What do we Really Know about State of the Art NER?

Ramya Balasubramaniam, Sowmya Vajjala
ramya.balasubramaniam@novisto.com sowmya.vajjala@nrc-cnrc.gc.ca

Overview

- Research in NER focuses on developing new models, with relatively less emphasis on resources and evaluation.
- Performed a broad evaluation of English NER using OntoNotes dataset, that has various text genres and sources,
- Created 6 adversarial test sets by performing small perturbations that replace select entities based on gender/geography while retaining the context.

Our Approach

1. Black box testing of pre-trained NER models with different test sets (standard test set split by source/genre, new adversarial test sets)

2. Training and testing NER architectures on non-standard splits (on random splits, genre-based splits)

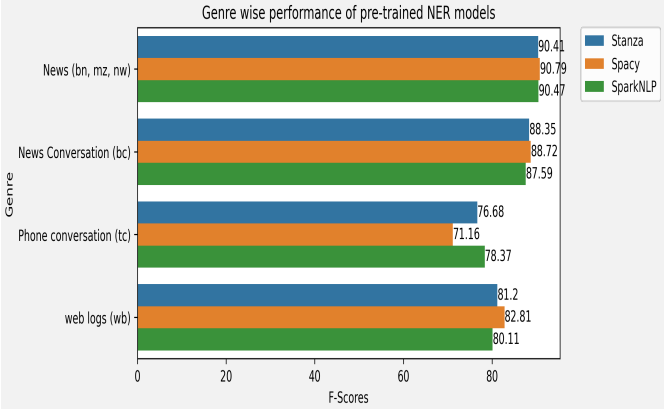
Libraries used for both cases: Spacy, Stanza, SparkNLP

New Adversarial test sets

Perturb	Sentence
None (Original)	Faced with massive demonstrations and Russia's backing of **Kostunica**, he agreed to step down in October.
Perturb_1	Faced with massive demonstrations and Russia's backing of **Dodo**, he agreed to step down in October.
Perturb_2	Faced with massive demonstrations and Russia's backing of **Kevin**, he agreed to step down in October.
Perturb_3	Faced with massive demonstrations and Russia's backing of **Arnav**, he agreed to step down in October.
Perturb_4	Faced with massive demonstrations and Russia's backing of **Pol**, he agreed to step down in October.
Perturb_5	Faced with massive demonstrations and Russia's backing of **Samaira**, he agreed to step down in October.
None (Original)	Previously we had a statistic, especially for the ringroads in Beijing.
Perturb_6	Previously we had a statistic, especially for the ringroads in **Galway**.

Table 1: Examples of sentences in perturbed test sets

Performance variation by Genre

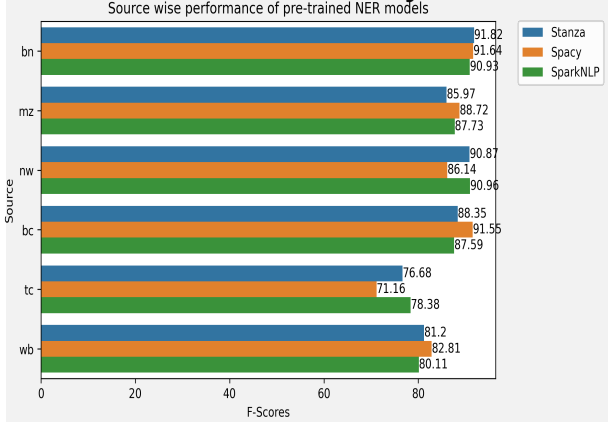


Sensitivity to Adversarial Input

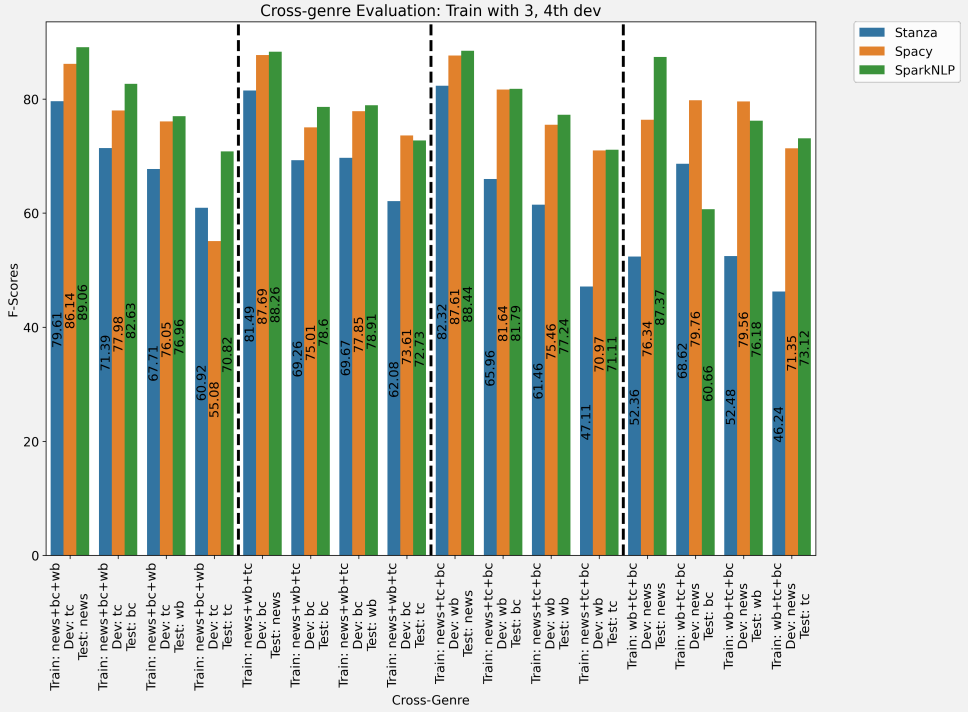
NER Model	All	GPE
Stanza	80.87 (88.71)	68.37 (95.2)
Spacy	82.48 (89.09)	73.47 (95.61)
SparkNLP	80.56 (88.6)	68.84 (95.61)

Table 7: Perturb_6 Performance
(Numbers in brackets indicate the performance on the original test set, without any perturbation)

Performance variation by Source



Training NER on non-standard split



Summary

1. Three SOTA models performed very differently across various entity types with huge variations.
2. For all models, performance on telephone conversation and web blog genres was much lower than the rest.
3. All models were sensitive to small input perturbations, with stark drop for some of the perturbations (4 and 6).
4. Training with 10 randomly generated splits showed over 4% variation in performance across the splits.
5. Trained models performed poorly on genres unseen during training even with multi-genre training data.

Recommendations

- Report the full performance table in the appendix.
- Experiment with random splits and report variation
- Report results on subsets of the test set
- Test the models with adversarial input