Multilingual Text Annotation of Slovenian, Croatian and Serbian with WebLicht

Nikola Ljubešić[2,1]
[1] University of Zagreb
Croatia
nikola.Ljubešić@ffzg.hr

Tomaž Erjavec
[2] Jožef Stefan Institute
Ljubljana, Slovenia
tomaz.erjavec@ijs.si

Darja Fišer[3,2]
[3] University of Ljubljana
Slovenia
darja.fiser@ff.uni-lj.si

Erhard Hinrichs
Tübingen University
Germany
erhard.hinrichs@uni-tuebingen.de

Marie Hinrichs
Tübingen University
Germany
marie.hinrichs@uni-tuebingen.de

Cyprian Laskowski
[2] Jožef Stefan Institute
Ljubljana, Slovenia
cyp@trojina.si

Filip Petkovski
Corner Case Ltd.
Skopje, Macedonia
filip.petkovsky@gmail.com

Wei Qui
Tübingen University
Germany
wei.qiu@uni-tuebingen.de

Abstract

Linguistic annotation of text corpora is a prerequisite for corpus linguistics or any advanced explorations of language. In this paper we first introduce three of the CLARIN.SI suite of open source trainable tools, namely diacritic restoration, word-normalisation and part-of-speech tagging with lemmatisation, trained for three Slavic languages: Slovene, Croatian and Serbian. We then present the trial integration of the tagger/lemmatiser with WebLicht, which has so far offered annotation workflows mainly for German and other Western European languages.

1 Introduction

With the increasing availability of language technologies for various languages, social sciences and humanities (SSH) have started to perceive their usefulness for their own research. Given the lower level of technical competence of most SSH researchers compared to computer scientists, a significant technological gap has to be filled, which would enable SSH scholars to include the developed technologies in their own research. And while this process is under way for English and major Western European languages, it is less advanced for smaller European languages.

This paper presents an effort to make annotation tools for three western South Slavic languages – Slovene, Croatian and Serbian – more widely accessible. We present a subset of the CLARIN.SI suite of open source state-of-the-art machine-learning tools trained for the three languages and the initial attempt to expose one of them as a web service endpoint integrated into the WebLicht corpus annotation workflow design and execution platform.

2 Annotation Tools and Models

The presented tools either improve on existing solutions or are completely novel for some or all of the three languages (Ljubešić et al., 2016a). They are based on the machine learning paradigm, and comprise the learning and execution components. The tools are thus reasonably language independent, and have currently been trained to work with Slovene, Croatian and Serbian. For training these models the best resources available for the task were used, in most cases available under one of the CC licences in the CLARIN.SI repository. All the developed tools are made available under the CLARIN.SI GitHub organisation[1].

2.1 Diacritic Restoration

In computer-mediated communication (CMC), Latin-based scripts users often omit diacritics when writing. Such text is typically easily understandable to humans but very difficult for computational processing because many words become ambiguous or unknown.

For restoring diacritics to words we developed the ReDi tool (Ljubešić et al. 2016b), which can be trained on large corpora of texts to produce high-quality rediacritisation, as the tool takes into account not only individual words but also their context. The tool considerably outperforms charlifter, so far the only open source tool available for this task. On CMC (Twitter) data without diacritics, the tool achieves 99.12% per-token accuracy for Slovene, 99.38% for Croatian and 99.17% for Serbian, removing thereby more than 95% of token-level error.

2.2 Non-Standard Text Normalisation

CMC is often written in non-standard language, where users use phonetic and dialectal spelling. However, annotation tools, such as PoS taggers and lemmatisers, are typically trained on standard language and perform poorly on non-standard texts. A standard morphosyntactic tagger of Slovene applied on various CMC texts, for instance, experiences a token-level accuracy drop from 94.27% to 68.67%, increasing thereby the error rate more than five times (Ljubešić et al, 2017). As developing new text tools for each language variety is very time consuming and expensive, a typical approach is to first normalise the spelling of words to the contemporary standard and only then apply further processing on them.

For normalising words we developed a tool called CSMTiser that uses character-level statistical machine translation (CSMT), the Slovene variant of which has been applied both to CMC and historical texts (Ljubešić et al., 2016c). It is based on the well-known SMT system Moses (Koehn et al., 2007). The tool is, for CMC processing of the three languages, trained on manually normalised collections of tweets split into sequences of characters. Experiments on Slovene showed that for less standard tweets the token-level error reduction obtained when applying CSMT is ~70% while for more standard tweets it is ~50% (Ljubešić et al., 2016c).

