A Neural NLP toolkit for Greek

Prokopis Prokopidis
prokopis@athenarc.gr
Institute for Language and Speech Processing
Athena Research and Innovation Center
Athens, Greece

Stelios Piperidis
spip@athenarc.gr
Institute for Language and Speech Processing
Athena Research and Innovation Center
Athens, Greece

ABSTRACT

We present a neural NLP toolkit for Greek, currently integrating modules for POS tagging, lemmatization, dependency parsing and text classification. The toolkit is based on language resources including web crawled corpora, word embeddings, large lexica, and corpora manually annotated at different levels of linguistic analysis. We show how we integrate these diverse resources in the development of the toolkit, achieving high accuracies on held-out evaluation data for each module.

CCS CONCEPTS

• **Computing methodologies → Natural language processing**.

KEYWORDS

NLP, Greek language, neural networks

ACM Reference Format:
Prokopis Prokopidis and Stelios Piperidis. 2020. A Neural NLP toolkit for Greek. In *11th Hellenic Conference on Artificial Intelligence (SETN 2020), September 2–4, 2020, Athens, Greece.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3411408.3411430

1 INTRODUCTION

Neural network methods have revolutionized the field of natural language processing [6]. Although modules and models for many NLP tasks are constantly integrated in popular frameworks for many languages, task accuracy can often be improved if one focuses on specific languages only. In this work, we present a set of tools that have been developed at the Institute for Language and Speech Processing (ILSP) and target the accurate and efficient processing of Greek texts. Our toolkit is based on diverse types of language resources, including web crawled corpora, word embeddings, large lexical resources, and corpora manually annotated at different levels of linguistic analysis. After briefly listing some indicative work related to Greek NLP in section 2, in the rest of this paper we provide details and evaluation regarding specific modules of the toolkit.

SETN 2020, September 2–4, 2020, Athens, Greece

ACM ISBN 978-1-4503-8878-8/20/09…$15.00
https://doi.org/10.1145/3411408.3411430

2 RELATED WORK

Among NLP research for Greek, an error-driven transformation-based part-of-speech tagger with an accuracy of 96.28 for basic POS was described in [13]. Support vector machines were used for named entity recognition in [9], where a 93.34 F1-score was reported for the PER class. An accuracy of 92.54 (on tags with POS and morphosyntactic features) was reported for a tagger integrated in a Greek NLP toolkit based on both machine learning algorithms and rule-based approaches [16]. For syntactic analysis, experiments with a graph-based dependency parser led to a 90.29 Labeled Attachment Score (LAS) against the Greek Dependency Treebank [17]. A Greek edition of the BERT [4] contextual language model is available from the AUEB NLP group [3]. Models for POS tagging, lemmatisation and dependency parsing of Greek have been trained on the UD_Greek-GDT treebank[1] and are integrated in a number of multilingual NLP toolkits, including UDPipe [21], Stanza [19] and Spacy. An accuracy of 91.78 for POS and morphosyntactic features, and a 85.36 LAS are reported for the Greek models in Spacy[2]. Scores of 94.33/96.49/88.78 are reported for the Greek tagging/lemmatization/parsing models in Stanza[3].

3 TOOLKIT MODULES

3.1 Sentence splitting and tokenization

For text segmentation, we use the TreebankWordTokenizer and the pre-trained Punkt [8] sentence boundary detection model for Greek included in the NLTK package [1]. We enhance the functionalities of these libraries with the use of in-house, language specific resources and heuristics. In more detail, we integrate a) a set of 2379 extra abbreviations and initials appearing frequently in online news articles b) generated patterns that match token internal punctuation in upper and lower case tokens (e.g. α.ε.κ.τ.ε) that may not be included in the set of known abbreviations c) patterns of punctuation combinations that are relatively safe indicators of sentence splits and typically include Greek quotation marks, e.g ;», »;, !» d) 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) and should not be processed as tokens appearing at the beginning or the end of quotations and e) patterns that trigger joining sentences wrongly split by the Punkt model (e.g. when sentence-final punctuation is followed by sentence-initial lower case tokens). In the end of the segmentation process, each token is assigned a token type[4] on the basis of patterns matching the token and its context.

[1] https://github.com/UniversalDependencies/UD_Greek-GDT
[2] https://spacy.io/models/el, el_core_news_lg model, v. 2.3.0
[3] https://stanfordnlp.github.io/stanza/performance.html, accessed July 2020
[4] See http://nlp.ilsp.gr/nlp/tagset_examples/tagset_en/tokenizer.html for the set of token types with indicative examples.

