MultiCoNER II: **Multi**lingual **Co**mplex **N**amed **E**ntity **R**ecognition

Asal Shavandi, Chahyon Ku, London Lowmanstone, Josh Spitzer-Resnick

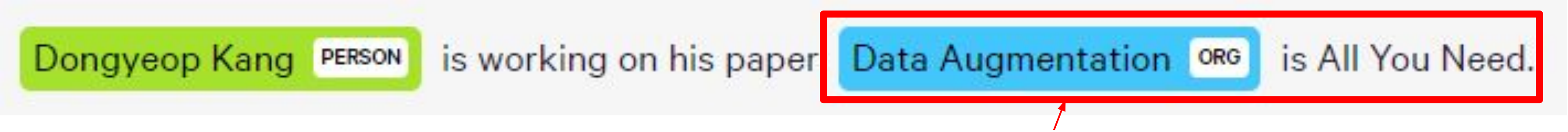
Oct 13, 2022

UNIVERSITY OF MINNESOTA
Driven to Discover℠

We are Polyglots!

Named Entity Recognition (NER)

Named entity recognition is a critical part of **information extraction** in which every word in a sentence is classified to named entity types such as names (people, organization, location) and numbers (time, date, and money).



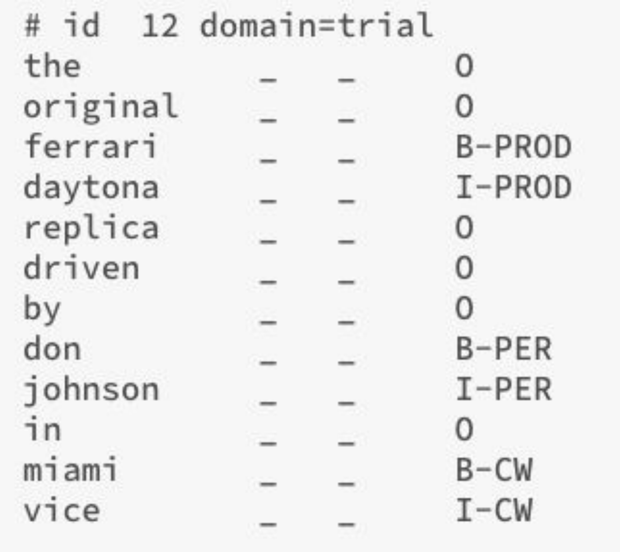
Dongyeop Kang PERSON is working on his paper Data Augmentation ORG is All You Need.

This needs work!

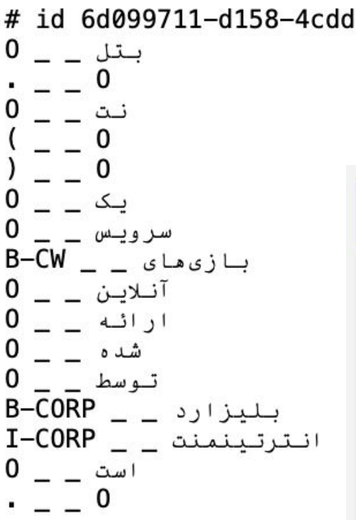
Among named entities, **complex named entities** are the more syntactically complex named entities–often names of **creative works**–that existing systems have a hard time recognizing.

