Jehangir et al (2023): A survey on Named Entity Recognition – datasets, tools, and methodologies

by Adam Kellich

Contents

- Datasets
- Tools
- Evaluation
- Approaches
  - unsupervised
  - rule-based
  - supervised
    deep learning
    transfer learning
- Challenges, future directions

Datasets

- Overview of popular ones, propose joint benchmark
- CoNLL (2002)
- WNUT (2017)
- OntoNotes (2013)
- NCBI (2014)
- BioCreative V (2016)
- Genia (2003)
- Wikipedia (many, 2015-ish)

Table 1
Datasets and their corresponding domains.

S.No.	Dataset	Type	Link	Scale (Datasize)			
				Train set (MB)	[Test set] (MB)	Dev. set (MB)	Total (MB)
1	CoNLL-2002	Multilingual	https://www.clips.uantwerpen.be/conll2002/ner/	–	–	–	14.8
2	CoNLL-2003	Multilingual	https://www.clips.uantwerpen.be/conll2003/ner/	–	–	–	16.7
3	WNUT-2017	Social-Media	http://noisy-text.github.io/2022/	–	–	–	1.74
4	OntoNotes	Multilingual	https://catalog.ldc.upenn.edu/LDC2013T19	–	–	–	947.4
5	NCBI Disease	Biomedical	https://www.ncbi.nlm.nih.gov/CBBresearch/Dogan/DISEASE/	0.938	0.171	0.162	–
6	BioCreative	Biomedical	https://biocreative.bioinformatics.udel.edu/tasks/biocreative-v/track-3-cdr/	0.2461	0.2650	0.2840	–
7	Genia Corpus	Biomedical	http://www.geniaproject.org/home	1.75	0.491	0.366	–
8	Wiki Gold	Wikipedia	https://figshare.com/articles/dataset/Learning_multilingual_named_entity_recognition_from_Wikipedia/5462500	–	–	–	102
9	Wiki Coref	Wikipedia	http://rali.iro.umontreal.ca/rali/?q=en/wikicoref	–	–	–	172
10	Hyena	Wikipedia	https://www.mpi-inf.mpg.de/departments/databases-and-information-systems/research/yago-naga/hyena	–	–	–	–
11	BC4CHEMD	Biomedical	https://biocreative.bioinformatics.udel.edu/resources/biocreative-iv/chemdner-corpus/	–	–	–	855
12	JNLPBA	Biomedical	http://www.geniaproject.org/shared-tasks/bionlp-jnlpba-shared-task-2004	9.44	3.11	–	–
13	BC2GM	Biomedical	https://biocreative.bioinformatics.udel.edu/accounts/login/?next=/resources/corpora/biocreative-ii-corpus/	3.55	1.35	–	–

CoNLL

- 2002: dutch, spanish; people, organizations, places, others development, training, testing
- 2003: german, English, same as 2002 + unannotated data (8 files total)

WNUT-2017

- Location, product, corporation, person, creative work (song, movie…), group
- From twitter, reddit, stack exchange, youtube (noisy)
- +++ geospecific places
- Comments over 140 words

OntoNotes

- Large (2.9 mil words) corpus in 3 of the biggest languages: *Chinese, Arabic, English*
- Weblogs, telephone speech (huh?), talk shows, news
- Structural information (predicate-argument syntax), shallow semantics

I ground the rye on number 6 click-LRB-out of 8-RRB-in my Champion Juicer grinder.

```
------------------------------------------------------------------------------------------------------------

Plain sentence:
---------------
    I ground the rye on number 6 click-LRB-out of 8-RRB-in my Champion Juicer grinder.

Treebanked sentence:
--------------------
    I ground the rye on number 6 click -LRB- out of 8 -RRB- in my Champion Juicer grinder .

Tree:
-----
    (TOP (S (NP-SBJ (PRP I))
            (VP (VBD ground)
                (NP (DT the)
                    (NN rye))
                (PP-MNR (IN on)
                        (NP (NP (NML (NN number)
                                     (CD 6))
                                (NN click))
                            (-LRB- -LRB-)
                            (PP (IN out)
                                (PP (IN of)
                                    (NP (CD 8))))
                            (-RRB- -RRB-)))
                (PP-LOC (IN in)
                        (NP (PRP$ my)
                            (NML (NNP Champion)
                                 (NNP Juicer))
                            (NN grinder))))
            (. .)))

Leaves:
-------
    0   I
    1   ground
          sense: ground-v.6
    2   the
    3   rye
    4   on
    5   number
```

NCBI disease corpus

- 790 disease categories, 88 % contain OMIM identity

BioCreative V chemical disease relation

- Chemical articles on diseases,1400 of them chosen from an existing dataset attributed to CTD-Pfizer partnership + 100 new items

Genia

- Uni Tokyo (98-2012)
- Biomedics, keywords: blood cells, transcription factors, human -> 1999 papers, annotated (language and semantic layers)

Wikipedia

- WikiGold – manual annotation of 39k words
- WikiCoref – corpus with focus on anaphoric relationships, similar to OntoNotes format, coreference types
- HYENA – 2012, million entities, 50k randomly chosen wiki pages, 92 % are one of 4 kinds, 10k random for testing

Tools

- SpaCy (2017), high level NLP, production, pretrained/custom models, NER module (residual CNN)
- NLTK (2009)
- Apache openNLP (2015) + ML + NLP tasks
- Tensorflow, Pytorch (2016, 2019)

Evaluation

- Relaxed precision x Also accurate spans (strict)
- [Lionel Andrés] [Messi] x [Lionel Andrés Messi] ->
  100% vs 0% x 100% 100%
- Precision, Recall, F1
- “Only when the projected label for the entire predicted word matches
- the label’s precise wording does the exact match metrics deem a
- prediction to be accurate (Segura-Bedmar et al., 2013). If a portion
- of the predicted entity is accurately recognized, Relaxed F1 deems the
- prediction to be correct. Strict F1 demands that a prediction’s character
- offsets match perfectly with the input annotation’s.”

