A Neural NLP toolkit for Greek

Prokopis Prokopidis and Stelios Piperidis

Institute for Language and Speech Processing/Athena RC

{prokopis, spip}@athenarc.gr

2–4 September 2020

SETN 2020, Athens, Greece

Introduction

- Modules and models for several tasks integrated in popular NLP frameworks for many languages
- Task accuracy can often be improved if one focuses on specific languages
- In this work, we present a set of neural NLP tools
  - developed at the Institute for Language and Speech Processing
  - targeting the accurate processing of Greek texts
- Toolkit based on diverse types of language resources, including
  - web crawled corpora
  - word embeddings
  - large lexical resources
  - corpora manually annotated at different levels of linguistic analysis

Related Work

- Papageorgiou et al. (2000): error-driven transformation-based part-of-speech tagger, 96.28 accuracy for basic POS
- POS tagging, lemmatisation and dependency parsing experiments for Greek conducted on the UD_Greek-GDT treebank[1]
- Models integrated in a number of multilingual NLP toolkits, including UDPipe (Straka and Straková, 2017), Stanza (Qi et al., 2020) and Spacy
    - Spacy: 91.78 accuracy for POS and morphosyntactic features; 85.36 LAS for parsing[2]
    - Stanza: 94.33/96.49 accuracies for tagging/lemmatization; 88.78 LAS for parsing[3]
- In this conference: Papantoniou and Y. Tzitzikas (2020) provide overview of the Greek NLP scene

[1] https://github.com/UniversalDependencies/UD_Greek-GDT

[2] https://spacy.io/models/el

[3] https://stanfordnlp.github.io/stanza/performance.html

Sentence splitting and tokenization

- Text segmentation based on the TreebankWordTokenizer and the sentence boundary detection model for Greek included in the NLTK package (Bird et al., 2009)
- Enhancements of these libraries with the use of language specific resources and heuristics, including
  - A set of 2379 frequent abbreviations and initials
  - Patterns of punctuation combinations that are relatively safe indicators of sentence splits and typically include Greek quotation marks, e.g `;»`, `»;`, `!»`
  - Sets of tokens (typically verbs, but also prepositions and articles) that often undergo deletion of initial or final vowels (e.g. `‘φερες = έφερες`, `φερ' = φέρε`)

Pretrained embeddings

- Embeddings created using
  - fastText library (Bojanowski et al., 2017), which takes into account morphology and represents each word as a bag of character n-grams.
  - A 672M tokens corpus that comprises
    - a collection of articles collected from online archives of Greek newspapers (556M tokens; articles published in the 2003-2020 period)
    - the Greek part of the w2c corpus (Majliš and Žabokrtský, 2012) (116M tokens).
    - Sentence splitting and tokenization module described above
  - 100-dimensional skip-gram (Mikolov et al., 2013) model with character n-grams from 3 to 6 characters, and a vocabulary size of 717.2K tokens.

POS tagging

- Corpus of 23007/427475 sentences/tokens (Papageorgiou et al., 2000), annotated using a tagset of 632 combinations of basic POS tags and features (e.g. `AjBaMaSgNm`: Adjective of basic degree, masculine gender, singular number and nominative case)
- BiLSTM tagger using
  - the StanfordNLP library (Qi et al., 2018)
  - skip-gram vectors described above
  - word embeddings for words with a frequency > 5 in the training set
  - character embeddings for all words
- Accuracy on the test set: 97.75/94.27 for POS/POS and morphosyntactic features
- The drop in accuracy when all features are included is mainly due to errors regarding the case of nominative and accusative noun homographs (e.g. `το/the κόκκινο/red βιβλίο/book`).

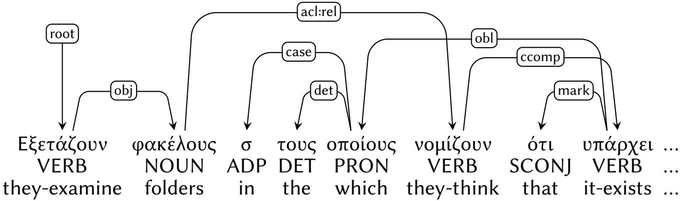
Lemmatization

- Lexicon-based lemmatizer based on a morphological lexicon[4] of 66K lemmas/2M inflected forms
- For words connected with two or more lemmas, POS tags are used for disambiguation
  - `προσχωρήσεις/Verb` -> `προσχωρώ`
  - `προσχωρήσεις/Noun` -> `προσχώρηση`
- For OOV words, we use a BiLSTM lemmatization model trained and evaluated on the same resources used for the POS tagger
- Accuracy of 97.94 for the neural lemmatizer without the use of the morphological lexicon, on all words of the test set

[4] `http://www.ilsp.gr/en/services-products/langresources/item/32-ilektronikomorfologiko`

Greek Dependency Treebank

- Greek Dependency Treebank: a 7400/178648 sentences/tokens resource (Prokopidis et al., 2005)
- Annotation in GDT adheres to the guidelines of the Universal Dependencies project, a collection of 300 similar resources for circa 100 languages (Nivre et al., 2020)
- Dependency annotations in UD focus in linking content words and aim to support downstream language understanding tasks (relation extraction, reading comprehension, etc.)



Εξετάζουν	φακέλους	σ	τους	οποίους	νομίζουν	ότι	υπάρχει	...
VERB	NOUN	ADP	DET	PRON	VERB	SCONJ	VERB	...
they-examine	folders	in	the	which	they-think	that	it-exists	...

