medical texts, like the one described in (Uzuner et al., 2011)[1]. However, the domain is not as limited, and corpus size is not as small as to allow this task to be done manually; rather, it calls for extracting more fine-grained knowledge from the present texts under human supervision and reinforcement.

MACSPro'20: Modeling and Analysis of Complex Systems and Processes, October 22–24, 2020, Venice, Italy
EMAIL: notextinhere@gmail.com (D. Tikhomirov); torselllo@yandex.ru (N. Nikitinsky); iamakarov@hse.ru, corresponding author (I. Makarov)
URL: http://hse.ru/en/staff/iamakarov (I. Makarov)
ORCID: 0000-0002-3308-8825 (I. Makarov)

(Yadav and Bethard, 2018)[2] offers a comprehensive survey of developments in NER systems and reviews different approaches to this problem, from early rule- and dictionary-based approaches to the more recent methods that rely on feature engineering and supervised learning, as well as state-of-the-art neural network systems. Besides describing examples of every major method for NER, the survey compares their effectiveness. This allows to make an informed decision of which method would suit the present task. The gazetteer-based approaches described there were discarded for obvious reasons: while it is possible to annotate a portion of the corpus, one would also like to infer other possible entities that were not found in the annotated part.

Of the machine learning methods, the most consistently good results were achieved by feature-inferring neural network models, namely by Bi-LSTM models, both word- and character-level [2]. It should be noted that feature-engineered models models, such as (Agerri and Rigan, 2017) [3], achieved similarly good results, but they have a strong disadvantage compared to the feature-inferring models: the feature engineering for a new domain requires a lot of time and resources, and constitutes a separate topic in NLP tasks all on its own. As such, when choosing among the ready-made decisions, the RNN models look much more favourable, offering the same good quality while being easy to adapt to suit one's needs.

The suitability of neural network methods is supported by the abundance of libraries that implement this strategy. One of the more popular and available options is the customizable NER model included into SpaCy (spacy.io), a multi-purpose NLP module for Python. The model used there, however, differs from the best-performing models described in (Yadav and Bethard, 2018) [2] in that they use a CNN approach rather than RNN. The SpaCy model calculates word representations using both subword features (prefix, suffix, general shape of the word) and "Bloom" embeddings (which is a way to assign hash IDs to vectors to reduce dimensionality and speed up the model while using several successive hashing functions to avoid hash collisions; the technique was described in (Serra and Karatzoglou, 2017) [4]). These embeddings are then passed to the trigram CNN with residual connections, where they are transformed in accrodance with their context. Finally, the prediction layer for the model is composed of a standard multi-layer perceptron. The architecture of this model is heavily inspired by (Lample et al., 2016) [5], which describes a transition-based model for chunking and labeling a sequence of inputs using a stack data structure, which allows for a "summary embedding" of several previous words. Among other things, this allows for an easy representations of multi-word named entities, as they are included in the stack together. The good results $F = 86.4$, the ease of implementation and customization makes SpaCy a viable instrument for the named entity recognition in our system.

Another popular library that we will inspect is DeepPavlov (https://deeppavlov.ai/). DeepPavlov employs the more "standard" RNN method for the named entity recognition task. It has also been trained on the OntoNotes 5.0 dataset [6], and shows similarly good results as the model used in SpaCy. The quality of these models on our dataset will be compared in Section 4.2.

2. Dataset

Our data consists of a corpus of English-language documentary works concerned with the Chernobyl disaster and, more generally, with effects of hazards connected to such catastrophes on the environment and human health. The corpus includes printed and internet articles and books. These materials come in different ebook formats: PDF, .mobi, .epub and .djvu, some of them containing a separate text layer, and some without one. In the case of ebooks and PDF files with OCR, the text layer is simply extracted using specialized Python libraries; in the latter case, such a layer is first created using image-to-text methods, in particular the Python OCR library PyTesseract.

The present task does not call for usual preprocessing techniques, such as removal of punctuation or stopwords: relation extraction and argumentation detection methods demand full preservation of all elements of written text that would allow to distinguish between sentences that contain opinions and/or argumentation and those that do not, such as commas, colons, conjunctions, etc. Therefore, the corpus preprocessing was primarily concerned with noise that comes from book formatting, that is done to suit the human reader, but does not translate well into any kind of automatic parsing. The sources of this noise include:

- line breaks that separate lines on PDF pages rather than paragraphs;
- tables, figures, images and their captions, that are not parsed effectively and often break the sentence in the middle;
- page numbers;
- repetitions of the book or chapter title;
- footnotes and references.

