TRAINING NER MODELS: KNOWLEDGE GRAPHS IN THE LOOP

Sotiris Karampatakis, Alexis Dimitriadis, Artem Revenko, Christian Blaschke

{firstname.secondname}@semantic-web.com
https://semantic-web.com

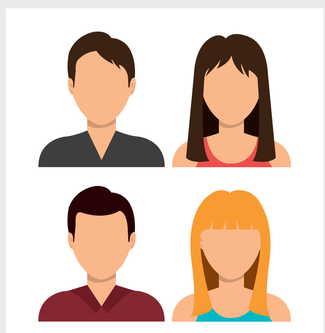
SEMANTIC WEB COMPANY

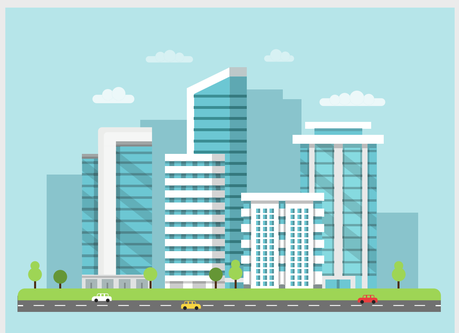
What is NER?

Named Entity Recognition (NER) is a sub-task of information extraction with the objective to identify and classify named entities mentioned in unstructured text. It is commonly approached as a supervised classification problem. This means that annotated training materials are required.

Common NE types are available

Pre-annotated corpora covering common cases such as Person, Organization, Location etc. are easy to obtain.





Are those Production Ready?

What about my case?

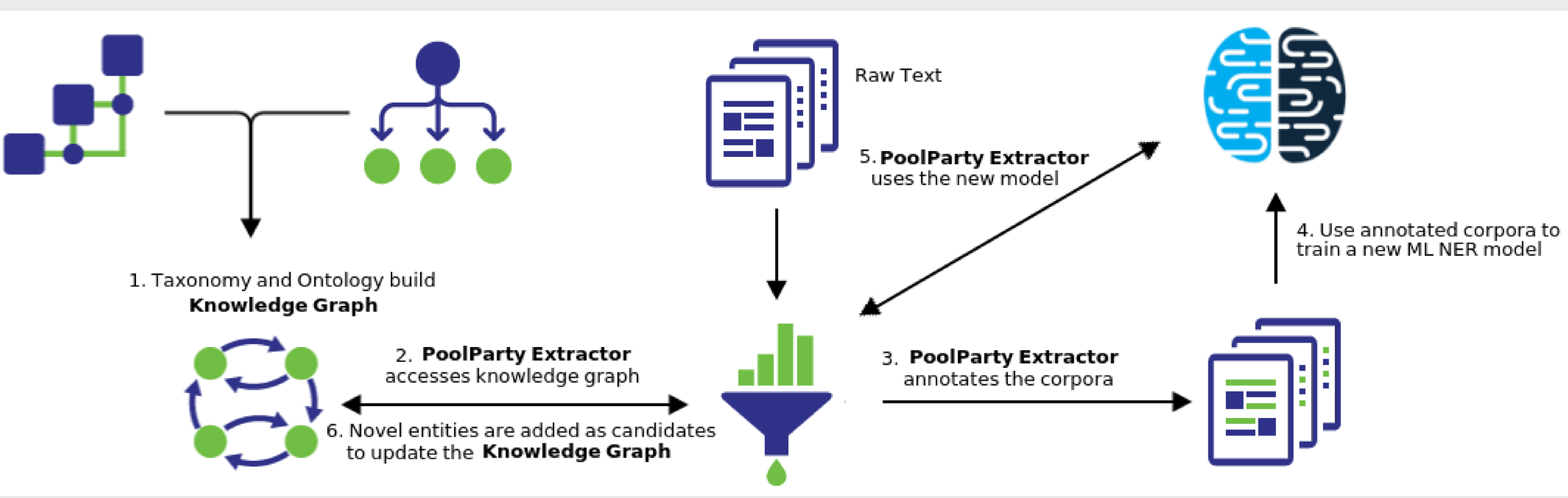
The recognition of more exotic types presents a bootstrapping problem. How can we train a classifier without the time and resource costs associated with manually annotating and curating a sizeable training dataset?

TODO:
- Case Law
- Product
- Disease

Our approach

We aim at producing annotated training data semi-automatically, distantly supervised using a Knowledge Graph. As the pre-requisite we require an initial vocabulary for a domain and raw text of the same domain of interest.

Workflow



Evaluation method

To set a baseline for our evaluation we used the **CoNLL-2003** shared task corpus and the **NCBI-disease** corpus.

- Use the human annotated training corpus to train models.
- Use the evaluation corpus for each dataset to evaluate the models in terms of Precision (PR), Recall (RE) and F_1 score.
- For each of the NE types, create a taxonomy based on the labels of the NE found on the training corpus.
- Re-annotate the raw training corpora using the PoolParty Extractor API, configured to use the corresponding Concept Scheme
- Finally, use the re-annotated corpora to train NER models and evaluate the new models using the corresponding human annotated evaluation corpus.

Results

Dataset	Vocabulary	Entity Type	Annotation Method						ΔF_1
			Human			Automatically			
			PR	RE	F_1	PR	RE	F_1	
CoNLL-2003	Extracted	Person	96.2	86.2	90.9	90.7	72.1	80.3	-10.6
CoNLL-2003	Extracted	Location	94.9	89.1	91.9	81.2	78.3	79.8	-12.2
CoNLL-2003	Extracted	Organization	94.2	65.4	77.2	55.1	70.2	61.7	-15.5
NCBI-disease	Extracted	Disease	82.7	62.1	70.9	75.6	67.1	71.1	0.2
NCBI-disease	MeSH-2019	Disease	82.7	62.1	70.9	55.5	27.7	36.9	-34.0

Evaluation results of OpenNLP NER on human annotated test corpora. Annotation method refers to the training corpora in each case. ΔF_1 is the difference in F_1 scores between automatic and human annotations. Vocabulary identifies how the controlled vocabulary for automatic annotations was created: either already provided human annotations were collected and used for automatic re-annotation or Disease branch of MeSH-2019.

Observations

- Models trained on automatically annotated corpus can achieve comparable results to models trained on human annotated corpus;
- The process allowed us to identify common pitfalls in the automated annotation task.

Acknowledgment

This work has been partially funded by the project LYNX which has received funding from the EU's Horizon 2020 research and innovation programme under grant agreement no. 780602, see http://www.lynx-project.eu