Lengoaia eta Sistema Informatikoak Saila

eman ta zabal zazu

Universidad del País Vasco
Euskal Herriko Unibertsitatea

On the Role of Morphology in Contextual Lemmatization

Rodrigo Agerri Gascón

informatika fakultatea
facultad de informática

Trabajo Original de Investigación

Presentado para optar a plaza de Personal Doctor Investigador Permanente con código IDPTCL1-D00141-1 en el Área de conocimiento de Lenguajes y Sistemas Informáticos (Resolución de la Universidad del País Vasco del 18 de diciembre de 2023; BOPV nº 246 del 28 de diciembre de 2023)

Donostia-San Sebastián, 22 de abril de 2024.

Contents

List of Figures

List of Tables

CHAPTER 1

On the Role of Morphology in Contextual Lemmatization

This chapter is based on the following publication:

Olia Toporkov and **Rodrigo Agerri** (2024). On the Role of Morphological Information for Contextual Lemmatization. In *Computational Linguistics* (MIT Press). Talk presented at the main conference of *EMNLP 2023*.

Abstract: Lemmatization is a Natural Language Processing (NLP) task which consists of producing, from a given inflected word, its canonical form or lemma. Lemmatization is one of the basic tasks that facilitate downstream NLP applications, and is of particular importance for high-inflected languages. Given that the process to obtain a lemma from an inflected word can be explained by looking at its morphosyntactic category, including fine-grained morphosyntactic information to train contextual lemmatizers has become common practice, without considering whether that is the optimum in terms of downstream performance. In order to address this issue, in this paper we empirically investigate the role of morphological information to develop contextual lemmatizers in six languages within a varied spectrum of morphological complexity: Basque, Turkish, Russian, Czech, Spanish and English. Furthermore, and unlike the vast majority of previous work, we also evaluate lemmatizers in out-of-domain settings, which constitutes, after all, their most common application use. The results of our study are rather surprising. It turns out that providing lemmatizers with fine-grained morphological features during training is not that beneficial, not even for agglutinative languages. In fact, modern contextual word representations seem to implicitly encode enough morphological information to obtain competitive contextual lemmatizers without seeing any explicit morphological signal. Moreover, our experiments suggest that the best lemmatizers out-of-domain are those using simple UPOS tags or those trained without morphology. Finally, we demonstrate that current evaluation practices for lemmatization are not adequate to clearly discriminate between models and to manifest the shortcoming of current lemmatization techniques.

1.1 Introduction

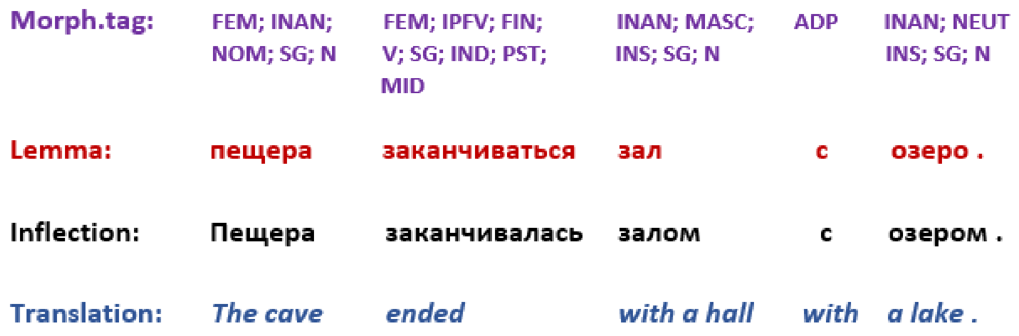
Lemmatization is one of the basic NLP tasks which consists of converting an inflected word form (e.g., *eating, ate, eaten*) into its canonical form (e.g., *eat*), usually known as the lemma, as formulated by the SIGMORPHON 2019 shared task (Aiken et al., 2019). Lemmatization is commonly used when expanding search criteria in information retrieval or to reduce dimensionality of problems in NLP tasks such as information extraction, text classification, and others. For example, for morphologically rich languages named entities are often inflected, which means that lemmatization is required as an additional process. Lemmatization is more challenging for languages with rich inflection as the number of variations for every different word form in such languages is very high. Table 1.1 illustrates this point by showing the differences in inflections of the word 'cat' for four languages with different morphological structure. This language sample offers a spectrum of varied complexity, ranging from the more complex ones, Basque and Russian, to the less inflected ones, such as Spanish and English, in that order.

English	Spanish	Russian	Basque
cat	gato	кот	katu
cats	gata	коты	katuak
	gatos	кота	katua
	gatas	коту	katuari
		котом	katuarekin
		коте	katuek
		котов	katuekin
		котам	katuei
		котами	katuen
		котах	katurik
			katuarentzat
			katuentzat
			...

Table 1.1: Examples of inflected forms of the word 'cat' in Basque, English, Spanish and Russian.

As we can see in Table 1.1, the word 'cat' can vary in English by changing from singular to plural. In Spanish gender (masculine/feminine) is also marked. Things get more complicated with languages that mark case. For example, in Russian there are six cases while for Basque there are 16, some of which can be doubly inflected.

Both the context in which it occurs and the morphosyntactic form of a word play a crucial role to approach automatic lemmatization (McCarthy et al., 2019). Thus, in Figure 1.1 we can see a fragment of a Russian sentence in which each inflected word form has a



Morph.tag:	FEM; INAN; NOM; SG; N	FEM; IPFV; FIN; V; SG; IND; PST; MID	INAN; MASC; INS; SG; N	ADP	INAN; NEUT INS; SG; N
Lemma:	пещера	заканчиваться	зал	с	озеро .
Inflection:	Пещера	заканчивалась	залом	с	озером .
Translation:	*The cave*	*ended*	*with a hall*	*with*	*a lake .*

Figure 1.1: Example of a morphologically tagged and lemmatized sentence in Russian using the UniMorph annotation scheme.

corresponding lemma (in red). Furthermore, each inflected form has an associated number of morphosyntactic features (expressed as tags) depending on its case, number, gender, animacy and others. Morphological analysis is crucial for lemmatization as it explains the process required to produce the lemma from the word form, which is why it has traditionally been used as a stepping stone to design systems to perform lemmatization.

As many other tasks in NLP, the first approaches to lemmatization were rule-based, but nowadays the best performing models address lemmatization as a supervised task in which learning in context is crucial. Regardless of the learning method used, three main trends can be observed in current contextual lemmatization: (i) those that use gold standard or learned morphological tags to generate features to learn lemmatization in a pipeline approach (Chrupala et al., 2008; Yildiz and Tantuğ, 2019); (ii) those that aim to jointly learn morphological tagging and lemmatization as a single task (Müller et al., 2015; Malaviya et al., 2019; Straka et al., 2019); (iii) systems that do not use any explicit morphological signal to learn to lemmatize (Chakrabarty et al., 2017; Bergmanis and Goldwater, 2018).

Research on contextual (mostly neural) lemmatization was greatly accelerated by the first release of the Universal Dependencies (UD) data (de Marneffe et al., 2014; Nivre et al., 2017), but specially by the contextual lemmatization shared task organized at SIGMORPHON 2019, which included UniMorph datasets for more than 50 languages (McCarthy et al., 2019). It should be noted that the best models in the task used morphological information either as features (Yildiz and Tantuğ, 2019) or as part of a joint or a multitask approach (Straka et al., 2019). However, the large majority of previous approaches have used all the morphological tags from UniMorph/UD assuming that fine-grained morphological information must be always beneficial for lemmatization, especially for highly inflected languages, but without analyzing whether that is the optimum in terms of downstream performance.

In order to address this issue, in this paper we empirically investigate the role of morphological information to develop contextual lemmatizers in six languages within a varied spectrum of morphological complexity: Basque, Turkish, Russian, Czech, Spanish and English. Furthermore, previous work has shown that morphological taggers substantially degrade when evaluated out-of-domain, be that any type of text different from the data used for training in terms of topic, text genre, temporality, etc. (Manning, 2011). This point led us to research whether lemmatizers based on fine-grained morphological information will degrade more when used out-of-domain than those requiring only coarse-grained UPOS tags. We believe that this is also an important point because lemmatizers are mostly used out-of-domain, namely, to lemmatize data from a different distribution with respect to the one that was employed for training.

Taking these issues into consideration, in this paper we set to investigate the following research questions with respect to the actual role of morphological information to perform contextual lemmatization. First, is fine-grained morphological information really necessary, even for high-inflected languages? Second, are modern context-based word representations enough to learn competitive contextual lemmatizers without including any explicit morphological signal for training? Third, do morphologically enriched lemmatizers perform worse out-of-domain as the complexity of the morphological features increases? Four, what is the optimal strategy to obtain robust contextual lemmatizers for out-of-domain settings? Finally, are current evaluation practices adequate to meaningfully evaluate and compare contextual lemmatization techniques?