2.3 Morphosyntactic Tagging and Lemmatisation

For Slavic languages, morphosyntactic tagging is probably the most important step in text annotation, and is still an active topic of research as such languages with their large tagsets of morphosyntactic descriptions (MSDs) and often limited training data still offer significant room for improvement in tagging accuracy. A similar point holds for lemmatisation, the process of assigning the base form to a word form in running text. On one hand, the rules for predicting the lemma of a word form for Slavic languages are complex and have many exceptions, while on the other hand, the word forms are often ambiguous and their MSD tag is needed to correctly determine the lemma.

We developed a new tagger combined with a lemmatiser called ReLDI-tagger (Ljubešić et al. 2016d), explicitly developed for high-quality processing of Slavic and other highly inflected languages. The tagger uses the CRF sequential classifier with a carefully engineered set of features that results in currently the best accuracy for all three languages, obtaining an error reduction on Slovene of ~25% compared to previous results with the Obeliks tagger (Grčar et al., 2012). The reported tagging accuracy is 94.27% for Slovene, 92.53% for Croatian and 92.33% for Serbian.

[1] https://github.com/clarinsi

3 Integration with WebLicht

WebLicht (Hinrichs et al. 2010) is an execution environment for automatic annotation of text corpora. Linguistic tools such as PoS taggers and parsers are encapsulated as web services, which can be combined by the user into custom processing chains. The resulting annotations can then be visualized in an appropriate way, such as in a table or tree format. WebLicht services can be run from within a web application with a graphical user interface. The web application supports execution of predefined processing chains or the assembly of customized chains by the users themselves. Alternatively, WebLicht as a Service (WaaS) can be used to execute WebLicht chains directly from a UNIX shell, script, or software program.

3.1 Text Corpus Format

WebLicht services are REST-style web services, where the output of one tool in a processing chain is sent as input to the next. In order for the WebLicht services to be interoperable with each other, an XML exchange format called TCF[2] (Text Corpus Format) is used. TCF is fully compatible with the Linguistic Annotation Format (LAF) and Graph-based Format for Linguistic Annotations (GrAF) developed in the ISO/TC37/SC4 technical committee (Ide and Suderman 2007). It should be stressed that TCF is not meant as "yet another standard", but rather as an internal processing format to support efficient data sharing and web service execution. In the first step, we have adapted our tools to support, in addition to its native tabular format, also TCF as its input and output format.

3.2 Exposing the tools on the Web

Next, we made some of our tools, in particular the ReDi diacritic restoration tool, the ReLDI-tagger and the mate-tools parser[3] trained on the respective Universal Dependencies corpora (Agić and Ljubešić, 2015; Dobrovoljc et al. 2017; Samardžić et al. 2017) for the three languages available on the Web[4] (Ljubešić et al., 2016a). The tool can be used, after logging in, via a web form, which enables prospective users to test it before downloading it or using it as a web service. Authentication is done with a username and password, or by having the client IP whitelisted in the system. Furthermore the tool is also made available as a web service. All requests, regardless of the mode of use, are performed through the HTTP protocol. The tool models are loaded in memory when the HTTP server is started which allows for fast processing of client requests. We use Jetty for hosting the parser, and Flask and Gunicorn for hosting the tagging, lemmatization and parsing tools.[5]

3.3 Registering with WebLicht

To make a web service available for use in WebLicht, the metadata describing the service must be created and made available in one of the CLARIN centre repositories, in our case in the CLARIN.SI repository. This requires the creation of a CMDI metadata file for each web service, and associating this metadata file with a new unique entry in the CLARIN centre repository. We have currently registered the developmental tool for Slovenian, Croatian and Serbian as separate entries. In the future, we will rather use HTTP parameters to pass the language to WebLicht in order not to have to create an entry for each tool / language combination.

The CMDI file can be created using Comet (CMD Orchestration Metadata Editing Tool). It is recommended to put all such web services in a dedicated set (in the OAI sense) of the repository, so that managing and harvesting the WebLicht web services is more efficient. Finally, this set must be registered in the CLARIN Centre Registry, in order for WebLicht to be aware of the set and to subsequently harvest it from the CLARIN Centre repository and so expose the web service in the scope of WebLicht. These steps have been taken during the integration of the ReLDI-tagger for Slovenian.