As it is impossible to find a one-size-fits-all solution to such problems, because each source is characterized by its own particular style of formatting, a separate set of preprocessing rules was created for each source work. These include removal of line breaks that do not come after an end of a sentence, figures, tables, running titles and author names, as well as page numbers. Apart from that, if the work is constructed as an anthology that covers various topics (e.g. various anthropogenic disasters or types of pollution), only a relevant part of such work was manually extracted and included in the corpus. All in all, our corpus contains 462843 tokens (without punctuation) split over 23090 sentences. The bibliographical information about books that were included in the corpus can be found in Appendix A.

3. Experiment Settings

The NER module was used in order to construct the domain ontology of the Chernobyl Disaster - objects and entities that are likely to be a subject of controversy and that played a somehow significant role in the events of the Chernobyl disaster. The entities and objects are included into the ontology in two ways: first, all named entities extracted by an NER model,

and second, common nouns that designate objects and events that cannot be considered named entities, but are, nevertheless, relevant to the topic at hand. These objects include, but not limited to:

- occupations of people that were involved in the catastrophe ("operators", "liquidators", "workers");
- different parts of the Chernobyl facility or the reactor ("graphite rods", "turbines");
- various health hazards ("radiation");
- consequences of exposure to such hazards ("cancer", "sarkoma");
- physical phenomena associated with the catastrophe and the reactor operation ("run-away", "void coefficient");
- umbrella terms for different causes of the disaster ("safety violations", "design faults")

4. NER Models Comparison

As stated above, two popular NER libraries were considered: a hybrid Bi-LSTM-CRF model which comes as a part of a DeepPavlov AI library, and a CNN model adopted by the SpaCy library. Both of these libraries are pre-trained on the OntoNotes 5.0 dataset, using the tagset described in (Weischedel et al., 2013) [6], and both demonstrate state-of-the-art results for the NER task with $F = 86.4$. Both of them can be adapted to different domains, easy to train and simple to integrate into any application. In order to choose between to approaches to the NER task, we have tested both libraries on a manually annotated text - a Wikipedia article on the Chernobyl disaster. The total size of the test corpus is 9700 tokens split over 317 sentences, with 561 entities annotated with OntoNotes labels that used in both models. We have excluded from the annotation labels usually associated with numeric strings, namely tags "ORDINAL", "CARDINAL", "QUANTITY", "PERCENT" "TIME", "DATE" and "MONEY": relations and events associated with those entities fall outside the scope of the present research, as it is oriented towards the extraction of facts and actors' involvement in them, rather than reconstructing of the temporal sequence of events. We have also excluded certain tags that were considered irrelevant for the domain we apply them to: "WORK_OF_ART", "PRODUCT" and "LANGUAGE". The list of OntoNotes tags used during manual annotation of the Wikipedia article, together with the native description and our domain-specific interpretation is provided in Table 1.

The use of certain tags differs slightly from the meaning originally intended by the creators of the tagset. The "EVENT" tag was applied to the phrase "Chernobyl disaster" and the word "Chernobyl" when used in such a sense; the "FAC" tag was used to designate the Chernobyl Nuclear Power Plant and different parts of the plant, e.g. names of reactors, such as "Unit Four"; finally, the LAW tag was applied to the documents created by various commissions following the catastrophe, such as the INSAG-7 report, which is included into the corpus as one of the sources. Such an, arguably frivolous, use of these tags was tried as an experiment to see whether the extraction of such objects and events can be delegated to the NER module, or should be found and listed manually.

Table 1
OntoNotes named entity tags and relevant examples from the Wikipedia article

Tag	OntoNotes 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
LOC	Non-GPE locations, mountain ranges, bodies of water
EVENT	Named hurricanes, battles, wars, sports event s, etc.
LAW	Named documents made into laws

	Precision	Recall	F1
EVENT	0.29	0.06	0.09
FAC	0.88	0.20	0.33
GPE	0.95	0.95	0.95
LAW	1.00	0.08	0.14
LOC	0.95	0.79	0.86
NORP	0.96	0.94	0.95
ORG	0.63	0.81	0.71
PERSON	0.92	0.92	0.92
Average	0.82	0.59	0.67

Table 2
Evaluation metrics for Deep Pavlov RNN model

	Precision	Recall	F1
EVENT	0.38	0.08	0.14
FAC	1.00	0.11	0.20
GPE	0.92	0.89	0.90
LAW	0.00	0.00	0.00
LOC	0.95	0.79	0.86
NORP	1.00	1.00	1.00
ORG	0.48	0.87	0.62
PERSON	0.83	0.73	0.77
Average	0.70	0.56	0.62

Table 3
Evaluation metrics for SpaCy CNN model

For evaluating the results, we have calculated the usual quality metrics: Recall, Precision and F1-score. The label was considered a true positive only in cases of exact entity matches.