The conclusions from our experimental study are the following: (i) fine-grained morphological features do not always benefit, not even for agglutinative languages; (ii) modern contextual word representations seem to implicitly encode enough morphological information to obtain state-of-the-art contextual lemmatizers without seeing any explicit morphological signal; (iii) the best lemmatizers out-of-domain are those using simple UPOS tags or those trained without explicit morphology; (iv) current evaluation practices for lemmatization are not adequate to clearly discriminate between models, and other evaluation metrics are required to better understand and manifest the shortcomings of current lemmatization techniques. The generated code and datasets are publicly available to facilitate the reproducibility of the results and further research on this topic.[1]

The rest of the paper is structured as follows. The next section discusses the most relevant work related to contextual lemmatization. The systems and datasets used in our experiments are presented in Sections 1.4 and 1.3, respectively. Section 1.5 presents the experimental setup applied to obtain the results, which are reported in Section 1.6. Section 1.7 provides a discussion and error analysis of the results. We finish with some concluding remarks in Section 1.8.

[1] https://github.com/oltoporkov/morphological-information-datasets

1.2 Background

First approaches to lemmatization consisted of systems based on dictionary lookup and/or rule-based finite state machines (Karttunen et al., 1992; Oflazer, 1993; Alegria et al., 1996; van den Bosch and Daelemans, 1999; Dhonnchadha, 2002; Segalovich, 2003; Carreras et al., 2004; Stroppa and Yvon, 2005; Jongejan and Dalianis, 2009). Grammatical rules in such systems, either hand-crafted or learned automatically by using machine learning, were leveraged to perform lemmatization together with the use of lexicons or morphological analyzers that returned the correct lemma. The problem of unseen and rare words was solved by generating a set of exceptions added to the general set of rules (Karttunen et al., 1992; Oflazer, 1993) or by using a probabilistic approach (Segalovich, 2003). Such systems resulted in very language-dependent approaches, and in most of the cases they required huge linguistic knowledge and effort, especially in the case of those languages with more complex, high-inflected morphology.

The appearance of large annotated corpora with morphological information and lemmas facilitated the development of machine learning methods for lemmatization in multiple languages. One of the core projects that gathered annotated corpora for more than 90 languages is the Universal Dependencies (UD) initiative (Nivre et al., 2017). This project offers a unified morphosyntactic annotation across languages with language-specific extensions when necessary. Based on the UD data, the Universal Morphology (UniMorph) project (McCarthy et al., 2020) converted the UD annotations into UniMorph, a universal tagset for morphological annotation (based on Sylak-Glassman (2016)), where each inflected word form is associated with a lemma and a set of morphological features. The current UniMorph dataset includes 118 languages, including extremely low-resourced languages such as Quechua, Navajo and Haida.

The assumption that context could help with unseen and ambiguous words led to the creation of supervised contextual lemmatizers. The pioneer work on this topic is perhaps the statistical contextual lemmatization model provided by Morfette (Chrupala et al., 2008). Morfette uses a Maximum Entropy classifier to predict morphological tags and lemmas in a pipeline approach. Interestingly, instead of learning the lemmas themselves, Chrupala et al. (2008) propose to learn automatically induced lemma classes based on the shortest edit script (SES), which consists of the number of edits necessary to convert the inflected word form into its lemma. Morfette has influenced many other works on contextual lemmatization, such as the system of Gesmundo and Samardžić (2012), IXA pipes (Agerri et al., 2014; Agerri and Rigau, 2016), Lemming (Müller et al., 2015) and the system of Malaviya et al. (2019). The importance of using context to learn lemmatization is investigated in the work of Bergmanis and Goldwater (2018). They compare context-free and context-sensitive versions of their neural lemmatizer Lematus and evaluate them across 20 languages. Results show that including context substantially improves lemmatization accuracy and it helps to better deal with the out-of-vocabulary problem.

The next step in the development of contextual lemmatization systems came with the supervised approaches based on deep learning algorithms and vector-based word representations (Chakrabarty et al., 2017; Dayanik et al., 2018; Bergmanis and Goldwater, 2018; Malaviya et al., 2019). The parallel development of the Transformer architecture (Vaswani et al., 2017) and the appearance of BERT (Devlin et al., 2019) and other Transformer-based masked language models (MLMs) offered the possibility to significantly improve lemmatization results. Thus, most of the participating systems in the SIGMORPHON 2019 shared task on contextual lemmatization for 66 languages were based on MLMs (McCarthy et al., 2019). The baseline provided by the task was based on the work of Malaviya et al. (2019), a system which performs joint morphological tagging and lemmatization.

To the best of our knowledge, current state-of-the-art results in contextual lemmatization are provided by those models that achieved best results in the SIGMORPHON 2019 shared task. The highest overall accuracy was achieved by UDPipe (Straka et al., 2019). Using UDPipe 2.0 (Straka, 2018) as a baseline, they added pre-trained contextualized BERT and Flair embeddings as an additional input to the network. The overall accuracy (average across all languages) was 95.78, the best among all the participants.

The second best result (95 overall word accuracy) in the task was obtained by the CHARLES-SAARLAND system (Kondratyuk, 2019). This system consists of a combination of a shared BERT encoder and joint lemma and morphology tag decoder. The model uses a two-stage training process, in which it first performs a multilingual training over all treebanks, and then they execute the same process monolingually, maintaining the previously learned multilingual weights. Morphological tags in this case are calculated jointly and lemmas are also represented as SES. The experiments are performed using multilingual BERT in combination with the methods introduced by UDify (Kondratyuk and Straka, 2019) for BERT fine-tuning and regularization.

The third best result (94.76) was reported by Morpheus (Yildiz and Tantuğ, 2019). Morpheus uses a two-level LSTM network which gets as input the vector-based representations of words, morphological tags and SES. Morpheus then aims to jointly output, for a given sequence, their corresponding morphological labels and the SES representing the lemma class which is later decoded into its lemma form.

Thus, it can be seen that a common trend in current contextual lemmatization is to use the morphological information provided by the full UniMorph labels without taking into consideration whether this is the optimal setting. Furthermore, lemmatization techniques are only evaluated in-domain, resulting in extremely, and perhaps deceptive, high results for the large majority of the 66 languages included in the SIGMORPHON 2019 data.

1.3 Languages and Datasets

In order to address the research questions formulated in the Introduction, we selected the following six languages: Basque, Turkish, Russian, Czech, Spanish and English. Such a choice will allow us to compare the role of fine-grained morphological information to learn contextual lemmatization within a range of languages of varied morphological complexity. In this section we briefly describe general morphological characteristics of each language as well as the specific datasets used.

1.3.1 Languages

Basque and Turkish are agglutinative languages with morphology mostly of the suffixing type. Basque is a language isolate and does not belong to any language group while Turkish is a member of the Oghuz group of the Turkic family. These two languages have no grammatical gender, with some particular exceptions for domestic animals, people and foreign words (Turkish) or in some colloquial forms when the gender of the addressee is expressed for the second person singular pronoun (Basque). Turkish and Basque have two number types (singular and plural), and in Basque there is also the unmarked number (undefined or *mugagabea*). In both Turkish and Basque the cases are expressed by suffixation.

Basque is an ergative-absolutive language containing 16 cases, meaning that the grammatical case marks both the subject of an intransitive verb and the object of a transitive verb. The verb conjugation is also specific for this language: the majority of the verbs are formed by a combination of a gerund form and a conjugated auxiliary verb.

Turkish has six general cases; nouns and adjectives are not distinguished morphologically and adjectives can also be used as adverbs without modifications or by doubling of the word. For verbs there are 9 simple and 20 compound tenses. There is a relatively small set of core vocabulary and the majority of Turkish words originate from applying derivative suffixes to nouns and verbal stems.

The two Slavic languages, namely, Russian and Czech, which have a fusional morphological system, exhibit a highly inflectional morphology and a wide number of morphological features. Russian belongs to the East Slavic language group, while Czech is a West Slavic language. These two languages have nominal declension which involves six main grammatical cases for Russian and seven for Czech. Both languages distinguish between two number (singular and plural) and three gender types (masculine, feminine and neuter). Furthermore, the masculine gender is subdivided into animate and inanimate. Verbs are conjugated for tense (past, present or future) and mood.

Spanish is a Romance language that belongs to Indo-European language family. It is a fusional language, which has a tendency to use a single inflectional morpheme to denote multiple grammatical, syntactic or semantic features. Nouns and adjectives in Spanish have two gender (male, female) and two number types (singular and plural). Besides, some articles, pronouns and determiners also possess a neuter gender. There are 3 main verb tenses (past, present and future) and each verb has around fifty conjugated forms. Apart from that, Spanish has 3 verboid forms (infinitive, gerund, past participle), perfective and imperfective aspects for past, 4 moods and 3 persons.

