# Lemmatize without morphology

**Proposers:** Rodrigo Agerri
**Contact:** rodrigo.agerri@ehu.eus

## Description

Lemmatization is a task in Natural Language Processing (NLP) which consists of producing, from a given inflected word form, its canonical form or lemma. In other words, the form that corresponds to its entry in the dictionary. Most human languages display inflected morphology, namely, the word form changes according to its morphosyntactic category. Lemmatization is one of the basic tasks that facilitate downstream NLP-based applications such as Machine Translation, Parsing or Information Retrieval. It is furthermore considered that lemmatization is of particular importance for high inflected languages such as Basque or Russian. Current approaches in NLP stress the importance of learning lemmatization in context. Moreover, previous work in contextual-based lemmatization mostly assumes that morphological information is crucial for the lemmatization task, and even more for morphologically rich languages, such as Basque. This is illustrated by the following:

| *Basque lemmatized* | *Basque* | *Spanish lemmatized* | *Spanish* |
|---|---|---|---|
| **etxe** | **etxe** | **casa** | **casa** |
| | **etxe**a | | **casa**s |
| | **etxe**ak | | |
| | **etxe**an | | |
| | **etxe**aren | | |
| | **etxe**ek | | |
| | **etxe**en | | |
| | **etxe**etako | | |
| | **etxe**etan | | |
| | **etxe**etara | | |
| | **etxe**ko | | |
| | **etxe**koak | | |
| | **etxe**ra | | |
| | **etxe**tatik | | |
| | **etxe**tik | | |
| | **etxe**z | | |

The way in which this task can be learned varies, but mostly current approaches combine two components: (i) trying to use the rich morphological information associated to each inflected form and, (ii) assuming that it is possible to directly learn the string edits required to convert the inflected form into its lemma; these approaches are usually based on calculating **edit distance scripts**, namely, **calculating the edit distance between the inflected form and its lemma and learning how many edits are required to transform the form into the lemma** (Chrupala, 2008).

As for many other tasks in NLP, current context-based supervised approaches to lemmatization are based on deep learning algorithms and vector-based word representations or word embeddings (Bergmanis and Goldwater 2018, Kondratyiuk 2019, Malaviya et al. 2019, McCarthy et al. 2019, Straka et al. 2019, Yildiz and Tantug 2019, among others). In any case, the large majority of approaches to neural context-based lemmatization use such morphological information, even arguing that assuming the lack of such annotation is not realistic (Malaviya et al. 2019). This particular claim is supported by the existence of the Universal Dependencies (UD) corpus (Nivre et al. 2017) which contains gold annotated data with lemmas and fine-grained morphological information (such as the one shown in the example above) for 90 human languages.

Previous work has shown (Toporkov 2020) that complex morphological tagging degrades too much when applied to out-of-domain data for which no gold morphological annotation is available, generating in turn cascading errors in lemmatization. This would be especially true for high-inflected languages such as Basque, Hungarian, Russian or Turkish. Taking this issue into account we hypothesize that developing neural contextual lemmatizers without morphological information must help to improve their out-of-domain performance. In order to do so, we propose to learn lemmatization using only **edit distance scripts,** such as the example shown in the following table (Straka et al. 2019).

| Lemma Rule | Casing Script | Edit Script | Most Frequent Examples |
|---|---|---|---|
| ↓0;d¦ | all lowercase | do nothing | the→the to→to and→and |
| ↑0¦↓1;d¦ | first upper, then lower | do nothing | Bush→Bush Iraq→Iraq Enron→Enron |
| ↓0;d¦- | all lowercase | remove last character | your→you an→a years→year |
| ↓0;abe | all lowercase | ignore form, use be | is→be was→be 's→be |
| ↑0;d¦ | all uppercase | do nothing | I→I US→US NASA→NASA |
| ↓0;d¦-- | all lowercase | remove last 2 chars | been→be does→do called→call |
| ↓0;d¦--- | all lowercase | remove last 3 chars | going→go being→be looking→look |
| ↓0;d--+b¦ | all lowercase | change first 2 chars to b | are→be 're→be Are→be |
| ↓0;d¦-+v+e | all lowercase | change last char to ve | has→have had→have Has→have |
| ↓0;d¦---+e | all lowercase | change last 3 chars to e | having→have using→use making→make |
| ↓0;d¦-+o→ | all lowercase | change last but 1 char to o | n't→not knew→know grew→grow |

Table 1: Eleven most frequent lemma rules in English EWT corpus, ordered from the most frequent one.

There are many methods to produce such edit distances between form and lemma (Chrupala 2008, Malaviya et al. 2019, Straka et al. 2019, Yildiz and Tantug 2019), and in this master thesis we propose to analyze the best performing ones as well as propose new variants of them.

# Objectives

The candidate may choose between the following objectives:

1. Evaluate edit distance methods for lemmatization using deep learning systems for sequence tagging.
2. Propose variants or new method for encoding edit distance between form and lemma.
3. Analyze performance of edit distance methods across domains and across language families.
4. Multilingual lemmatization using edit distance methods.

The master thesis can be written in Basque or English.

# Tasks and Plan

- Month 1: Start of the project, defining the objectives and tasks.

- Month 2: Start experiments. Optionally, it is recommended for the candidates to attend the "Seminar on language technologies. Deep Learning (LAP 18). https://ixa.si.ehu.es/master/programa_html
- Months 3-5: Experiments and final development.
- Final month: Writing up.

# References


Akbik, A.; Blythe, D.; Vollgraf, R. (2018). Contextual string embeddings for sequence labeling. In COLING 2018.

Toms Bergmanis and Sharon Goldwater. 2018. Context sensitive neural lemmatization with Lematus. In NAACL 2018.

Grzegorz Chrupała. 2008. Towards a machine-learning architecture for lexical functional grammar parsing. Ph.D. thesis, Dublin City University.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. In NAACL 2019.

Kondratyuk, D. (2019). Cross-Lingual Lemmatization and Morphology Tagging with Two-Stage Multilingual BERT Fine-Tuning. In Proceedings of the 16th Workshop on Computational Research in Phonetics, Phonology, and Morphology (pp. 12-18).

Martin Haspelmath and Andrea Sims. 2013. Understanding morphology. Routledge.

McCarthy, A. D., Vylomova, E., Wu, S., Malaviya, C., Wolf-Sonkin, L., Nicolai, G., & Cotterell, R. (2019). The SIGMORPHON 2019 shared task: Morphological analysis in context and cross-lingual transfer for inflection. In SIGMORPHON 2019.

Malaviya, C., Wu, S., & Cotterell, R. (2019). A simple joint model for improved contextual neural lemmatization. In NAACL 2019.

Joakim Nivre et al. 2017. Universal dependencies 2.0. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Straka, M., Straková, J., & Hajič, J. (2019). UDPipe at SIGMORPHON 2019: Contextualized embeddings, regularization with morphological categories, corpora merging. In SIGMORPHON 2019.

Yildiz, E., & Tantuğ, A. C. (2019). Morpheus: A neural network for jointly learning contextual lemmatization and morphological tagging. In Proceedings of the 16th Workshop on Computational Research in Phonetics, Phonology, and Morphology (pp. 25-34).