Multilingual Named Entity Recognition with SpaCy and Stanza

A comparison of two NLP tools

Index

1. Introduction
2. Data and Resources
3. Method (the Code)
4. Results and Discussion
5. Challenges and Open Issues
6. Summary and Conclusion
7. Sources

1. Introduction

Investigate runtime, accuracy and differences on **NER**

Investigate runtime, accuracy and differences on **NER**

Using Spanish and English data

- **Europarl corpus** (parallel, annotated)
- **OpenSubtitles** (largely parallel, unannotated)

Investigate runtime, accuracy and differences on **NER**

Using Spanish and English data

- **Europarl corpus** (parallel, annotated)
- **OpenSubtitles** (largely parallel, unannotated)

Stanza slower in all instances

SpaCy is slightly less accurate in English, but does an large amount of mistakes on Spanish data

2. Data and Resources

Main source for NER: **“Evaluation Corpus for Named Entity Recognition using Europarl”**

Main source for NER: **“Evaluation Corpus for Named Entity Recognition using Europarl”**

Europarl corpus:

- consists of transcriptions of European Parliament sessions
- plenty of various detailed sentences
- contain different entities
- 4 entity types: LOC, PER, ORG, MISC
- manually annotated by Nora Aranberri

Main source for NER: **“Evaluation Corpus for Named Entity Recognition using Europarl”**

Europarl corpus:

- consists of transcriptions of European Parliament sessions
- plenty of various detailed sentences
- contain different entities
- 4 entity types: LOC, PER, ORG, MISC
- manually annotated by Nora Aranberri

Movie subtitles from “opensubtitles.org”

- using subtitles from “Back to the Future” and “El Hoyo”
- everyday language is used
- largely parallel
- contain many names and places, also fictional ones
- unannotated

Software:

- Python version 3.11.4
- NLP libraries:
  - SpaCy v3.6
  - Stanza 1.4.0.
- SpaCy language models:
  - en_core_web_md in English
  - es_core_news_md in Spanish

SpaCy and **Stanza** over **NLTK**:

Unable to find on which data the language model is trained on, therefore we can’t prove that the data used in our project wasn’t used for the training.

SpaCy and **Stanza** over **NLTK**:

Unable to find on which data the language model is trained on, therefore we can’t prove that the data used in our project wasn’t used for the training.