Finally, English is a Germanic language, also part of Indo-European language family. It has lower inflection in comparison to previously mentioned languages. Only nouns, pronouns and verbs are inflected, while the rest of the parts of speech are invariable. In English animate nouns have two genders (masculine or feminine) and the third person singular pronouns distinguish three gender types: masculine, feminine, and neuter, while for most of the nouns there is no grammatical gender. Nouns have only a genitive case and personal pronouns are mostly declined in subjective and objective cases. English has a variety of auxiliary verbs that help to express the categories of mood and aspect and participate in the formation of verb tenses.

1.3.2 Datasets

The datasets we used are distributed as part of the data used for the SIGMORPHON 2019 shared task (McCarthy et al., 2019). The source of the original datasets comes from the Universal Dependencies (UD) project (de Marneffe et al., 2014), but the morphological annotations are converted from UD annotations to the UniMorph schema (Kirov et al., 2018) with the aim of increasing agreement across languages. As our experiments will include both in-domain and out-of-domain evaluations, we selected some datasets for each of the settings.

With respect to *in-domain*, we chose one corpus per language using the standard train and development partitions. For Basque we used the Basque Dependency Treebank (BDT) (Aldezabal et al., 2008), which contains mainly literary and journalistic texts. The corpus was manually annotated and then automatically converted to UD format. For Czech we used the CAC treebank (Hladká et al., 2008) based on the Czech Academic Corpus 2.0. This corpus includes mostly unabridged articles from a wide range of media such as newspapers, magazines and transcripts of spoken language from radio and TV programs. The corpus was annotated manually and then converted to UD format. With respect to English we chose English Web Treebank (EWT) (Silveira et al., 2014). This corpus includes different Web sources: blogs, various media, e-mails, reviews and Yahoo! answers. In the EWT corpus the lemmas were assigned by UD-converter and manually corrected. UPOS tags were also converted to UD format from manual annotations. For Russian we used GSD corpus, extracted from Wikipedia and manually annotated by native speakers. In the case of Spanish we selected the GSD corpus as well, consisting of texts from blogs, reviews, news and Wikipedia. Finally, for Turkish we used ITU-METU-Sabanci Treebank (IMST) (Sulubacak et al., 2016). It consists of well-edited sentences from a wide range of domains, manually annotated and automatically converted to UD format.

For the *out-of-domain* evaluation setting we picked the test sets of other datasets included in UniMorph, different from the ones selected for in-domain experimentation. In the case of Basque, only one corpus was available in the Universal Dependencies project, so we used the Armiarma corpus which consists of literary critics semi-automatically annotated using Eustagger (Alegria et al., 1996). For Czech and Turkish we used the PUD data – part of the Parallel Universal Dependencies treebanks created for the CoNLL 2017 shared task (Zeman et al., 2017). The corpora consist of 1,000 sentences from the news domain and Wikipedia annotated for 18 languages. The Czech language PUD data was manually annotated and then automatically converted to UD format. For Turkish the original data was automatically converted to UD format, but later manually reannotated (Türk et al., 2019). In the case of English we used the Georgetown University Multilayer (GUM) corpus (Zeldes, 2017). This corpus presents a collection of annotated Web texts from interviews, news, travel guides, academic writing, biographies and fiction from such sources as Wikipedia, Wikinet and Reddit. Its lemmas were manually annotated, while UPOS tags were converted to UD format from manual annotations. In the case of Russian we used SynTagRus (Lyashevkaya et al., 2016), which consists of texts from a variety of genres, such as contemporary fiction, popular science, as well as news and journal articles from the 1960-2016 period. Its lemmas, UPOS tags and morphological features were manually annotated in non-UD style and then automatically converted to UD format. For Spanish we chose the AnCora corpus (Taulé et al., 2008), which contains mainly texts from news. All the elements of this corpus were converted to UD format from manual annotations.

1.4 Systems

In this section we present the systems that we will be applying in our investigation. First, research on the role of fine-grained morphological information for contextual lemmatization will be performed in-domain using the statistical lemmatizer from the IXA pipes toolkit (Agerri and Rigau, 2016) and Morpheus, the third best system in the SIGMORPHON 2019 shared task. These two systems were chosen due to several reasons: (i) both use morphological information as features to learn lemmatization and, (ii) both systems use SES to represent automatically induced lemma classes; and (iii), they both address contextual lemmatization as sequence tagging.

In order to investigate whether modern contextual word representations are enough to learn, without any explicit morphological signal, competitive lemmatizers both in- and out-of-domain, we train baseline models using Flair (Akbik et al., 2018), multilingual MLMs mBERT and XLM-RoBERTa (Devlin et al., 2019; Conneau et al., 2020) as well as language-specific MLMs for each of the languages: BERTeus for Basque (Agerri et al., 2020), slavicBERT for Czech (Arkhipov et al., 2019), RoBERTa for English (Liu et al., 2019), Russian ruBERT (Kuratov and Arkhipov, 2019), Spanish BETO (Cañete et al., 2020) and BERTurk for Turkish.[2] As with Morpheus and IXA pipes, we treat contextual lemmatization as a sequence tagging task and fine-tune the language models by adding a single linear layer to the top of the model. The experiments were implemented using the HuggingFace Transformers API (Wolf et al., 2020).

[2] https://github.com/stefan-it/turkish-bert

1.4.1 Systems using morphology

IXA pipes is a set of multilingual tools which is based on a pipeline approach (Agerri et al., 2014; Agerri and Rigau, 2016). IXA pipes learns perceptron (Collins, 2002) models based on shallow local features combined with pre-trained clustering features induced over large unannotated corpora. The lemmatizer implemented in IXA pipes is inspired by the work of Chrupala et al. (2008), where the model learns the SES between the word form and its lemma. IXA pipes allows to learn lemmatization using gold-standard or learned morphological tags.

Morpheus is a neural contextual lemmatizer and morphological tagger which consists of two separate sequential decoders for generating morphological tags and lemmas. The input words and morphological features are encoded in context-aware vector representations using a two-level LSTM network and the decoders predict both the morphological tags and the SES, which are later decoded into its lemma (Yildiz and Tantuğ, 2019). Morpheus obtained the third best overall result in the SIGMORPHON 2019 shared task (McCarthy et al., 2019).

1.4.2 Systems without explicit morphological information

We train a number of models that use modern contextual word representations by addressing lemmatization as a sequence tagging task. Thus, the input consists of words encoded as contextual vector representations and the task is to assign the best sequence of SES to a given input sequence.

Flair is a NLP framework based on a BiLSTM-CRF architecture (Huang et al., 2015; Ma and Hovy, 2016) and pre-trained language models that leverage character-based word representations which, according to the authors, capture implicit information about natural language syntax and semantics. Flair has obtained excellent results in sequence labelling tasks such as named entity recognition, POS tagging and chunking (Akbik et al., 2018). The library includes pre-trained Flair language models for every language except Turkish.

With respect to the MLMs, we use two multilingual models and 6 language models trained specifically for each of the languages included in our study. Multilingual BERT (Devlin et al., 2019) is a Transformer-based masked language model, pre-trained on the Wikipedias of 104 languages with both the masking and next sentence prediction objectives. Furthermore, we also use XLM-RoBERTa (Conneau et al., 2020), trained on 2.5TB (295K millions of tokens) of filtered CommonCrawl data for 100 languages. XLM-RoBERTa is based on the BERT architecture but (i) trained only on the MLM task, (ii) on larger batches (iii) on longer sequences and (iv), with dynamic mask generation. Thus, multilingual BERT was trained with a batch size of 256 and 512 sequence length for 1M steps, using both the MmLM and NSP tasks. Regarding XLM-RoBERTa, both versions (base and large) were trained over 1.5M steps with batch 8192 and sequences of 512 length.

Language	Model	Architecture	Training corpus and number of tokens
Basque	BERTeus	BERT	35M tokens (Wikipedia) + 191M tokens (online)
Czech	slavicBERT	BERT	Russian news and Wikipedia in Russian, Bulgarian, Czech and Polish
English	RoBERTa	BERT	BookCorpus (800M tokens),
			CC-News (16,000M tokens), OpenWebText (8,706M tokens), CC-Stories (5,300M tokens)
Russian	ruBERT	BERT	Dataset for original BERT (BookCorpus(800M tokens)), English Wikipedia (2,500M tokens), Russian news and Wikipedia for subword vocabulary
Spanish	BETO	BERT	Wikipedia and OPUS project in Spanish (3,000M tokens)
Turkish	BERTurk	BERT	OSCAR corpus, Wikipedia, OPUS corpora, corpus of Kemal Oflaizer (4,404M tokens total)

Table 1.2: List of language-specific models used in the experiments for each of the target languages.