3.2 Pre-trained embeddings

Except for text segmentation, most modules of our toolkit make use of pre-trained word embeddings. We create embeddings with the fastText library [2], which takes into account morphology and represents each word as a bag of character n-grams. We train fastText embeddings on 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); and the Greek part of the w2c corpus [10] (116M tokens). We sentence split and tokenize the corpus with the text segmentation module described in 3.1, and then train a 100-dimensional skip-gram[11] model with character n-grams from 3 to 6 characters, and a vocabulary size of 717.2K tokens.

3.3 POS tagging

Following text segmentation, the next module in the toolkit is a neural tagger that assigns part of speech and morphosyntactic features to each token in a sentence. The tagger is trained on a corpus of 23007/427475 sentences/tokens that is a derivative of the resource described in [13]. The corpus is manually annotated for sentence and token segmentation, part of speech and morphosyntactic features, and lemmas. The tagset used for the morphology annotation layer of the corpus contains 632 combinations of basic POS tags and features that capture the rich morphology of the Greek language.[5] As an example, the full tag `AjBaMaSgNm` for a word like εύφορος/`fertile` denotes an adjective of basic degree, masculine gender, singular number and nominative case. The three last features are also used for nouns, articles, pronouns, and passive participles. Verb tags include features for tense and aspect, while articles are distinguished for definiteness.

We split the corpus in 80/10/10% train/dev/test partitions and train a BiLSTM tagger using the StanfordNLP library [18]. The tagger takes into account the skip-gram vectors of §3.2, word embeddings for words with a frequency > 5 in the training set, and character embeddings for all words. The accuracy of the tagger on the test set is 97.75 (POS tags) and 94.27 (tags with POS and all features). The drop in accuracy when all features are included in the evaluation is mainly due to errors regarding the case of nominative and accusative noun homographs. Nouns of neuter gender in particular are difficult to disambiguate, especially due to the fact that articles and adjectives potentially preceding them are also invariant for nominative and accusative case (e.g. το/`the` κόκκινο/`red` βιβλίο/`book`).

3.4 Lemmatization

In the lemmatization module of our toolkit, we use a hybrid approach. A lexicon-based lemmatizer retrieves lemmas from ILSP's Morphological lexicon,[6] which consists of 2M entries (inflected forms) that correspond to 66K lemmas. When a word under examination is connected in the lexicon with two or more lemmas, the lemmatizer uses information from the POS tags for disambiguation. For example, the word προσχωρήσεις will be assigned the lemma προσχωρώ, if tagged as a verb, and the lemma προσχώρηση, if tagged as a noun. In the case of words that do not appear in the lexicon, we use a BiLSTM lemmatization model trained and evaluated on the same resources used for the POS tagger. We observe an accuracy of 97.94 for the neural lemmatizer without the use of the morphological lexicon, on all words of the test set.

3.5 Dependency parsing

Dependency parsing has become a preeminent paradigm in automatic syntactic analysis during the past 15 years. In dependency parsing, analyzers generate tree representations for each input sentence, where each word depends on a head word and is assigned a label depicting its relation to the head word (for an example, cf. Fig. 1). The Universal Dependencies (UD) community effort, a project that started in 2014, has grown into a collection of 300 such language resources representing close to 100 languages [12]. One of the advantages of this effort is that it has led to the development and collection of treebanks that adhere to common and extensible annotation guidelines. Thus, these resources can be used in training and evaluating multilingual dependency parsers for languages belonging to a large number of language families. The UD guidelines aim to ease the annotation process and improve the accuracy of parsers and downstream NLP applications. The Greek UD treebank (UD_Greek-GDT) consists of 2521 manually annotated sentences (61673 tokens) derived from primary texts that are in the public domain, including wikinews articles and European parliament sessions. UD_Greek-GDT is derived from the Greek Dependency Treebank, a 7400/178648 sentences/tokens resource [15]. Among the structures that may pose problems to a parser for a morphologically rich language like Greek is word order. The relatively free word order of the language can be inferred when examining typical head-dependent structures in the treebank. Although determiners and adjectives almost always precede their nominal heads, the situation is different for arguments of verbs. Of the 2776 explicit subjects in UD_Greek-GDT, 32.89% occur to the right of their parent, while the percentage rises to 46.12% for subjects of verbs heading dependent clauses.