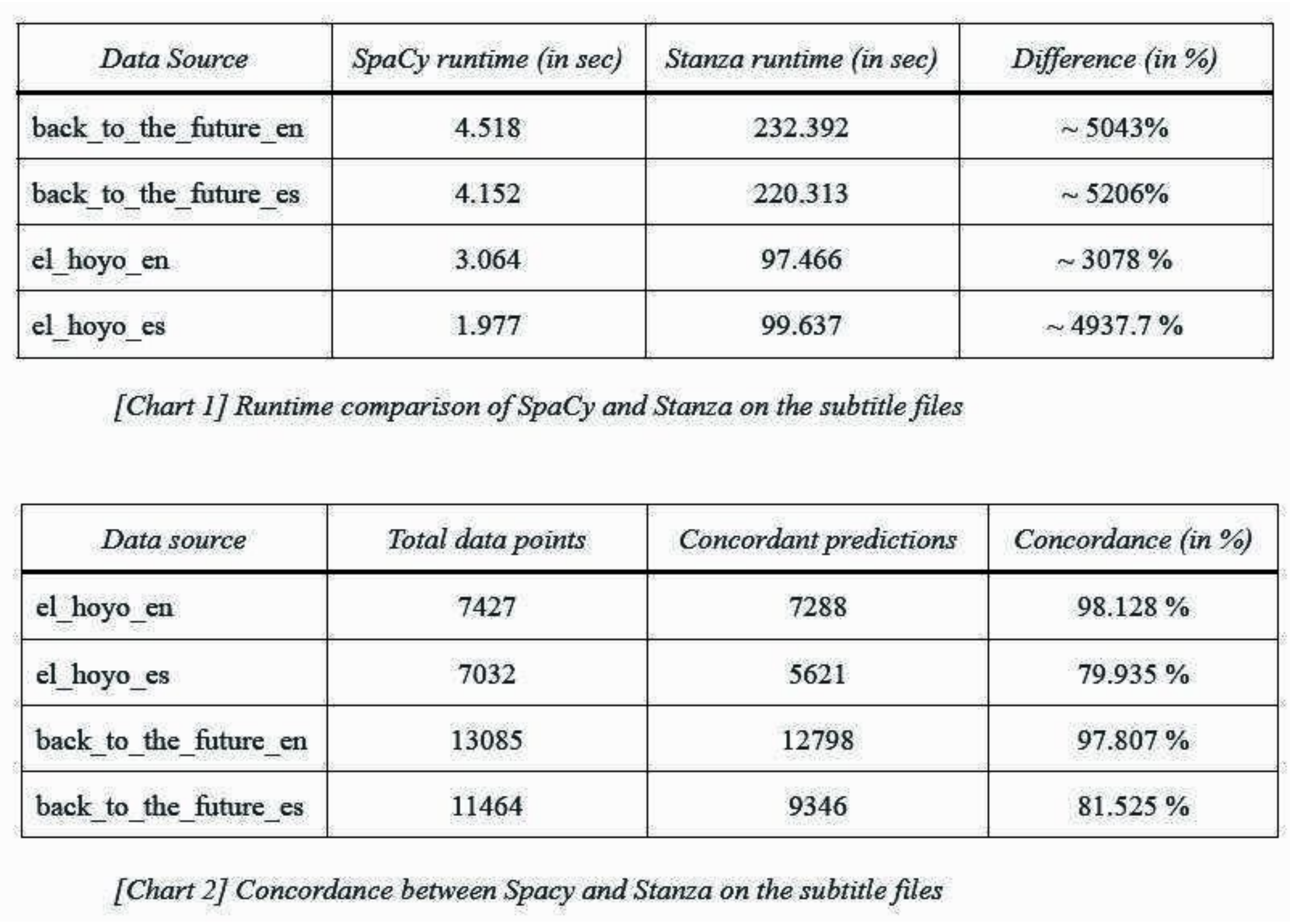
Also used the **re-library** for regular expressions.

3. Method (the Code)

- utils.py:
    - contains auxiliary methods
    - load the data
    - perform Named Entity Recognition
    - evaluate the results

- utils.py:
  - contains auxiliary methods
  - load the data
  - perform Named Entity Recognition
  - evaluate the results
- el_hoyo.py & back_to_the_future.py, europarl_en.py & europarl_es.py:
  - evaluations of the different files executed with the help of methods implemented in utils.py
  - results loaded into text files with respective file names, such as europarl_en_eval.txt