Details about the six language-specific MLMs used are provided in Table 1.2. BERTeus (Agerri et al., 2020) is a BERT-base model trained on the BMC Basque corpus, which includes the Basque Wikipedia and news articles from online newspapers. Apart from the training data, the other difference from original BERT is the subword tokenization, which is closer to linguistically interpretable strings in Basque. BERTeus significantly outperforms multilingual BERT and XLM-RoBERTa in tasks such as POS tagging, named entity recognition, topic modelling and sentiment analysis.

BERTurk[3] is a cased BERT-base model for Turkish. This model was trained on a filtered and sentence segmented version of the Turkish OSCAR corpus (Ortiz Suárez et al., 2019), together with Wikipedia, various OPUS corpora (Tiedemann, 2016) and data provided by Kemal Oflazer, which resulted in total size of 35GB (4,404M tokens total).

For Czech we used slavicBERT (Arkhipov et al., 2019), developed by taking

[3] https://github.com/stefan-it/turkish-bert

multilingual BERT as a basis and further pre-trained using Russian news and the Wikipedias of four Slavic languages: Russian, Bulgarian, Czech and Polish. The authors also rebuilt the vocabulary of subword tokens, using the subword-nmt repository.[4]

RuBERT was developed in a similar fashion as slavicBERT but only with Russian as target language using the Russian Wikipedia and news corpora (Kuratov and Arkhipov, 2019). They generated a new subword vocabulary obtained from subword-nmt which contains longer Russian words and subwords.

For Spanish we used BETO (Cañete et al., 2020) – a BERT-base language model, trained on a large Spanish corpus. The authors of this model upgraded the initial BERT model by using the Dynamic Masking technique, introduced in RoBERTa. BETO performed 2M steps in two different stages: 900K steps with a batch size of 2048 and maximum sequence length of 128, and the rest of the training with a batch size of 256 and maximum sequence length of 512. We use the version trained with cased data, which included the Spanish Wikipedia and various sources from the OPUS project (Tiedemann, 2012) in a final corpus size of around 3 billion words.

RoBERTa-base is the model chosen for English. RoBERTa (Liu et al., 2019) is an optimized version of BERT, as commented above. To train this model the authors, apart from the standard datasets used to train the BERT model, also used the CC-news dataset, including English news articles from all over the world published between January 2017 and December 2019. The total size of the training data exceeds 160GB of uncompressed text (more than 30 billion tokens).

1.4.3 Baselines

We use two models as baselines. First, the system used as a baseline for the SIGMORPHON 2019 shared task (McCarthy et al., 2019), a joint neural model for morphological tagging and lemmatization presented by Malaviya et al. (2019). This system performs morphological tagging by using a LSTM tagger described in Heigold et al. (2017) and Cotterell and Heigold (2017). The lemmatizer is a neural sequence-to-sequence model (Wu and Cotterell, 2019) which includes a hard attention mechanism with a training scheme based on dynamic programming. The tagger and lemmatizer are connected together by jackknifing (Agić and Schluter, 2017), which allows to avoid exposure bias and improve lemmatization results.

The second baseline is the winner of the SIGMORPHON'19 shared task (Straka et al., 2019). UDPipe is a multitask model which jointly learns morphological tagging and lemmatization. The system architecture consists of three bidirectional LSTMs that process the input and softmax classifiers that generate lemmas and morphosyntactic features. Lemmatization is performed as a multiclass classification task, where the system predicts the correct lemma rule or SES.

[4] https://github.com/rsennrich/subword-nmt/

1.5 Experimental Setup

The systems described above were trained on the datasets listed in Section 1.3.2 using the following methodology. For the two IXA pipes models (using gold-standard and learned morphology) we used the default feature set, with and without clustering features, specified in Agerri and Rigau (2016). The default hyperparameters were also applied to train Morpheus (Yildiz and Tantuğ, 2019). The input character embedding length d_a is set to 128, the length of the word vectors d_e to 1024 and the length of the context-aware word vectors d_c to 2048. Moreover, the length of character vectors in the minimum edit prediction component d_u and the length of the morphological tag vectors d_v are set to 256. The hidden unit sizes in the decoder LSTMs d_g and d_q are set to 1024. The Adam optimization algorithm is used with learning rate 3e-4 to minimize the loss (Kingma and Ba, 2015).

Flair is used off-the-shelf with FastText CommonCrawl word embeddings (Grave et al., 2018) combined with Flair contextual embeddings for each of the languages. The hidden size of the LSTM is set to 256 with a batch of 16.

The MLMs were fine-tuned for lemmatization as a sequence tagging task by adding a single linear layer on top of the model being fine-tuned. A grid search of hyperparameters was performed to pick the best batch size (16, 32), epochs (5, 10, 15, 20, 25) and learning rate (1e-0, 2e-5, 3e-5, 5e-5). We pick the best model on the development set in terms of word accuracy and loss. A fixed seed is used to ensure reproducibility of the results.

For multilingual BERT we used a maximum sequence length of 128, batch size 32 and 5e-5 as learning rate while for XLM-RoBERTa we used the same configuration but with a batch of 16. For Russian we perform grid search on two language-specific models, namely, ruBERT and slavicBERT. RuBERT obtained the best results with a maximum sequence length of 128, batch size 16, and a 5e-5 value for learning rate over 15 epochs. For the rest of the models the best configuration was that of XLM-RoBERTa over 5 epochs for BETO and RoBERTa-base, 10 epochs for BERTeus, 15 epochs for BERTurk and 20 epochs with slavicBERT for Czech.

1.6 Experimental Results

In this section we present the experiments to empirically address the following research questions with respect to the actual role of morphological information to perform contextual lemmatization, namely, (i) is fine-grained morphological information really necessary, even for agglutinative languages? (ii) are modern context-based word representations enough to learn competitive contextual lemmatizers without including any explicit morphological signal during training? (iii) do morphologically enriched lemmatizers perform worse out-of-domain as the complexity of the morphological features increases? (iv) what is the optimal strategy to obtain robust contextual lemmatizers for out-of-domain settings? and (v), are current evaluation practices adequate to meaningfully evaluate and compare contextual lemmatization techniques?

Unlike the vast majority of previous work on contextual lemmatization, which has been mostly evaluated *in-domain* (McCarthy et al., 2019), we also report results in *out-of-domain* settings. It should be noted that by *out-of-domain* we mean to evaluate the model on a different data distribution from the data used for training (Manning, 2011).

First, Section 1.6.1 studies the in-domain performance of contextual lemmatizers depending on the type of morphological features used to inform the models during training. The objective is two-fold: to determine whether complex (or any at all) morphological information is required to obtain competitive lemmatizers and, secondly, to establish whether modern contextual word representations and MLMs allow us to perform lemmatization without any morphological information.

Second, in the *out-of-domain* evaluation presented in Section 1.6.2 we analyze the performance of morphologically informed lemmatizers. Furthermore, comparing them with contextual lemmatizers developed without an explicit morphological signal would allow us to obtain a full picture as to what is the best strategy for out-of-domain settings (the most common application scenario).

1.6.1 In-domain evaluation

For the first experiment we train the two variants of the IXA pipes statistical system, ixa-pipe-gs and ixa-pipe-mm (Agerri and Rigau, 2016), and one neural lemmatizer, Morpheus (Yildiz and Tantuğ, 2019). As explained in Section 1.4, all three require explicit morphological information and they all apply shortest edit scripts (SES) to automatically induce lemma classes from the training data.

Morphological label	**Short Form**
UPOS	UPOS
UPOS+Case+Gender	UCG
UPOS+Case+Number	UCN
UPOS+Case+Gender+Number	UCGN
UPOS+AllFeaturesOrdered	UAllo

Table 1.3: List of UniMorph morphological tags used.

Furthermore, we combined the UniMorph morphological tags to generate labels of different complexity. Thus, taking UPOS tags as a basis we obtain 5 different morphological tags, as shown in Table 1.3. The first 4 are combinations of UPOS, case, gender and number. The last label includes UPOS and every feature present for a given word in UniMorph in the following order: {UPOS+Case+Gender+Number+All}. For some word types, such as prepositions or infinitives, UniMorph only includes the UPOS tag. In order to illustrate this, Table 1.4 provides an example originally in Russian

Word form	**Morphological label**	**Lemma**
Проект[*Project*]	NNOMMASC	проект[*project*]
сильно[*a lot*]	ADV	сильно[*a lot*]
отличался[*differed*]	VMASC	отличаться[*to differ*]
от[*from*]	ADP	от[*from*]
предыдущих[*previous*]	ADJGEN	предыдущий[*previous*]
подлодок[*submarines*]	NGENFEM	подлодка[*submarine*]
.	_	.