Tables 2 and 3 indicate that on average, both models show high precision and low recall, which can be interpreted as a tendency to have many false positives - mistaking the label or identifying as an entity a string that is not. The only entity label for which this pattern does

not hold is the "ORG" label, for which the evaluation metrics high recall with low precision. It seems that the criteria for identifying "ORG" entities are fairly broad in both models.

It can also be seen both models encounter considerable problems with detecting the entities of types "EVENT", "FAC" and "LAW". The most frequent source of mistakes was the polysemy of the word "Chernobyl", which can designate both the city of Chernobyl (assigned a GPE label), the Chernobyl disaster (EVENT) and Chernobyl Nuclear Power Plant (FAC). This polysemy was taken into account while preparing the test corpus, and, as the results show, was not handled well by either model. Other types of entities, which were used conventionally, have shown rather good results. It should be noted, however, that the CNN model employed by SpaCy shows much lower accuracy when identifying entities of the "PERSON" type, with both precision and recall metrics significantly lower than demonstrated by DeepPavlov's RNN model. This difference in accuracy for this particular label we consider to be crucial. There is a finite and rather small number of unique entities of the LAW, EVENT and FAC types that are mentioned in the corpus, and these entities can be specified by hand, as there was a limited number of facilities involved in Chernobyl catastrophe. In addition, most events that are of interest to our research are, in any case, mostly represented by common nouns. The same, however, cannot be said about human actors that took part in the events: the same person can be called by different variations of the same name, and the list of people who are mentioned in the corpus can hardly be exhausted as the significant part of this corpus consists of personal stories and analyses of individual involvement. Based on the results of NER models evaluation, it was decided to use the DeepPavlov RNN model for detection of locations, organizations and people names, while relying on the manually constructed list of interesting objects for the detection of documents, facilities and events.

5. Conclusion

In this paper we have described a way to extract named entities from documentary corpus on Chernobyl disaster. We have compared NER models: SpaCy, based on CNN, and RNN-based DeepPavlov, with the latter achieving a better quality on annotated Wikipedia article on the Chernobyl Disaster. The system can by no means be called a finished one, and this paper serves as a first attempt or a proof-of-concept for a large-scale project.

Acknowledgments

The research was supported by the Russian Science Foundation grant 19-11-00281.

References

[1] O. Uzuner, B. R. South, S. Shen, S. L. DuVall, 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text, Journal of the American Medical Informatics Association : JAMIA 18 (2011) 552–556. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3168320/. doi:10.1136/amiajnl-2011-000203.

[2] V. Yadav, S. Bethard, A Survey on Recent Advances in Named Entity Recognition from Deep Learning models, arXiv preprint arXiv:1910.11470 (2018) 14.
[3] R. Agerri, G. Rigau, Robust Multilingual Named Entity Recognition with Shallow Semi-supervised Features (Extended Abstract), in: Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, International Joint Conferences on Artificial Intelligence Organization, Melbourne, Australia, 2017, pp. 4965–4969. URL: https://www.ijcai.org/proceedings/2017/703. doi:10.24963/ijcai.2017/703.
[4] J. Serrà, A. Karatzoglou, Getting deep recommenders fit: Bloom embeddings for sparse binary input/output networks, arXiv:1706.03993 [cs] (2017). URL: http://arxiv.org/abs/1706.03993, arXiv: 1706.03993.
[5] G. Lample, M. Ballesteros, S. Subramanian, K. Kawakami, C. Dyer, Neural Architectures for Named Entity Recognition, arXiv:1603.01360 [cs] (2016). URL: http://arxiv.org/abs/1603.01360, arXiv: 1603.01360.
[6] Weischedel, Ralph, et al., OntoNotes Release 5.0 LDC2013t19. Web Download., 2013.