- Greek UD treebank (UD_Greek-GDT): 2521/61673 manually annotated sentences/tokens, from texts in the public domain (wikinews articles and European parliament sessions)

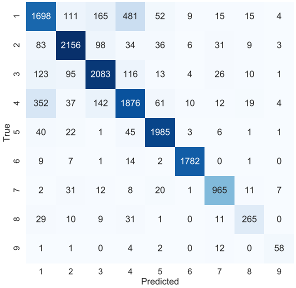
Greek Dependency Parsing

- We train attention-based graph dependency parsers (Dozat and Manning, 2017) using the skip-gram vectors
- We evaluate using Labeled Attachment Score, the percentage of words for which the correct syntactic relation (i.e. subject or object) to the correct head has been identified
- Training the parser on the 178.6K GDT dataset, with manual annotations for POS and lemmas, we observe a 93.42 LAS
- In experiments with automatic annotations for morphology and lemma, LAS is 92.58, i.e. a 0.84 drop in comparison to the manual annotation setting
- On the smaller UD_Greek-GDT, with manual annotation for morphology and lemmas, we obtain a 90.84 LAS

Text Classification

- Corpus: 154.9K newspaper articles with thematic categories extracted from metadata
- We train fastText classifiers and we observe a best 83.55 accuracy on a 15.4K articles test dataset (using bi-grams and an one-vs-all loss function)
- Categories 1 and 4 (”Greek politics” and ”Society”) often confuse the classifier, leading to 833 errors

ID	Category	#Articles	%
1	Greek politics	25000	16.13
2	International news	25000	16.13
3	Economy	25000	16.13
4	Society	25000	16.13
5	Culture	21164	13.66
6	Sports	18935	12.22
7	Science-technology	10528	6.79
8	Environment	3587	2.31
9	Health-medicine	775	0.50
	Total	154989	100.00

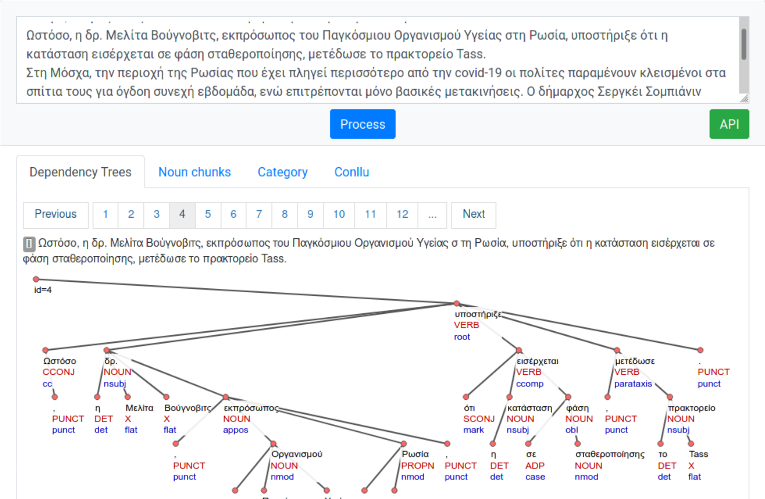


True \ Predicted	1	2	3	4	5	6	7	8	9
1	1698	111	165	481	52	9	15	15	4
2	83	2156	98	34	36	6	31	9	3
3	123	95	2083	116	13	4	26	10	1
4	352	37	142	1876	61	10	12	19	4
5	40	22	1	45	1985	3	6	1	1
6	9	7	1	14	2	1782	0	1	0
7	2	31	12	8	20	1	965	11	7
8	29	10	9	31	1	0	11	265	0
9	1	1	0	4	2	0	12	0	58

Accuracy=0.836

Availability

- Pre-trained embeddings and processed versions of crawled corpora available from `http://nlp.ilsp.gr/setn-2020/`
- NLP modules accessible via a web application and a REST API at `http://nlp.ilsp.gr/nws/`



Ωστόσο, η δρ. Μελίτα Βούγνοβιτς, εκπρόσωπος του Παγκόσμιου Οργανισμού Υγείας στη Ρωσία, υποστήριξε ότι η κατάσταση εισέρχεται σε φάση σταθεροποίησης, μετέδωσε το πρακτορείο Tass.
Στη Μόσχα, την περιοχή της Ρωσίας που έχει πληγεί περισσότερο από την covid-19 οι πολίτες παραμένουν κλεισμένοι στα σπίτια τους για όγδοη συνεχή εβδομάδα, ενώ επιτρέπονται μόνο βασικές μετακινήσεις. Ο δήμαρχος Σεργκέι Σομπιάνιν

Ωστόσο, η δρ. Μελίτα Βούγνοβιτς, εκπρόσωπος του Παγκόσμιου Οργανισμού Υγείας σ τη Ρωσία, υποστήριξε ότι η κατάσταση εισέρχεται σε φάση σταθεροποίησης, μετέδωσε το πρακτορείο Tass.

Conclusions and future work

- We presented a neural NLP toolkit for the Greek language
  - Tagger/lemmatizer accuracies: 97.75/94.27/97.94 for POS/POS+features/lemma, respectively
  - Dependency parser LAS: 92.58 on automatic POS and lemmas
- Plans for
  - Deployment of modules in clarin:el (Piperidis et al., 2017) and the European Language Grid (Rehm et al., 2020)
  - Use of contextual word representations
  - Integration of other NLP modules in the toolkit

Thank you! Questions?