Table 1.4: An example of the data used to train contextual lemmatizers with {UPOS+Case+Gender} (UCG) morphological information.

including the information required to train contextual lemmatizers, namely, the word, some morphological tag, and the lemma.

Putting it all together, Table 1.5 characterizes the final datasets used for in- and out-of-domain evaluation. The number of tokens, unique labels per category and unique SES (calculated using the UDPipe method) illustrate the varied complexity of the languages involved.[5] Thus, those languages with more complex morphology have a higher number of unique labels that include additional morphological features. The same pattern can be seen in the amount of lemma classes (SES), significantly larger for the languages with more complex morphology. In the case of Turkish the low number of lemmas could be explained by the fact that most Turkish words are formed by applying derivative suffixes to nouns and verbal stems. Moreover, the core vocabulary in this particular corpus is rather small. Finally, we decided to order the subtags comprising the full UniMorph labels as the number of unique labels decreased significantly.

Table 1.6 reports the in-domain results of training the three systems for the six languages with the 5 different types of morphological labels. First, the results show that the neural lemmatizer Morpheus outperforms the statistical lemmatizers for every language except English. In fact, for languages with more complex morphology, such as Basque and Turkish, the differences are larger. Second, if we look at the impact of including fine-grained morphological features it can be seen that no single morphological tag performs best across systems and languages. Thus, while adding case, number and/or gender seems to be slightly beneficial, differences in performance are substantial when training the statistical lemmatizer using gold-standard morphological labels (ixa-pipe-gs) and especially for languages with more complex morphology (Basque, Russian, Turkish). Third, the results clearly show that adding every available morphological feature is not beneficial *per se*. Fourth, the statistical lemmatizer trained with learned morphological tags (ixa-pipe-mm) performs significantly worse in every case except for English and

[5] Even though it is not required for out-of-domain evaluation, the UniMorph information is not available for the Basque Armiarma corpus because it is not part of the UniMorph project.

lang	data	#toks	UPOS	UCGN	UAllo	UnAllo	SES
Basque	train (BDT)	97,336	15	205	1,143	1,683	1,306
	dev (BDT)	12,206	14	148	556	787	432
	test (BDT)	11,901	14	153	545	773	428
	test (Armiarma)	299,206	-	-	-	-	1,495
Czech	train (CAC)	395,043	16	332	1,266	1,784	946
	dev (CAC)	50,087	16	298	876	1,129	536
	test (CAC)	49,253	15	284	827	1,036	556
	test (PUD)	1,930	14	175	288	292	151
Russian	train (GSD)	79,989	14	241	851	1,384	553
	dev (GSD)	9,526	14	191	435	673	235
	test (GSD)	9,874	14	203	455	713	258
	test (SynTagRus)	109,855	15	247	757	1,243	896
Spanish	train (GSD)	345,545	25	116	287	510	310
	dev (GSD)	42,545	23	100	208	342	200
	test (GSD)	43,497	23	103	222	387	200
	test (AnCora)	54,449	15	75	178	309	298
English	train (EWT)	204,857	16	43	94	173	233
	dev (EWT)	24,470	16	41	88	160	120
	test (EWT)	25,527	16	41	85	156	115
	test (GUM)	8,189	17	42	72	124	80
Turkish	train (IMST)	46,417	15	124	1,541	1,897	211
	dev (IMST)	5,708	15	95	605	748	106
	test (IMST)	5,734	16	100	589	725	104
	test (PUD)	1,795	15	66	217	220	59

Table 1.5: Language complexity reflected in the number of labels according to the complexity of the morphological features, number of lemma classes and corpus tokens.

Spanish. Finally, adding a special label ‘no-tag’ with no morphological information shows that performance decreases significantly for every system and language.

Summarizing, *in-domain* performance for high-inflected languages improves when some fine-grained morphological attributes (case and number or gender) are used to train the statistical lemmatizers. However, for English and Spanish using UPOS seems to be enough. Thus, in the case of neural lemmatization with Morpheus (the best of the models using morphological information), we can see that no substantial gains are obtained by adding fine-grained morphological features to UPOS tags, not even for agglutinative languages such as Basque or Turkish.

This point is reinforced by the results of computing the McNemar test of statistical significance to establish whether the differences in the results obtained by Morpheus (the best among the models trained with morphology) informed only with UPOS labels or

English						
	no-tag	**UPOS**	**UCG**	**UCN**	**UCGN**	**UAllo**
ixa-mm	-	98.97	98.97	99.03	98.97	98.86
ixa-gs	96.98	99.51	99.49	99.58	99.59	**99.65**
morpheus	97.60	98.20	98.12	98.13	98.19	98.14
Spanish						
ixa-mm	-	98.75	98.74	98.71	98.78	98.74
ixa-gs	98.36	98.82	98.78	98.82	98.80	98.88
morpheus	98.17	98.09	98.93	**98.96**	98.92	98.91
Russian						
ixa-mm	-	94.85	95.37	95.69	95.50	95.53
ixa-gs	91.85	95.05	96.95	96.45	96.99	97.04
morpheus	96.50	96.92	96.91	97.10	97.18	**97.24**
Basque						
ixa-mm	-	93.19	93.22	93.14	93.30	93.49
ixa-gs	91.68	93.50	94.33	94.58	94.58	96.50
morpheus	95.48	96.30	96.43	**96.54**	96.37	96.42
Czech						
ixa-mm	-	97.76	97.17	97.29	97.10	97.10
ixa-gs	95.64	97.68	98.10	97.93	98.09	98.20
morpheus	98.37	98.78	**98.84**	98.83	98.82	98.80
Turkish						
ixa-mm	-	84.83	84.51	85.06	85.06	83.95
ixa-gs	85.97	88.81	88.89	89.14	89.14	90.52
morpheus	96.04	96.41	**96.53**	95.95	96.27	96.50

Table 1.6: In-domain lemmatization results on the development sets for systems that use morphology to train contextual lemmatizers. ixa-mm: IXA pipes with learned morphological tags; ixa-gs: IXA pipes with gold standard morphology.

with the best morphological label (as by Table 1.6 above) are statistically significant or not (null hypothesis). The result of the test showed that for every language the differences were not significant ($\alpha = .05$, with 0.936 p-value for Basque, 0.837 for Czech, 0.511 for Russian and 0.942 for Spanish).

Taking this into consideration, the next natural step is to consider whether it is possible to learn good contextual lemmatizers without providing any explicit morphological signal during training. Previous work on probing contextual word representations and Transformer-based masked language models (MLMs) suggests that such models implicitly encode information about part-of-speech and morphological features (Manning et al., 2020; Akbik et al., 2018; Conneau et al., 2018; Belinkov et al., 2017). Following this, for this experiment we fine-tune various well-known multilingual and monolingual language models (detailed in Section 1.4) by using only the word forms and the automatically induced shortest edit scripts (SES) as implemented by UDPipe (Straka et al., 2019).

Figure 1.2 reports the results. From left-to-right, the first three bars correspond to the best statistical and Morpheus models using explicit morphological information as previously reported in Table 1.6. The next four list the results from Flair, mBERT, XLM-RoBERTa-base and a language-specific monolingual model (none of these four use any explicit morphological signal) whereas *base* (dark purple) refers to the system of Malaviya et al. (2019), employed as a baseline for the SIGMORPHON 2019 shared task (McCarthy et al., 2019). For state-of-the-art comparison, the last column on the right provides the results from UDPipe (Straka et al., 2019) (light purple color). Finally, the dark blue bars represent the best result for each language without considering either the baseline system or UDPipe.

The first noticeable trend is that every model beats the baseline except the IXA pipes-based statistical lemmatizers, which perform over the baseline and comparatively to the other models for English and Spanish only, the languages with the less complex morphology.

The second and, perhaps, most important fact is that the four models (Flair, mBERT, XLM-RoBERTa and mono) which do not use any morphological signal for training, obtain a remarkable performance across languages, XLM-RoBERTa-base being the best overall, even better than language-specific monolingual models. In fact, XLM-RoBERTa-base outperforms Morpheus for 4 out of the 6 languages, a neural model which was the third best system in the SIGMORPHON 2019 benchmark and which uses all the morphological information available in the UniMorph data. The McNemar test of significance shows that the differences in results obtained by Morpheus and XLM-RoBERTa are statistically significant ($\alpha = .05$) for Russian, Spanish and English (in XLM-RoBERTa's favour), and for Basque and Turkish (Morpheus over XLM-RoBERTa).

An additional observation is our XLM-RoBERTa-base lemmatization models perform competitively with respect to UDPipe, which obtains the best results for 5 out of the 6 languages included in our study. XLM-RoBERTa also outperforms the monolingual models, a behaviour that has been reported for other NLP tasks Agerri and Agirre (2023).