---

2 http://weblicht.sfs.uni-tuebingen.de/weblichtwiki/index.php/The_TCF_Format

3 http://www.ims.uni-stuttgart.de/forschung/ressourcen/werkzeuge/matetools.en.html

4 The entry point is currently http://nl.ijs.si/services but will soon migrate to the new CLARIN.SI server.

5 The documentation for the Web service is available from https://github.com/clarinsi/reldi-lib-doc.

4 Conclusion

This paper presented a subset of the CLARIN.SI tools for linguistic annotation, namely a diacritic restorer, normaliser, tagger and lemmatiser, most of them trained for three South Slavic languages, Slovene, Croatian and Serbian. In the second part of the paper a prototype integration of a subset of these tools into the WebLicht platform was described. Our future plans are focused primarily on exposing the remaining CLARIN.SI tools described in this paper via WebLicht, as well as additional tools that were recently developed, like a named entity recognizer covering both standard and non-standard language, and a morphosyntactic tagger developed exclusively for annotating non-standard language. During the process of adding additional tools into WebLicht, we will draft documentation on that process, helping thereby similar projects in the future.

References

[Agić and Ljubešić 2015] Željko Agić, Nikola Ljubešić (2015). Universal Dependencies for Croatian (that Work for Serbian, too). *Proceedings of the 5th Workshop on Balto-Slavic Natural Language Processing*.

[Dobrovoljc et al. 2017] Kaja Dobrovoljc, Tomaž Erjavec and Simon Krek (2017). The Universal Dependencies Treebank for Slovenian. *Proceedings of the 6th Workshop on Balto-Slavic Natural Language Processing (BSNLP 2017)*, Association for Computational Linguistics (ACL), Valencia, Spain.

[Grčar et al. 2012] Miha Grčar, Simon Krek and Kaja Dobrovoljc (2012). Obeliks: statistični oblikoskladenjski označevalnik in lematizator za slovenski jezik (Obeliks: a statistical morphosyntactic tagger and lemmatiser for Slovene). In *Proceedings of the Eight Conference on Language Technologies*, Ljubljana, Slovenia.

[Hinrichs et al. 2010] Hinrichs, E., M. Hinrichs & T. Zastrow (2010). WebLicht: Web-Based LRT Services for German. In: *Proceedings of the Systems Demonstrations at the 48th Annual Meeting of the Association for Computational Linguistics (ACL-2010)*. Uppsala. pp. 25-29.

[Ide and Suderman 2007] Nancy Ide and Keith Suderman. 2007. Graf: a graph-based format for linguistic annotations. In Proceedings of the Linguistic Annotation Workshop, LAW '07, pages 1–8.

[Koehn 2017] Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. (2007). Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions, ACL '07*, pp. 177–180, Stroudsburg, PA, USA. Association for Computational Linguistics.

[Ljubešić et al. 2016a] Nikola Ljubešić, Tomaž Erjavec, Darja Fišer, Tanja Samardžić, Maja Miličević, Filip Klubička, Filip Petkovski. (2016). Easily Accessible Language Technologies for Slovene, Croatian and Serbian. *Proceedings of the Conference on Language Technologies and Digital Humanities*. pp. 120–124, Ljubljana, Slovenia, 2016, Academic Publishing Division of the Ljubljana Faculty of Arts. Ljubljana, Slovenia, http://www.sdjt.si/jtdh-2016/en/.

[Ljubešić et al. 2016b] Nikola Ljubešić, Tomaž Erjavec, Darja Fišer. (2016). Corpus-Based Diacritic Restoration for South Slavic Languages. In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016), European Language Resources Association (ELRA), Paris, France.

[Ljubešić et al. 2016c] Nikola Ljubešić, Katja Zupan, Darja Fišer, Tomaž Erjavec. (2016). Normalising Slovene data: historical texts vs. user-generated content. In *Proceedings of Konferenz zur Verarbeitung natürlicher Sprache (KONVENS 2016)*, Bochum, Germany.

[Ljubešić et al. 2016d] Nikola Ljubešić, Tomaž Erjavec. (2016). Corpus vs. Lexicon Supervision in Morphosyntactic Tagging: the Case of Slovene. *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, European Language Resources Association (ELRA), Paris, France.

[Ljubešić et al. 2017] Nikola Ljubešić, Tomaž Erjavec, Darja Fišer. (2017) Adapting a State-of-the-Art Tagger for South Slavic Languages to Non-Standard Text. *Proceedings of the 6th Workshop on Balto-Slavic Natural Language Processing (BSNLP 2017)*, Association for Computational Linguistics (ACL), Valencia, Spain.

[Samardžić et al. 2017] Tanja Samardžić, Mirjana Starović, Željko Agić, Nikola Ljubešić (2017). Universal Dependencies for Serbian in Comparison with Croatian and Other Slavic Languages. *Proceedings of the 6th Workshop on Balto-Slavic Natural Language Processing (BSNLP 2017)*, Association for Computational Linguistics (ACL), Valencia, Spain.