We train a neural attention-based parser [5] using the train/dev/test partitions of UD_Greek-GDT version 2.5[7] and the skip-gram vectors of §3.2. In an experiment involving manual annotation for morphology and lemmas, we obtain a Labeled Attachment Score (LAS, the percentage of words for which a relation with the correct label to the correct head has been identified) of 90.84 on the test set of the UD_Greek-GDT. When we train the parser on the 80% of the 178.6K GDT dataset, we obtain a higher 93.42 LAS. After replacing the 100 dimensional, 717.2K vocabulary skip-gram vectors with 300 dimensional, 2M vocabulary cbow vectors,[8] we observe a nearly identical 93.40 LAS. When focusing on "content dependency relations" (e.g. *subject* and *object*) and omitting evaluation on easier, "functional" relations (e.g. *determiner* and *auxiliary*), we observe lower scores of 86.27/90.03 in the UD_Greek-GDT and GDT settings, respectively. In experiments with automatic annotations for morphology and lemma on the GDT datasets, we obtain a LAS

[5] See http://nlp.ilsp.gr/nlp/tagset_examples/tagset_en/ for a full description of the tagset, including all morphosyntactic features and indicative examples.
[6] http://www.ilsp.gr/en/services-products/langresources/item/32-ilektronikomorfologiko
[7] https://lindat.mff.cuni.cz/repository/xmlui/handle/11234/1-3105
[8] Available from https://fasttext.cc/docs/en/crawl-vectors.html and described in [7]

of 92.58, i.e. a 0.84 drop in comparison to the manual annotation setting.

3.6 Text classification

From the newspaper articles corpus used in the creation of word embeddings, we pre-process and select a subset for experiments in text classification. Pre-processing of each article includes metadata extraction, removal of boilerplate (menu text, disclaimers, etc.), text extraction, and paragraph segmentation. We select articles based on the availability of metadata information including publication date and thematic categories, with the latter mapped to a custom newspaper-independent scheme. We create a 154.9K/51M articles/tokens dataset with a distribution of 9 categories in the articles shown in Table 1.

Table 1: Distribution of categories in the articles of the text classification dataset

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

We train fastText classifiers using different combinations of loss functions and word n-gram lengths 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. As shown in the confusion matrix of Figure 2, categories 1 and 4 ("Greek politics" and "Society") often confuse the classifier, leading to 833 errors (5.4% of all testing instances). When evaluating on the best two labels returned by the model with a probability over 0.8, we obtain a 88.1/74.8 precision@2/recall@2, respectively.

4 AVAILABILITY

Pre-trained embeddings and processed versions of crawled corpora that have been acquired from web sites with open content, are available from http://nlp.ilsp.gr/setn-2020/ as open resources for reuse by the community. All the NLP modules described above are accessible via a web application and a REST API at http://nlp.ilsp.gr/nws/. In Figure 1 we give an example of a syntactic tree visualisation generated by the current interface of the web application.

5 CONCLUSIONS

We presented a neural NLP toolkit for the Greek language. In the evaluation of the BiLSTM tagger and lemmatizer of the toolkit, we observed accuracies of 97.75/94.27/97.94 in assigning the correct POS/POS+features/lemma, respectively. The Labeled Attachment Score of the best neural dependency parser in the toolkit is 92.58 on automatic POS and lemmas. In future work, we plan to deploy most of the modules in the clarin:el infrastructure [14] and the European Language Grid [20]. We also aim to investigate how contextual word representations can improve results in specific tasks and how other NLP modules can be integrated in the toolkit.

ACKNOWLEDGMENTS

We acknowledge support of this work by the project "APOLLONIS: Greek Infrastructure for Digital Arts, Humanities and Language Research and Innovation" (MIS 5002738), "Reinforcement of the Research and Innovation Infrastructure" Action, Operational Programme "Competitiveness, Entrepreneurship and Innovation" (NSRF 2014-2020), co-financed by Greece and the EU (European Regional Development Fund), and by the EU Horizon 2020 Programme under GA no. 825627 (European Language Grid).

REFERENCES

[1] Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python* (1st ed.). O'Reilly Media, Inc.

[2] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics* 5 (2017), 135–146. https://doi.org/10.1162/tacl_a_00051

[3] Ilias Chalkidis. 2020. A Greek edition of BERT pre-trained language model. https://github.com/nlpaueb/greek-bert. [Online; accessed May-2020].

