ITU Turkish NLP Web Service

Gülşen Eryiğit
Department of Computer Engineering
Istanbul Technical University
Istanbul, 34469, Turkey
gulsen.cebiroglu@itu.edu.tr

Abstract

We present a natural language processing (NLP) platform, namely the "ITU Turkish NLP Web Service" by the natural language processing group of Istanbul Technical University. The platform (available at tools.nlp.itu.edu.tr) operates as a SaaS (Software as a Service) and provides the researchers and the students the state of the art NLP tools in many layers: preprocessing, morphology, syntax and entity recognition. The users may communicate with the platform via three channels: 1. via a user friendly web interface, 2. by file uploads and 3. by using the provided Web APIs within their own codes for constructing higher level applications.

1 Introduction

ITU NLP research group is devoted to produce Turkish NLP tools for more than 10 years. The group offers many NLP courses in graduate level and core NLP research components to different research groups both in NLP field and other disciplines: e.g. linguistics, data mining, web mining and information retrieval. The motivation of the presented platform in this paper comes from the real word problems of sharing the produced NLP resources by different people from varying level of computer background (starting from undergraduates to PhD students or researchers, people from other fields (e.g.linguistics)). These may be categorized under the following main problems:

1. Need to provide assistance for the installation and the usage of different tools, all posing different technological requirements in the users' computers.
2. Difficulty to share the updates and the new modules introduced into the pipeline.
3. Difficulty of using the tools for educational purposes within the classrooms and term projects.
4. licensing issues of the underlying technologies (such as FST and machine learning softwares)

The difficulty in the ease-of-use of Turkish NLP tools and their inconsistencies with each others were also causing the replication of the same effort in different places and preventing the community from working on less-studied higher level areas for the Turkish language. A good example to this may be the efforts for creating Turkish morphological analyzers: some outstanding ones among many others are (Oflazer, 1994; Eryiğit and Adalı, 2004; Akın and Akın, 2007; Sak et al., 2008; Çöltekin, 2010; Şahin et al., 2013))

In this paper, we present our new web service which provides both a whole Turkish NLP pipeline (from raw data to syntax, example given in Figure1 priorly defined in (Eryiğit, 2012)) and its atomic NLP components for stand-alone usage, namely:

- Tokenizer
- Deasciifier
- Vowelizer
- Spelling Corrector
- Normalizer
- isTurkish
- Morphological Analyzer
- Morphological Disambiguator
- Named Entity Recognizer
- Dependency Parser

2 Provided Components

The provided components via our web service may be grouped under 4 layers: preprocessing, morphological processing, multiword expression handling and syntactic processing.

2.1 Preprocessing

The preprocessing layer consists of many sub components specifically developed for unformat-

Proceedings of the Demonstrations at the 14th Conference of the European Chapter of the Association for Computational Linguistics, pages 1–4,
Gothenburg, Sweden, April 26-30 2014. ©2014 Association for Computational Linguistics