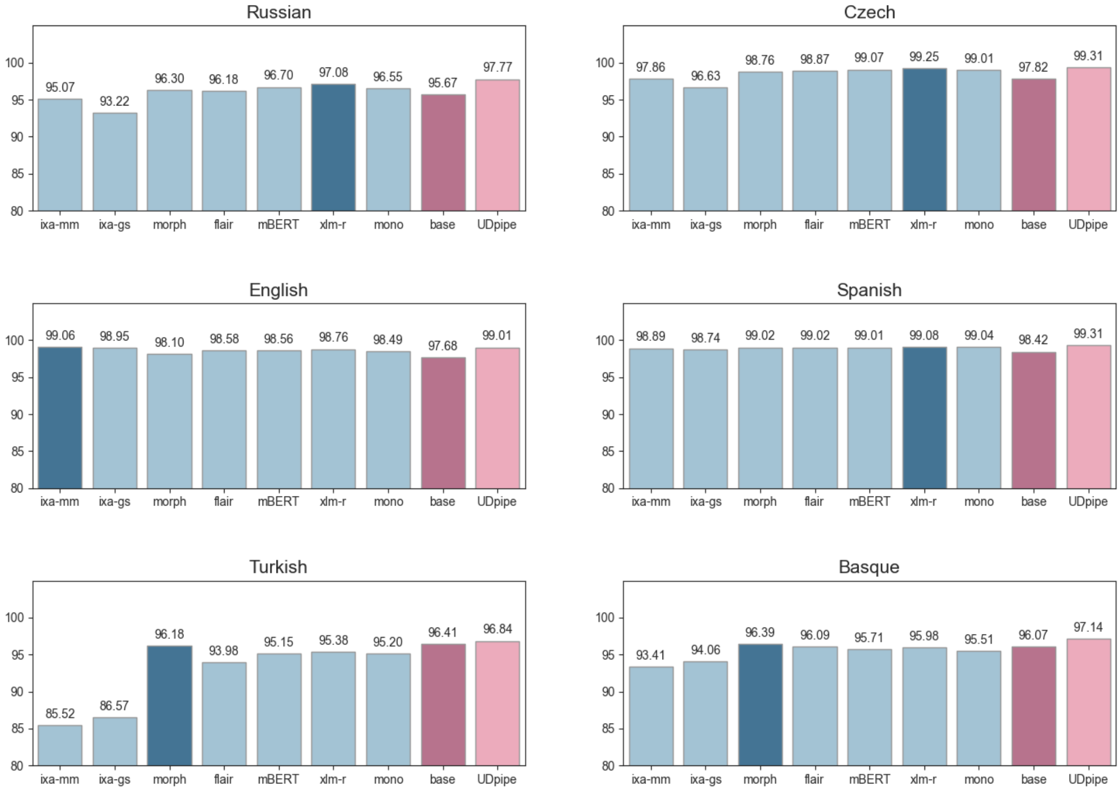


Figure 1.2: Overall *in-domain* lemmatization results on the test data for models trained with and without explicit morphological features; monolingual Transformers: Russian - ruBERT, Czech - slavicBERT, Basque - BERTeus, Turkish - BERTurk, English - RoBERTa, Spanish - BETO.

In any case, UDPipe's strong performance is somewhat expected as it was the overall winner of the SIGMORPHON 2019 lemmatization task. It should be noted that UDPipe is a rather complex system consisting of a multitask model to predict POS tags, lemmas and dependencies by applying three shared bidirectional LSTM layers which take as input a variety of word and character embeddings, the final model being an ensemble of 9 possible embedding combinations. However, the results obtained by the language models we trained without any explicit morphological signal, such as XLM-RoBERTa-base, are based on a simple baseline setting, where the Transformer models are fine-tuned using the automatically induced SES as the target labels in a token classification task. These results seem to confirm that, as it was the case for POS tagging and other tasks (Manning et al., 2020), contextual word representations implicitly encode morphological information which made them perform strongly for lemmatization.

However, we can see that for agglutinative languages such as Basque and Turkish, the neural models using explicit morphological features (Morpheus, Malaviya et al. 2019 and UDPipe) still outperform those without it (although for Basque the differences are much smaller). Still, the overall results show that, apart from Basque and Turkish, differences between XLM-RoBERTa and the best model for each language are rather minimal. This demonstrates that it is possible to generate competitive contextual lemmatization without any explicit morphological information using a very simple technique, although a more sophisticated approach or larger language model may be required to be competitive with the state-of-the-art currently represented by UDPipe.

1.6.2 Out-of-domain evaluation

Although lemmatizers are mostly used out-of-domain, the large majority of the experimental results published so far do not take this issue into account when evaluating approaches to contextual lemmatization. In this section we empirically investigate the out-of-domain performance of the lemmatizers from the previous section to establish whether: (i) using fine-grained morphological information causes cascading errors in the lemmatization performance; (ii) whether the lack of morphological information helps to obtain more robust lemmatizers across domains.

For a better comparison, Table 1.7 presents both the in-domain results presented in the previous section together with their corresponding out-of-domain performance on the datasets presented in Section 1.3.

Table 1.7 allows to see the general trend in performance across domains and with respect to the type of morphological information used. First, and as it could be expected, out-of-domain performance is substantially worse for every evaluation setting and particularly significant for highly-inflected languages. Second, in terms of the type of morphological label, there are no clear differences between the models using just UPOS tags or those using more fine-grained information, the exception being Russian and Turkish with the *ixa-pipe-mm* system, for which the highest result with {UPOS+Case+Number} is around 1 point in word accuracy better than UPOS.

		In-Domain			**Out-of-Domain**		
		ixa-mm	**ixa-gs**	**morpheus**	**ixa-mm**	**ixa-gs**	**morpheus**
NO TAG	en	-	96.34	97.51	-	90.40	92.47
	es	-	98.53	98.17	-	89.75	89.70
	ru	-	92.81	95.31	-	83.95	86.84
	eu	-	90.61	95.69	-	85.64	88.25
	cs	-	96.37	98.31	-	91.50	91.61
	tr	-	87.11	95.62	-	77.16	84.07
UPOS	en	**99.11**	98.91	98.10	**95.38**	95.25	92.92
	es	98.91	98.76	98.94	97.53	97.41	90.29
	ru	94.36	93.74	96.20	90.00	89.40	87.59
	eu	93.11	92.29	96.39	85.22	86.79	88.97
	cs	97.86	97.28	98.75	92.33	93.68	91.66
	tr	84.65	87.76	**96.44**	79.22	81.67	**84.96**
UCG	en	99.10	98.92	97.99	95.20	95.24	92.97
	es	98.94	98.70	98.98	97.54	97.43	90.31
	ru	94.85	93.30	96.21	90.97	89.33	87.67
	eu	92.65	92.39	96.34	85.23	86.74	89.09
	cs	97.29	96.64	**98.76**	91.61	91.35	91.92
	tr	85.09	87.09	96.18	80.06	81.23	84.74
		ixa-mm	**ixa-gs**	**morph**	**ixa-mm**	**ixa-gs**	**morpheus**
UCN	en	99.06	98.87	98.01	95.16	95.16	92.86
	es	98.92	98.75	**99.02**	97.56	97.44	90.35
	ru	95.07	93.70	96.20	**91.00**	89.60	87.58
	eu	93.03	92.35	96.39	85.47	86.36	89.03
	cs	97.44	96.87	98.71	91.04	92.07	**92.23**
	tr	85.52	87.18	96.11	80.33	81.00	84.40
UCGN	en	99.08	98.96	97.99	95.21	95.15	92.95
	es	98.89	98.71	98.97	**97.59**	97.44	90.38
	ru	95.00	93.08	**96.44**	90.80	89.13	87.66
	eu	93.03	92.28	96.39	85.38	86.55	88.86
	cs	97.17	96.68	98.70	91.71	91.50	91.97
	tr	85.52	87.18	96.20	80.33	81.00	84.46
UAllo	en	99.04	98.95	98.06	95.08	95.13	93.15
	es	98.86	98.74	99.00	97.54	97.45	90.34
	ru	94.75	93.22	96.30	90.88	88.66	87.57
	eu	93.41	94.06	**96.50**	85.33	86.31	**89.11**
	cs	97.03	96.63	98.70	91.19	91.81	92.02
	tr	84.90	86.57	96.22	79.39	80.50	84.96

Table 1.7: In-domain and out-of-domain test results for systems trained with explicit morphological information. Underline: Best model per language and type of label; **bold**: best overall per language.

Furthermore, there is not a common type of morphological information that works best across languages. Third, while the statistical lemmatizers are competitive for Spanish and English, they are clearly inferior for Basque and Turkish. Finally, when looking at the results in terms of the models using gold-standard morphological annotations (*ixa-pipe-gs* and Morpheus) it is interesting that they degrade less out-of-domain than the model using learned morphological tags for most of the cases except for Russian. Summarizing, we can conclude that adding fine-grained morphological information to UPOS does not in general result in better out-of-domain performance.