4. Results

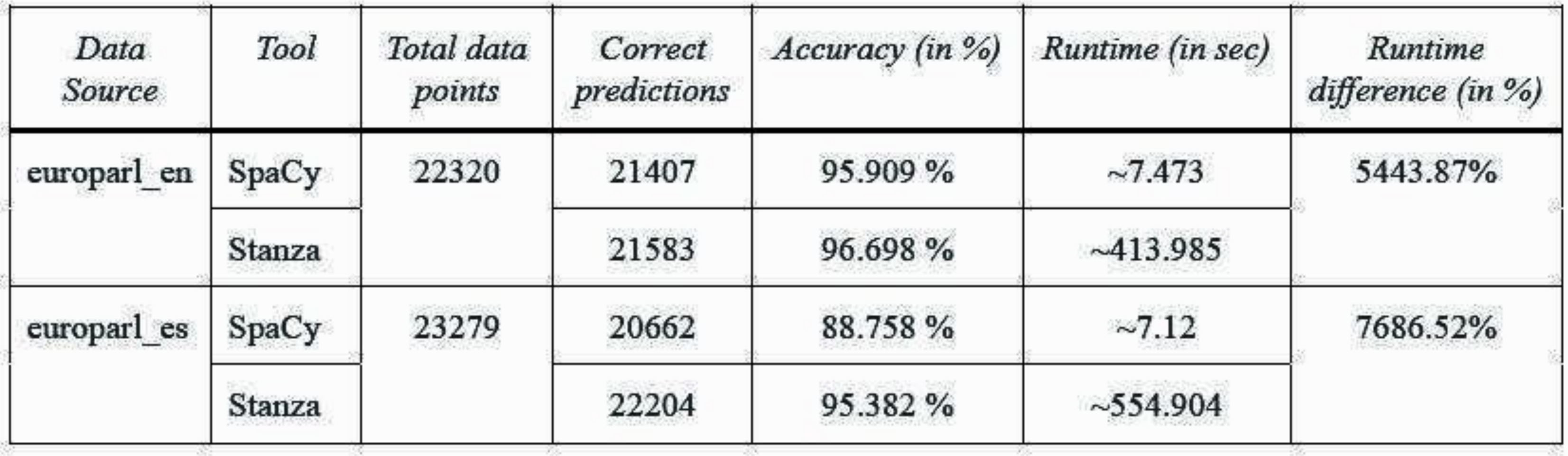


Data Source	*SpaCy runtime (in sec)*	*Stanza runtime (in sec)*	*Difference (in %)*
back_to_the_future_en	4.518	232.392	~ 5043%
back_to_the_future_es	4.152	220.313	~ 5206%
el_hoyo_en	3.064	97.466	~ 3078 %
el_hoyo_es	1.977	99.637	~ 4937.7 %

[Chart 1] Runtime comparison of SpaCy and Stanza on the subtitle files

Data source	*Total data points*	*Concordant predictions*	*Concordance (in %)*
el_hoyo_en	7427	7288	98.128 %
el_hoyo_es	7032	5621	79.935 %
back_to_the_future_en	13085	12798	97.807 %
back_to_the_future_es	11464	9346	81.525 %

[Chart 2] Concordance between Spacy and Stanza on the subtitle files



Data Source	*Tool*	*Total data points*	*Correct predictions*	*Accuracy (in %)*	*Runtime (in sec)*	*Runtime difference (in %)*
europarl_en	SpaCy	22320	21407	95.909 %	~7.473	5443.87%
	Stanza		21583	96.698 %	~413.985	
europarl_es	SpaCy	23279	20662	88.758 %	~7.12	7686.52%
	Stanza		22204	95.382 %	~554.904	

[Chart 3] Accuracy and runtime of SpaCy and Stanza on the Europarl corpus

Europarl - English

- problem with entities specific to context of the European Union, e.g. names of legal texts and organisations
- annotating salutations as PER
- not recognizing several names of persons
- recognizing capitalized words/multi word expressions as ORG

Europarl - Spanish

- Stanza: same errors like in English plus some additional cases
  - Russia labeled as ORG instead of LOC
  - declaring all other country adjectives non-entities (O) instead of MISC

Europarl - Spanish

- Stanza: same errors like in English plus some additional cases
  - Russia labeled as ORG instead of LOC
  - declaring all other country adjectives non-entities (O) instead of MISC
- SpaCy: shares errors and differences with Stanza, seemingly inexplicable mistakes occur as well
  - whole sentences labeled as MISC
  - non-entities recognized as persons

- SpaCy has problems recognizing names of persons: labeling non-entities as PER and some names as O
- both NLP tools struggle recognizing “Doc” (nickname of Doctor Emmett Brown) as PER
- specific names of the fictional characters, events and locations are difficult to recognize for SpaCy
- Stanza isn’t correctly labeling “DeLorean”
- SpaCy: labeling whole sentences in Spanish subtitles as MISC

Source: https://commons.wikimedia.org/wiki/File:Back-to-the-future-logo.svg

DeLorean time machine

Source: https://de.m.wikipedia.org/wiki/Datei:TeamTimeCar.com-BTTF_DeLorean_Time_Machine-OtoGodfrey.com-JMortonPhoto.com-07.jpg