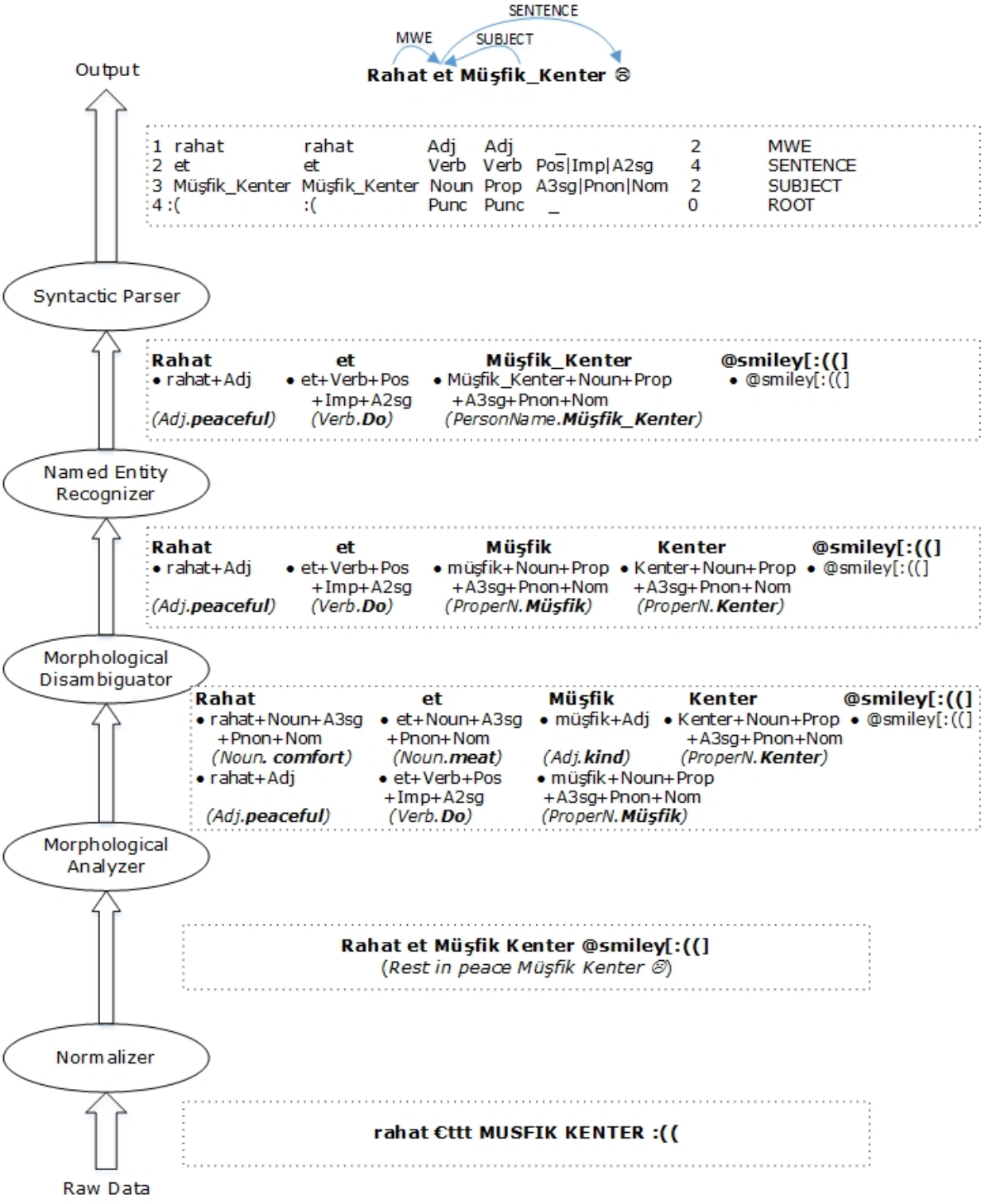

Figure 1: ITU Turkish NLP Pipeline

ted social media data in focus. These are a tokenizer, a diacritic restorer, a vowelizer, a spelling corrector and a normalizer. The diacritic restorer [1] is the component where the ASCII characters are transformed into their proper Turkish forms. The deasciifier (Adalı and Eryiğit, 2014) chooses the most probable candidate within the current context by using conditional random fields (CRFs). The **vocalizer** (Adalı and Eryiğit, 2014) restores the omitted vowels (generally within the social media messages for shortening purpose): e.g. "svyrm" will be converted to "seviyorum" (I love you). The **spelling corrector**[2] is kind of an adaptation of Wang et al.(2011) into agglutinative languages. The **normalizer** (Torunoğlu and Eryiğit, 2014) is constructed of the previous three components and many other modules and provides a state of the art text normalizer for Turkish.

2.2 Morphological Processing

This layer consists of a rule based **morphological analyzer** (Şahin et al., 2013; Şahin, 2014) which uses HFST-Helsinki Finite State Transducer (Lindén et al., 2009) and a hybrid **morphological disambiguator**[3]. This layer also provides the **isTurkish** component which validates a word by using the morphological analyzer.

2.3 Multi Word Expressions

As shown in Eryigit et al. (2011), the detection and unification of the named entities has the highest impact for the syntactic layer. That is why the following Turkish named entity recognizer (Şeker and Eryiğit, 2012) is included within the pipeline and the remaining multiword expressions are detected in the syntactic layer as shown in Figure 1 (dependency label MWE).

2.4 Syntactic Parsing

For the syntactic layer we are providing the state of the art dependency parser for Turkish presented in (Eryiğit et al., 2008; Nivre et al., 2007) which produces the ouputs in Conll format (Buchholz and Marsi, 2006).