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186. https://www.aclweb.org/anthology/N19-1423.pdf

[5] Timothy Dozat and Christopher D. Manning. 2017. Deep Biaffine Attention for Neural Dependency Parsing. In *5th International Conference on Learning Representations ICLR*. https://openreview.net/forum?id=Hk95PK9le

[6] Yoav Goldberg. 2017. *Neural Network Methods for Natural Language Processing*. Synthesis Lectures on Human Language Technologies, Vol. 37. Morgan & Claypool, San Rafael, CA. https://doi.org/10.2200/S00762ED1V01Y201703HLT037

[7] Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. 2018. Learning Word Vectors for 157 Languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. European Language Resources Association (ELRA), Miyazaki, Japan. https://www.aclweb.org/anthology/L18-1550.pdf

[8] Tibor Kiss and Jan Strunk. 2006. Unsupervised Multilingual Sentence Boundary Detection. *Computational Linguistics* 32, 4 (2006), 485–525. https://www.aclweb.org/anthology/J06-4003.pdf

[9] Georgios Lucarelli and Ion Androutsopoulos. 2006. A Greek Named-Entity Recognizer That Uses Support Vector Machines and Active Learning. In *SETN*.

[10] Martin Majliš and Zdeněk Žabokrtský. 2012. Language Richness of the Web. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis (Eds.). European Language Resources Association (ELRA), Istanbul, Turkey.

[11] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. In *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*, Yoshua Bengio and Yann LeCun (Eds.). http://arxiv.org/abs/1301.3781

[12] Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajic, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. 2020. Universal Dependencies v2: An Evergrowing Multilingual Treebank Collection. In *Proceedings of The 12th Language Resources and Evaluation Conference*. European Language Resources Association, Marseille, France, 4034–4043. https://www.aclweb.org/anthology/2020.lrec-1.497.pdf

[13] Harris Papageorgiou, Prokopis Prokopidis, Voula Giouli, and Stelios Piperidis. 2000. A Unified POS Tagging Architecture and its Application to Greek. In *Proceedings of the 2nd Language Resources and Evaluation Conference*. European Language Resources Association, Athens, 1455–1462.

[14] Stelios Piperidis, Penny Labropoulou, and Maria Gavrilidou. 2017. clarin:el: a language resources documentation, sharing and processing infrastructure.