- SpaCy has problems recognizing names of persons: labeling non-entities as PER and some names as O
- both NLP tools struggle recognizing “Doc” (nickname of Doctor Emmett Brown) as PER
- specific names of the fictional characters, events and locations are difficult to recognize for SpaCy
- Stanza isn’t correctly labeling “DeLorean”
- SpaCy: labeling whole sentences in Spanish subtitles as MISC

Source: https://commons.wikimedia.org/wiki/File:Back-to-the-future-logo.svg

- SpaCy
  - labels in Spanish subtitles whole sentences and words that are not entities as MISC, LOC and PER
  - doesn’t recognize names of the main characters Goreng and Trimagashi
- Stanza recognizes sometimes non-entities as PER

- SpaCy labels God and Christ as O, Stanza as PER

Source: https://www.netflix.com/de/title/81128579

5. Challenges and Open Issues

- it was difficult to ensure that both SpaCy and Stanza split the text data in the same way
- need to “post-process” predictions by SpaCy because of different tag set

- it was difficult to ensure that both SpaCy and Stanza split the text data in the same way
- need to “post-process” predictions by SpaCy because of different tag set
- more difficult than anticipated to find parallel annotated corpora for NER
- in future research, the NLTK NLP tool could be included in the comparison

6. Summary and Conclusion

- SpaCy significantly faster than Stanza, in some cases up to 5000% faster

- SpaCy significantly faster than Stanza, in some cases up to 5000% faster
- English: much shorter runtime of SpaCy outweighs the only slightly improved accuracy of Stanza (less than 1% higher)

- SpaCy significantly faster than Stanza, in some cases up to 5000% faster
- English: much shorter runtime of SpaCy outweighs the only slightly improved accuracy of Stanza (less than 1% higher)
- Spanish: SpaCy produces large number of errors compared to Stanza and also many seemingly inexplicable labels

- SpaCy significantly faster than Stanza, in some cases up to 5000% faster
- English: much shorter runtime of SpaCy outweighs the only slightly improved accuracy of Stanza (less than 1% higher)
- Spanish: SpaCy produces large number of errors compared to Stanza and also many seemingly inexplicable labels
- Stanza offers extensive knowledge of fictional names, applies to both English and Spanish

7. Sources

Agerri, R., Chung, Y., Aldabe, I., Aranberri, N., Labaka, G. & Rigau, G. (2018). Building Named Entity Recognition Taggers via Parallel Corpora. In *Proceedings of the 11th Language Resources and Evaluation Conference (LREC 2018)*.

Honnibal, M., Montani, I., Van Landeghem, S. & Boyd, A. (2020). *2020. spaCy: Industrial-strength Natural Language Processing in Python*.

ixa-ehu. (n.d.). *ner-evaluation-corpus-europarl: Manually annotated test set from Europarl for Named Entity Recognition*.

Koehn, P. (2005). *Europarl: A Parallel Corpus for Statistical Machine Translation*.

Qi, P.; Zhang, Yuhao; Zhang, Yuhui; Bolton, J. & Manning, C. D. (2020). Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. In *Association for Computational Linguistics (ACL) System Demonstrations*.

SpaCy · industrial-strength Natural Language Processing in Python. (n.d.). Spacy.Io. Retrieved September 12, 2023, from https://spacy.io/

Stanza – A Python NLP Package for Many Human Languages. (n.d.). Stanza. Retrieved September 12, 2023, from https://stanfordnlp.github.io/stanza/

Subtitle converter. (n.d.). Happy Scribe. Retrieved September 12, 2023, from https://www.happyscribe.com/de/untertitel-tools/untertitel-converter

Subtitles - download movie and TV Series subtitles. (n.d.). Opensubtitles.org. Retrieved September 12, 2023, from https://www.opensubtitles.org/

Thank you for your attention!