3 Conclusion and Future Work

We introduced our ITU Turkish NLP Web Platform which provides us easier administration, automatic updates and patch management, compatibility, easier usage, easier collaboration[4] and global accessibility by being designed as a SaaS. Any body from any discipline with any level of underlying computer background may easily use our web interface either for only analyzing language data or for constructing more complicated NLP systems. The platform already attracted many users from different universities in Turkey and it is now started to get used in many research projects and graduate theses. We believe as being the pioneer serving almost all of the available and top performing NLP tools for Turkish, ITU Turkish NLP Web Service will fed light to new research topics for this language.

For now, the pipeline is constructed by converting the input output formats of each individual tools. But our current effort is to transform the platform into a UIMA(Ferrucci and Lally, 2004) compliant architecture so that it can also integrate with other such platforms universally. We also plan to service the new version of ITU Data Annotation Tool (Eryiğit, 2007) from the same address where the users will also be able to see their data visually (e.g. dependency trees)

Acknowledgments

I want to thank my students without whose it would be impossible to produce the ITU Turkish NLP pipeline: Thomas Joole, Dilara Torunoğlu, Umut Sulubacak and Hasan Kaya. This work is part of a research project supported by TUBITAK 1001(Grant number: 112E276) as an ICT cost action (IC1207) project.

References

Kübra Adalı and Gülşen Eryiğit. 2014. Vowel and diacritic restoration for social media texts. In *5th Workshop on Language Analysis for Social Media (LASM) at EACL*, Gothenburg, Sweden, April. Association for Computational Linguistics.

Ahmet Afsin Akın and Mehmet Dündar Akın. 2007. Zemberek, an open source nlp framework for turkic languages. *Structure*.

Sabine Buchholz and Erwin Marsi. 2006. Conll-X shared task on multilingual dependency parsing. In *Proceedings of the 10th Conference on Computational Natural Language Learning*, pages 149–164, New York, NY. Association for Computational Linguistics.

Çağrı Çöltekin. 2010. A freely available morphological analyzer for Turkish. In *Proceedings of the 7th International conference on Language Resources and Evaluation (LREC2010)*, pages 820–827.

[1] named as "**deasciifier**" since the term is already adopted by the Turkish community

[2] Publication in preparation.

[3] Publication in preparation.

[4] The mailing list notifications are sent to registered users with each new broadcast.