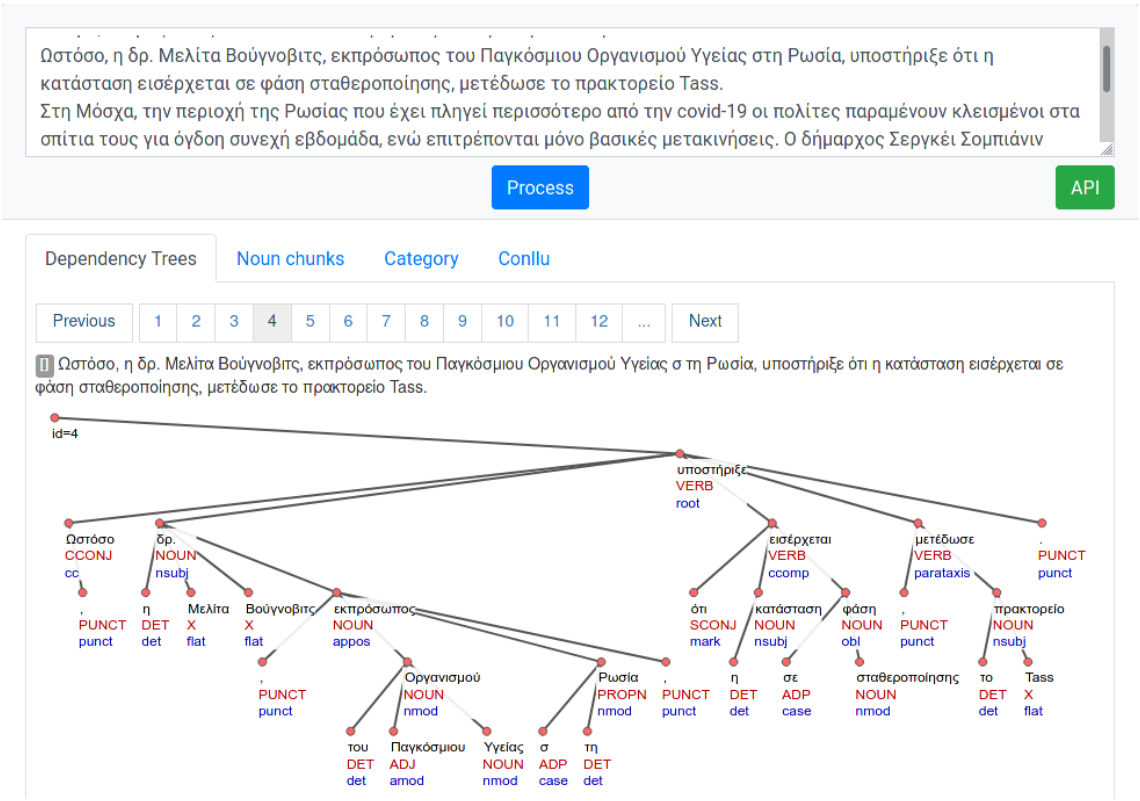

Figure 1: A tree visualization of a sentence analysed by the dependency parser of the toolkit

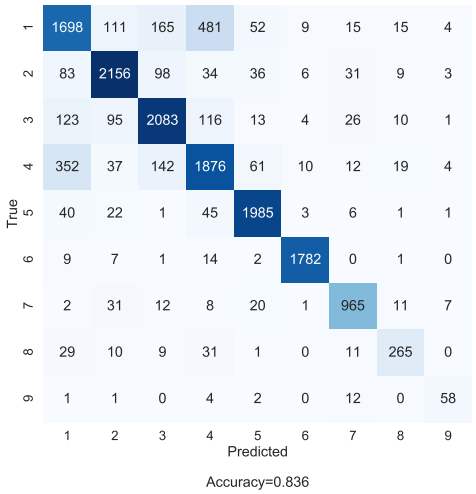

Figure 2: Confusion matrix for the text classifier

In *Proceedings of the 12th International Conference on Greek Linguistics*. Berlin, Germany, 851–869.

[15] Prokopis Prokopidis, Elina Desypri, Maria Koutsombogera, Haris Papageorgiou, and Stelios Piperidis. 2005. Theoretical and practical issues in the construction of a Greek Dependency Treebank. In *Proceedings of the Fourth Workshop on Treebanks and Linguistic Theories*. Barcelona, Spain.

[16] Prokopis Prokopidis, Byron Geograntopoulos, and Haris Papageorgiou. 2011. A suite of NLP tools for Greek. In *Proceedings of the 10th International Conference of Greek Linguistics*. Komotini, Greece. http://nlp.ilsp.gr/nlp/ICGL2011_Prokopidis_etal.pdf

[17] Prokopis Prokopidis and Haris Papageorgiou. 2017. Universal Dependencies for Greek. In *NoDaLiDa Workshop on Universal Dependencies*. http://universaldependencies.org/udw17/pdf/UDW13.pdf

[18] Peng Qi, Timothy Dozat, Yuhao Zhang, and Christopher D. Manning. 2018. Universal Dependency Parsing from Scratch. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*. Association for Computational Linguistics, Brussels, Belgium, 160–170. https://www.aclweb.org/anthology/K18-2016.pdf

[19] Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. *arxiv:2003.07082* (2020). https://arxiv.org/pdf/2003.07082.pdf

[20] Georg Rehm, Maria Berger, Ela Elsholz, Stefanie Hegele, Florian Kintzel, Katrin Marheinecke, Stelios Piperidis, Miltos Deligiannis, Dimitris Galanis, Katerina Gkirtzou, Penny Labropoulou, Kalina Bontcheva, David Jones, Ian Roberts, Jan Hajic, Jana Hamrlová, Lukáš Kačena, Khalid Choukri, Victoria Arranz, Andrejs Vasiljevs, Orians Anvari, Andis Lagzdiņš, Jūlija Meļņika, Gerhard Backfried, Erinç Dikici, Miroslav Janosik, Katja Prinz, Christoph Prinz, Severin Stampler, Dorothea Thomas-Aniola, Jose Manuel Gomez-Perez, Andres Garcia Silva, Christian Berrío, Ulrich Germann, Steve Renals, and Ondrej Klejch. 2020. European Language Grid: An Overview. In *Proceedings of The 12th Language Resources and Evaluation Conference*. European Language Resources Association, Marseille, France, 3359–3373. https://www.aclweb.org/anthology/2020.lrec-1.412.pdf

[21] Milan Straka and Jana Straková. 2017. Tokenizing, POS Tagging, Lemmatizing and Parsing UD 2.0 with UDPipe. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*. Association for Computational Linguistics, Vancouver, Canada, 88–99. http://www.aclweb.org/anthology/K/K17/K17-3009.pdf