Following this, we would like to evaluate the out-of-domain performance when not even UPOS labels are used for training. From what we have seen in-domain, the systems that operate without morphology achieve competitive results with respect to the models using morphological information. Figure 1.3 provides an overview of both the in- and out-of-domain results obtained for both types of systems, confirming this trend. Thus, it is remarkable that the XLM-RoBERTa model scores best out-of-domain for Turkish and Czech, and a very close second in Russian. The results for Spanish and English deserve further analysis, as the IXA pipes statistical models clearly outperform every other system for these two languages, with the differences around 7 points in word accuracy.

Figure 1.4[6] presents the reversed results of those presented in Figure 1.3, namely, the test set of the in-domain corpora becomes the out-of-domain test data while the models are fine-tuned on the training split of the out-of-domain data. Doing this experiment allows to discard that the out-of-domain behaviour exhibited in previous results could be due to differences in size between the training in-domain data and the testing out-of-domain test sets. Good examples of this are Russian and Spanish for which SynTagRus and AnCora are used as in-domain data in the reversed setting. These two datasets are much larger than the GSD corpora for those languages (used as in-domain data in the original setting). Thus, results in the reversed setting demonstrate that: (i) out-of-domain performance worsens substantially regardless of the language and model, (ii) language models fine-tuned without explicit morphological information outperform in-domain every other model for all languages except Turkish, and (iii), the out-of-domain results of XLM-RoBERTa-base are the best for Russian and Czech and similar to other models in English and Spanish.

In any case, Figures 1.3 and 1.4 show that the results of every model significantly degrade when evaluated out-of-domain, the most common application of lemmatizers. Thus, even for high-scoring languages such as English and Spanish, out-of-domain performance worsens between 3 and 5 points in word accuracy. For high-inflected languages the differences are around 8 for Basque and more than 10 for Turkish.

[6] Basque is not present in this evaluation due to the fact that the Armiarma corpus does not include UniMorph annotations.

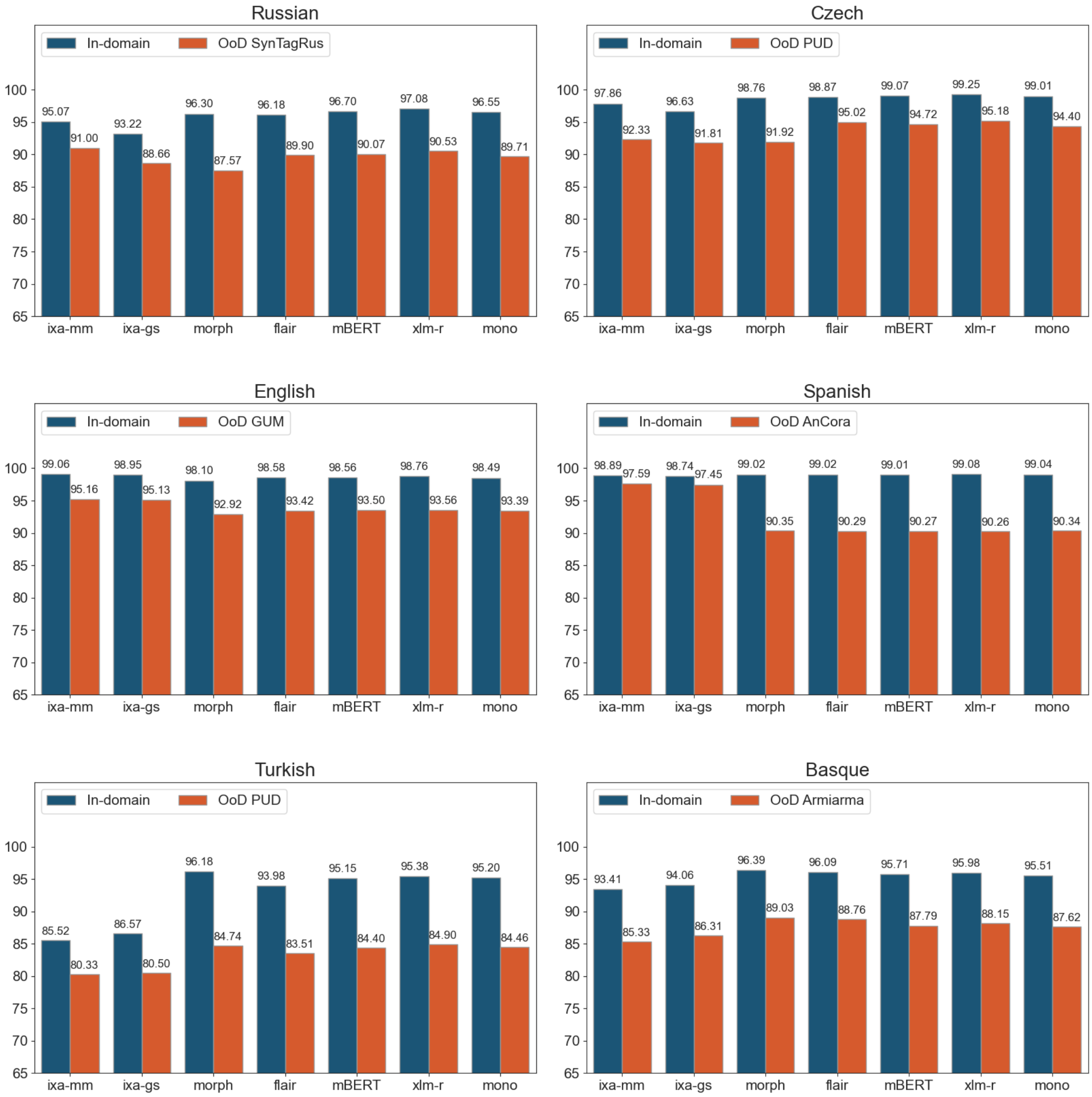


Figure 1.3: Overall in-domain and out-of-domain results.

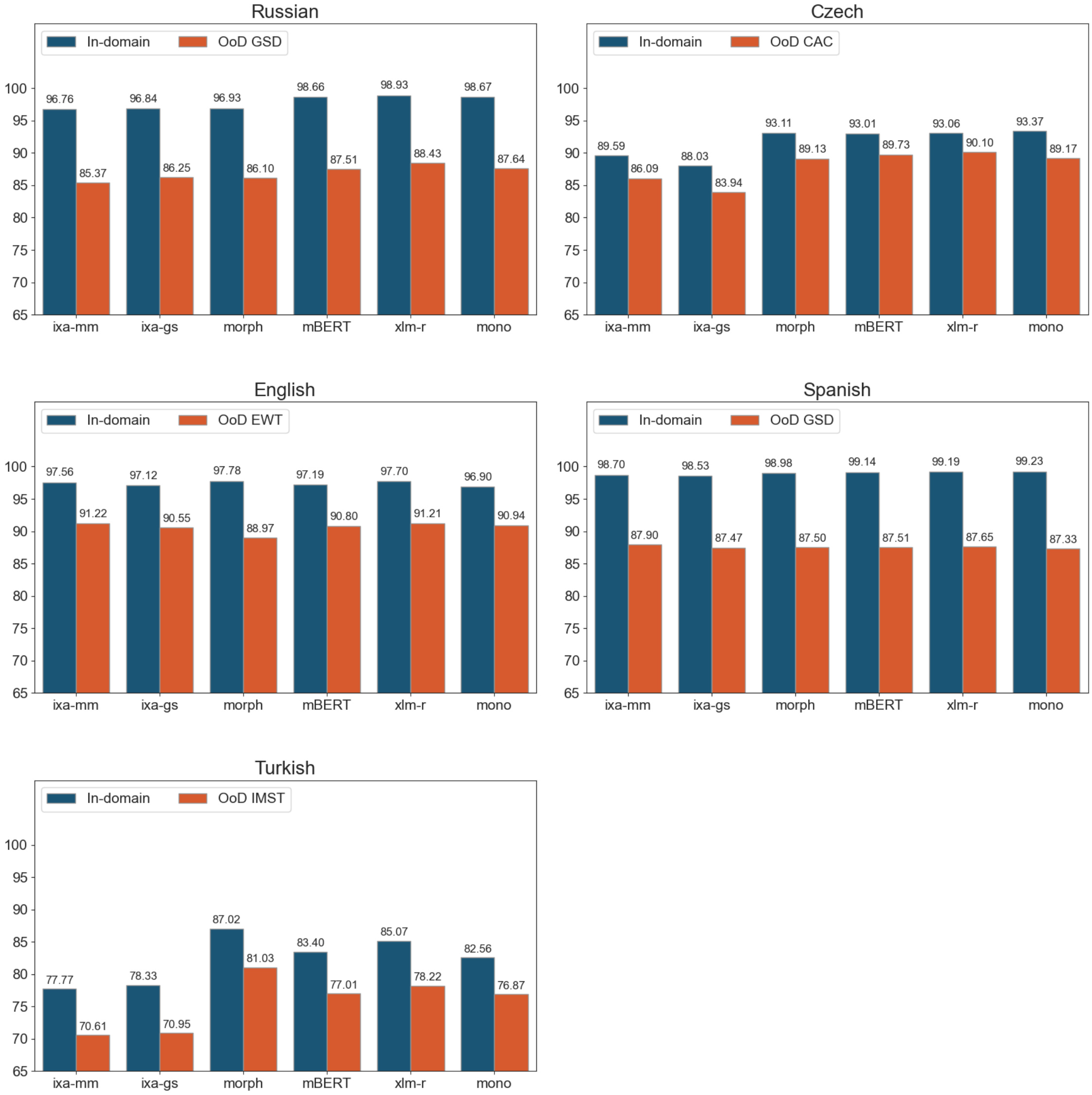


Figure 1.4: Overall in- and out-of-domain results in the reversed setting.