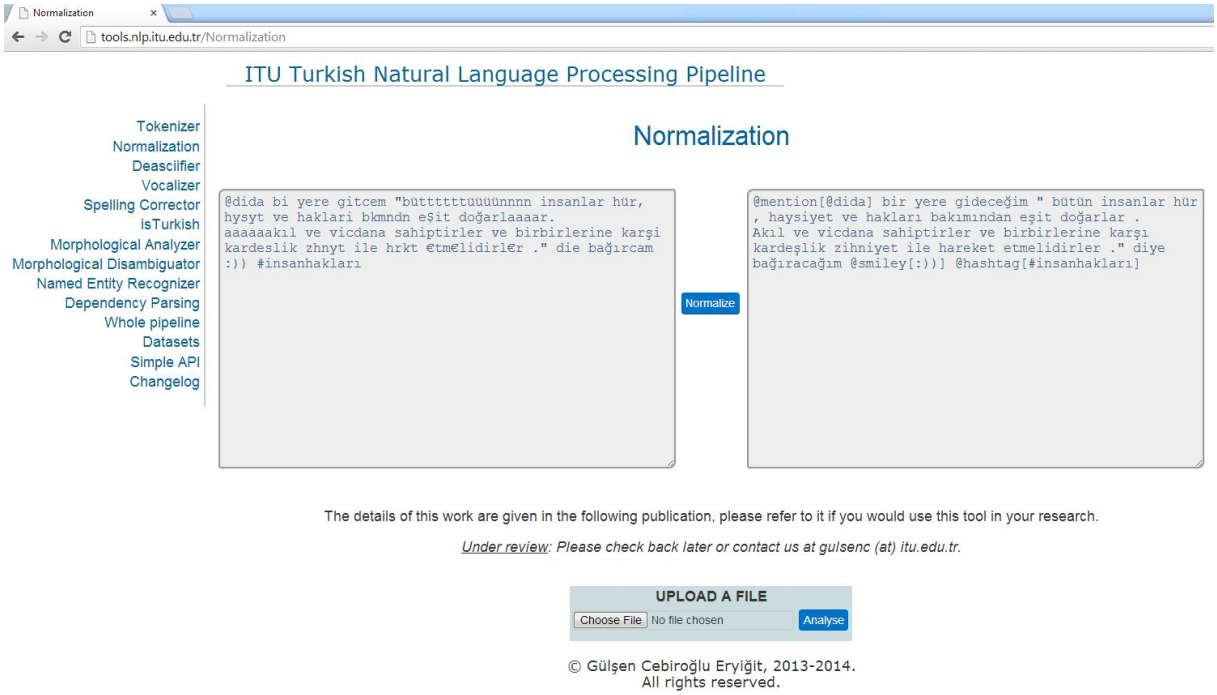

Figure 2: ITU Turkish NLP Web Interface

Gülşen Eryiğit and Eşref Adalı. 2004. An affix stripping morphological analyzer for Turkish. In *Proceedings of the International Conference on Artificial Intelligence and Applications*, pages 299–304, Innsbruck, 16-18 February.

Gulsen Eryigit, Tugay Ilbay, and Ozan Arkan Can. 2011. Multiword expressions in statistical dependency parsing. In *Proceedings of the Second Workshop on Statistical Parsing of Morphologically Rich Languages (IWPT)*, pages 45–55, Dublin, Ireland, October. Association for Computational Linguistics.

Gülşen Eryiğit. 2007. Itu treebank annotation tool. In *Proceedings of the ACL workshop on Linguistic Annotation (LAW 2007)*, Prague, 24-30 June.

Gülşen Eryiğit. 2012. The impact of automatic morphological analysis & disambiguation on dependency parsing of turkish. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC)*, Istanbul, Turkey, 23-25 May.

Gülşen Eryiğit, Joakim Nivre, and Kemal Oflazer. 2008. Dependency parsing of Turkish. *Computational Linguistics*, 34(3):357–389.

David Ferrucci and Adam Lally. 2004. UIMA: an architectural approach to unstructured information processing in the corporate research environment. *Natural Language Engineering*, 10(3-4):327–348.

Krister Lindén, Miikka Silfverberg, and Tommi Pirinen. 2009. Hfst tools for morphology–an efficient open-source package for construction of morphological analyzers. In *State of the Art in Computational Morphology*, pages 28–47. Springer.

Joakim Nivre, Johan Hall, Jens Nilsson, Atanas Chanev, Gülşen Eryiğit, Sandra Kübler, Stetoslav Marinov, and Erwin Marsi. 2007. Maltparser: A language-independent system for data-driven dependency parsing. *Natural Language Engineering Journal*, 13(2):99–135.

Kemal Oflazer. 1994. Two-level description of Turkish morphology. *Literary and Linguistic Computing*, 9(2):137–148.

Muhammet Şahin, Umut Sulubacak, and Gülşen Eryiğit. 2013. Redefinition of Turkish morphology using flag diacritics. In *Proceedings of The Tenth Symposium on Natural Language Processing (SNLP-2013)*, Phuket, Thailand, October.

Muhammet Şahin. 2014. ITUMorph, a more accurate and faster wide coverage morphological analyzer for Turkish. Master's thesis, Istanbul Technical University.

Haşim Sak, Tunga Güngör, and Murat Saraçlar. 2008. Turkish language resources: Morphological parser, morphological disambiguator and web corpus. In *GoTAL 2008*, volume 5221 of *LNCS*, pages 417–427. Springer.

Gökhan Akın Şeker and Gülşen Eryiğit. 2012. Initial explorations on using CRFs for Turkish named entity recognition. In *Proceedings of COLING 2012*, Mumbai, India, 8-15 December.

Dilara Torunoğlu and Gülşen Eryiğit. 2014. A cascaded approach for social media text normalization of Turkish. In *5th Workshop on Language Analysis for Social Media (LASM) at EACL*, Gothenburg, Sweden, April. Association for Computational Linguistics.

Ziqi Wang, Gu Xu, Hang Li, and Ming Zhang. 2011. A fast and accurate method for approximate string search. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 52–61.