MultiCoNER -> MultiCoNER II

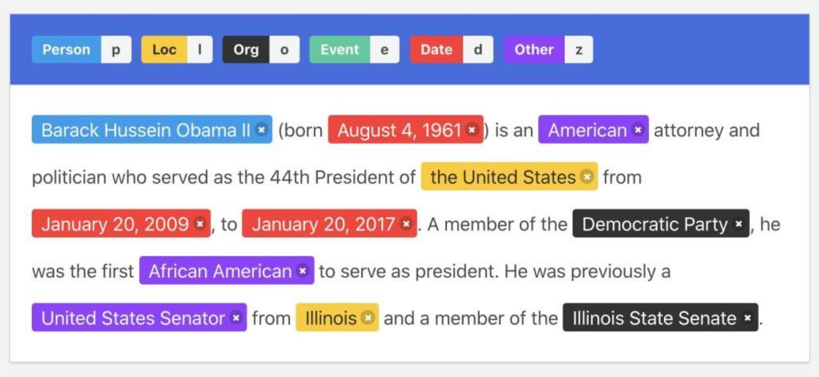
Even though **knowledge-retrieval-based systems** achieved great results last year, they are **sensitive to noisy** and **out-of-domain** entities.



```
# id  12 domain=trial
the           _   _    O
original      _   _    O
ferrari       _   _    B-PROD
daytona       _   _    I-PROD
replica       _   _    O
driven        _   _    O
by            _   _    O
don           _   _    B-PER
johnson       _   _    I-PER
in            _   _    O
miami         _   _    B-CW
vice          _   _    I-CW
```



```
# id 6d099711-d158-4cdd-9cc4-f38c0fb22433
O _ _ بتل
. _ _ O
O _ _ نت
( _ _ O
) _ _ O
O _ _ یک
O _ _ سرویس
B-CW _ _ بازی‌های
O _ _ آنلاین
O _ _ ارائه
O _ _ شده
O _ _ توسط
B-CORP _ _ بلیزارد
I-CORP _ _ انترتینمنت
O _ _ است
. _ _ O
```



Barack Hussein Obama II (born August 4, 1961) is an American attorney and politician who served as the 44th President of the United States from January 20, 2009, to January 20, 2017. A member of the Democratic Party, he was the first African American to serve as president. He was previously a United States Senator from Illinois and a member of the Illinois State Senate.

https://demos.explosion.ai/displacy-ent

Related works

General approaches:

- Fine-tuning pre-trained multilingual models such as mBERT & XLM-RoBERTa
- Ensemble of classifiers, conditional random field (CRF)
- Majority voting

Recent approaches:

They expand general approaches by adding:

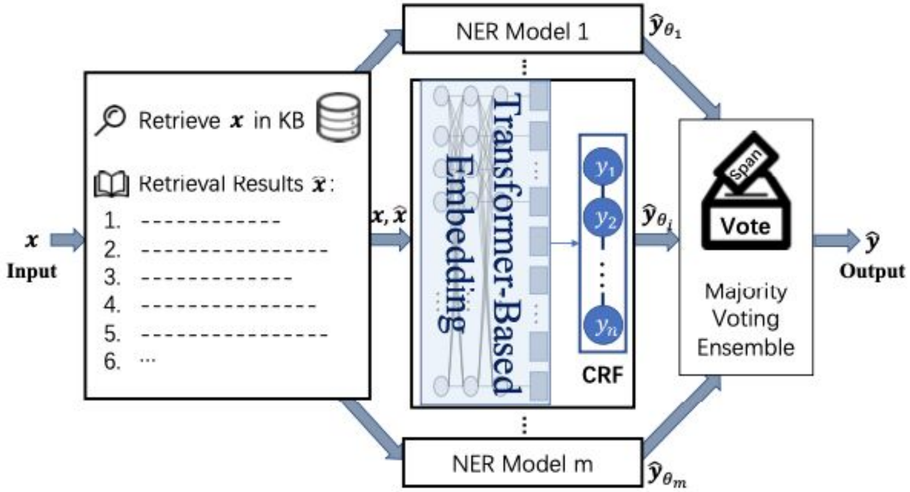
- Data augmentation techniques
- External knowledge retrieval techniques

Related works

SemEval 2022 1st team!