Approach 1: rule-based

- Designed by specialists, always domain specific
- Regex technique (*ičitý, *ečný, *ane, *amine)
- String similarity (suffixes) + genetic programming
- “In work (Eftimov et al., 2017),the NER method is proposed for extracting dietary information based on evidence which is the first of its kind. The first step is finding and identifying the entities mentioned, and the next entails selecting and obtaining the entities.”
- Rule based -> (semi)supervised validation

Approach 2: unsupervised

- Social network NER – noisy, short: labeling tweets with sequential CRF labeler -> clustering similar content -> finetuned CRF labeling

  I’ve been to Lady Gaga concert -> clustered (takes similar tweets into account) > I’ve been to Lady Gaga concert

- Biomedical NER – leveraging corpus statistics, shallow syntax, terminology
- Domain shift, external sources dependency -> Peng et. al (2021): **adversarial** (1 v 1) training with entity-aware attention
- *Problematic evaluation: lack labels + new patterns vs. imposing structure*

Approach 3: supervised

- HMMs – probab of words/transitions (2004 Zhang, cascaded, abbrev. NER, SOTAATT)
- SVMs – best hyperplane, good at high dim data
- CRFs: classifier normally predicts 1 label for 1 example, but here we consider the neighborhood
  discriminative undirected probabilistic graphical model
  NLP: linear chain – prediction depends on the immediate neighbors

Approach 3: deep learning

- Initial deep learning for NER (2008, Collobert) – first time doing NLP tasks without hand-extracted features (CNN)
- Now: BiLSTM, CNN, hybrid of those two, combination with other: SVMs…
- Text representations of “Hello everyone”:

BOW	[1, 0, 0, 0, 1]
Count Vectorizer	[1, 0, 0, 0, 1]
TF-IDF	[0.5, 0, 0, 0, 0.25]
Word2Vec-skip gram, cbow	[1, 2, 3, 4, 5]

Approach 3: deep learning: CNNs

- Input s= w1, w2, w3…
- Filters used on windows of words of size *m*,
  feature c_i = f(W . E_(i: i+*m))* -> creates high-level combinatorial feature embeddings -> to be combined
- Can be word-level, character-level
- E.g. (Cho for NER, 2020) CNN + bidirLSTM (word/character) chained and input into MLP
  the work also uses an attention technique to the bidirectional-long short-term memory-CRF model
- Gated CNNs, Dilated CNN, RNN

Approach 3: deep learning: CNNs

- (2023, Chang) 3D CNN for document-level is needed. BiLSTM is incapable of working with two sentences at once -> incomplete information
- They researched adding residual structure (layer-by-layer) to optimize the BiLSTM blocks (addresses the deterioration issues when increasing layers)

Approach 3: deep learning: CNNs

- OOV problem -> make educated guesses (breaking down) (Naet 2019)
- Char vectors –(CNN)> Compositional morphemes vectors -> concated

- Zhou (2020), ImExtr from drug package images. Two layers (correcting, CNN) output: determine the sentence

Approach 3: deep learning: RNNs

- RNN – W matrix, sequentially encodes input, vanishing gradient problem. “France, the country of good food… <million words>… “
- LSTMs
- Bi-LSTMs (text is passed in both directions, earliest 2015)

$$i_t = Sigmoid(W_i n_t + U_i h_{(t-1)} + b_i) \quad (7)$$

$$f_t = Sigmoid(W_f n_t + U_f h_{(t-1)} + b_f) \quad (8)$$

$$O_t = Sigmoid(W_o n_t + U_o h_{(t-1)} + b_o) \quad (9)$$

$$\tilde{c}_t = \tanh(W_c n_t + U_c h_{(t-1)} + b_c) \quad (10)$$

where W and U are weights and b is the bias. The hidden state h_t and current state c_t are calculated as

$$c_t = f_t \odot c_{t-1} + i_t \odot \tilde{c}_t \quad (11)$$

$$h_t = O_t \odot \tanh(c_t) \quad (12)$$

Approach 3: deep learning: RNNs

- (2019, Dic-Att-BiLSTM-CRF). String matching with dictionary of diseases, document-level attention -> bi-LSTM with CRF to do the NER. SOTA on NCBI (F1 88 %) and BioCreativeV (F1 88 %)

Challenges

- Data annotation, language ambiguity, noise (preprocessing)
- Complex biomedical texts (expertise, not clearly written, abbreviations)
  “‘Exon-intron structure of the human neuronal nicotinic acetylcholine receptor alpha 4 subunit (CHRNA4)”
- User-generated text (BFU)
- Multilingual NER (only frequency based techniques)
- Domain adaptation (orange vs orange)
- Entity linking (Barcelona vs Barcelona), coreference resolution