Knowledge-based System for Multilingual NER by Wang et al. 2022

- Knowledge retrieval module
  - Adopted by Wang et al. 2021
- NER module
  - XLM-RoBERTa pre-trained model
  - Linear-chain CRF classification layer
- Ensemble module
  - M models with different random seeds
  - Majority voting

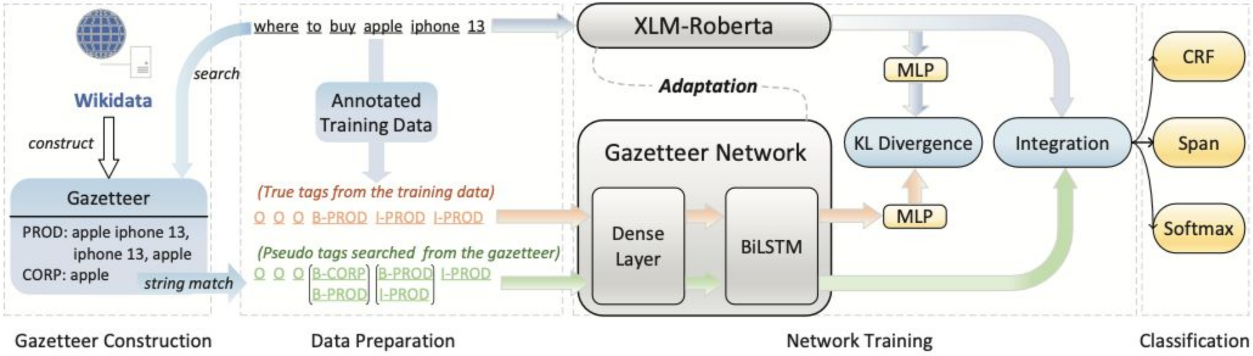


Related works

SemEval 2022 2nd team!

Gazetteer-Adapted Integration Network (GAIN) for Multilingual Complex NER by Chen et al. 2022

- NER module
  - XLM-RoBERTa
  - Two classic sequential labeling: softmax & CRF
  - Segment-based classification: Span
- Gazetteer module
  - Extract Wikidata entities
  - Dense layer & BiLSTM
- GAIN module
  - Divergence Loss



Baseline: XLM-RoBERTa + CRF

- **Used in all top performing models in the previous competition**
- Pre-trained on 2.5TB of filtered CommonCrawl data containing 100 languages (source: Hugging Face)
- Testing on previous competition data until new competition data is released
- Should provide easy-to-compare baseline, even if not SOTA

(Source)

Proposed Idea: two forms of data augmentation

1. Add spelling errors to the sentences
    a. Use NoiseMix and other natural language data augmentation packages
        i. Shown to increase performance on NLP tasks
    b. Swap words (with synonyms) and letters (with other letters) to make models more robust
2. MulDA
    a. “A Multilingual Data Augmentation Framework for Low-Resource Cross-Lingual NER”
    b. Translate sentences into different languages
    c. Specific to NER
        i. Uses placeholder mechanism to retain labels during translation
    d. Not used during previous competition

- **Why** do this?
    - Makes models **more robust** to spelling errors (noise)
    - Helps create **more data** across languages
    - Should **improve all top teams** from last competition

Broader impact

1. Marketing (**Amazon!**)
   a. Global e-commerce becomes easier through fluent translation of entities
   b. Products and reviews
   c. NER is part of AWS' Comprehend managed NLU product

2. Academia
   a. Recognition of academic papers
   b. Semantic Scholar: connecting papers and ideas in papers as "entities" recognized.

3. Domain-specific knowledge
   a. Medical, law, political systems become more accessible
   b. Regulation: security questions, disclosures, account authentication

Questions?

(Image created by DALL-E)

References

- https://multiconer.github.io/
  - https://assets.amazon.science/1a/b3/e091bdd94d0f9e5d2963e2dd6943/multiconer-a-large-scale-multilingual-dataset-for-complex-named-entity-recognition.pdf
  - dataset https://multiconer.s3.us-west-2.amazonaws.com/readme.html
- https://multiconer.github.io/multiconer_1/
  - https://aclanthology.org/2022.semeval-1.196.pdf
  - https://competitions.codalab.org/competitions/36044
- https://arxiv.org/pdf/2203.00545.pdf
- https://arxiv.org/pdf/2105.03654.pdf
- https://arxiv.org/pdf/2203.03216.pdf
- https://aclanthology.org/2021.acl-long.453.pdf