1.6.3 Is Explicit Morphology Required?

Given that pre-trained language models such as XLM-RoBERTa-base can be leveraged to learned competitive lemmatizers without using any explicit morphological signal, we propose a final experiment to address the following two additional research questions. First, will lemmatization results get closer to the state-of-the-art by using a larger Transformer-based model such as XLM-RoBERTa-large? Second, can we improve the performance of a language model such as XLM-RoBERTa by adding morphological information during fine-tuning?

	xlm-r-base				**xlm-r large**			
	in-domain		out-of-domain		in-domain		out-of-domain	
	without morph.	with morph.	without morph.	with morph.	without morph.	with morph.	without morph.	with morph.
en	98.76	98.74	93.56	93.72	98.85	**98.92**	93.82	**93.86**
es	99.08	99.10	90.26	90.42	99.12	**99.15**	90.48	**90.53**
eu	95.98	96.45	88.15	88.60	96.66	**96.70**	88.75	**88.81**
ru	97.08	97.25	90.53	90.92	97.63	**97.96**	91.60	**91.71**
cz	99.25	99.32	95.18	94.72	**99.40**	99.23	95.42	**96.06**
tr	95.38	95.19	84.90	85.34	**96.30**	96.13	85.18	**85.40**

Table 1.8: In- and out-of-domain results for XLM-RoBERTa-base and XLM-RoBERTa-large models with and without morphological features during training.

Table 1.8 shows the results of experimenting with XLM-RoBERTa-base and XLM-RoBERTa-large to learn lemmatization as a sequence labelling task with and without adding morphology as explicit handcrafted features. For each language we pick the best morphological configuration from Table 1.7 and encode the morphological labels as feature embeddings. Both feature and encoded text embeddings are then sent into a softmax layer for sequence labelling (Wang et al., 2022). The first observation is that the large version of XLM-RoBERTa obtains the best results both in- and out-of domain. It is particularly noteworthy that fine-tuning XLM-RoBERTa-large with only the SES classes helps to outperform any other model for every language and evaluation setting. Furthermore, adding morphology as a feature seems to be beneficial. In fact, the morphologically informed models are the best in 4 out of 6 in-domain evaluations and for all 6 out-of-domain cases.

We compute the McNemar test to establish whether the differences obtained with and without morphological features are actually statistically significant. It turns out that for XLM-RoBERTa-large results are rather mixed. Thus, only for Russian (p-value 0.003) and Czech (0.000) are the results significant at $\alpha = .05$. For Turkish and Basque the results are not conclusive (p-value 0.0495) while for the rest the null hypothesis cannot be rejected (0.423 for Spanish, 0.242 in English and 0.547 in Basque). Regarding

XLM-RoBERTa-base, in 4 out of 6 languages the results are statistically significant at $\alpha = .01$ (the McNemar test), failing to reject the null hypothesis for Russian and Turkish.

To sum up, our experiments empirically demonstrate that fine-grained morphological information to train contextual lemmatizers does not lead to substantially better in- or out-of-domain performance, not even for languages of varied complex morphology, such as Basque, Czech, Russian and Turkish. Thus, only for Basque and Turkish did Morpheus (using UPOS tags) outperformed XLM-RoBERTa models.

Taking this into account, and as previously hypothesized for other NLP tasks (Manning et al., 2020), modern contextual word representations seem to implicitly capture morphological information valuable to train lemmatizers without requiring any explicit morphological signal. We have proved this by training off-the-shelf language models to perform lemmatization as a token classification task obtaining state-of-the-art results for Russian and Czech, and very close performance to UDPipe in the rest. Finally, statistical models are only competitive to perform contextual lemmatization on languages with a morphology on the simple side of the complexity spectrum, such as English or Spanish.

Thus, the results indicate that XLM-RoBERTa-large is the optimal option to learn lemmatization without any explicit morphological signal for every language and evaluation setting.

1.7 Discussion

In this paper we performed a number of experiments to better understand the role of morphological information to learn contextual lemmatization. Our findings can be summarized as follows: (i) fine-grained morphological information does not help to substantially improve contextual lemmatization, not even for high-inflected languages; using UPOS tags seems to be enough for comparable performance; (ii) contextual word representations such as those employed in Transformer and Flair models seem to encode enough implicit morphological information to allow us to train good performing lemmatizers without any explicit morphological signal; (iii) the best-performing lemmatizers out-of-domain are those using either simple UPOS tags or no morphology at all; (iv) evaluating lemmatization on word accuracy is not the best strategy; results are too high and too similar to each other to be able to discriminate between models. By using word accuracy we are assigning the same importance to cases in which the lemma is equivalent to the word form (e.g. 'the') as to complex cases in which the word form includes case, number and/or gender information (e.g, 'medikuarenera', which in Basque means "to the doctor", with its corresponding lemma 'mediku'). We believe that this may lead to a high overestimation in the evaluation of the lemmatizers.

In this section, we address some remaining open issues with the aim of understanding better the main errors and difficulties still facing lemmatization. First, we discuss the convenience of using an alternative metric to word accuracy. Second, we analyze the performance of XLM-RoBERTa-base by evaluating accuracy per SES. Third, we examine the generalization capabilities of XLM-RoBERTa-base by computing word accuracy for in-vocabulary and out-of-vocabulary words. We also discuss any issues regarding test data contamination. Finally, we perform some error analysis on the out-of-domain performance of the XLM-RoBERTa-base model for Spanish, to see why it is different to the rest of the languages, as illustrated by Figure 1.3.

1.7.1 Sentence Accuracy

Looking at the in-domain results for lemmatization reported in the previous sections and in the majority of recent work (Malaviya et al., 2019; McCarthy et al., 2019; Yildiz and Tantuğ, 2019; Straka et al., 2019), with word accuracy in-domain scores around 96 or higher, it is not surprising to wonder whether contextual lemmatization is a solved task. However, if we look at the evaluation method a bit more closely, things are not as clear as they seem. As it has been argued for POS tagging (Manning, 2011), word accuracy as an evaluation measure is easy because you get many free points for punctuation marks and for the many tokens that are not ambiguous with respect to its lemma, namely, those cases in which the lemma and the word form are the same. Following this, a more realistic metric might consist of looking at the rate of getting the whole sentence correctly lemmatized, just as it was proposed for POS tagging (Manning, 2011).

Figure 1.5 reports the sentence accuracy of the six languages we used in our experiments both for in- and out-of-domain. In contrast to the word accuracies reported in Figure 1.3, we can see that the corresponding sentence accuracy results drop significantly. In addition to demonstrating that lemmatizers have a large margin of improvement, sentence accuracy allows us to better discriminate between different models. We can see this phenomenon in the English and Spanish results. Thus, while every model obtained very similar in-domain word accuracy in Spanish, using sentence accuracy helps to discriminate between the statistical and the neural lemmatizers. Furthermore, it also shows that among the neural models XLM-RoBERTa clearly outperforms the rest of the models by almost 1 percent.

The effect of sentence accuracy for the in-domain evaluation is vastly magnified when considering out-of-domain performance, with the extremely low scores across languages providing further evidence of how far lemmatization remains from being solved.

1.7.2 Analyzing word accuracy per SES

The next natural step in our analysis is identifying which specific cases are most difficult for lemmatizers. In order to do so, we look at the word accuracy for each of the SES labels automatically induced from the data. In order to illustrate this point, we took XLM-RoBERTa-base as an example use case and analyze their predictions for the languages which could be inspected in-house, namely, Basque, English, Spanish and

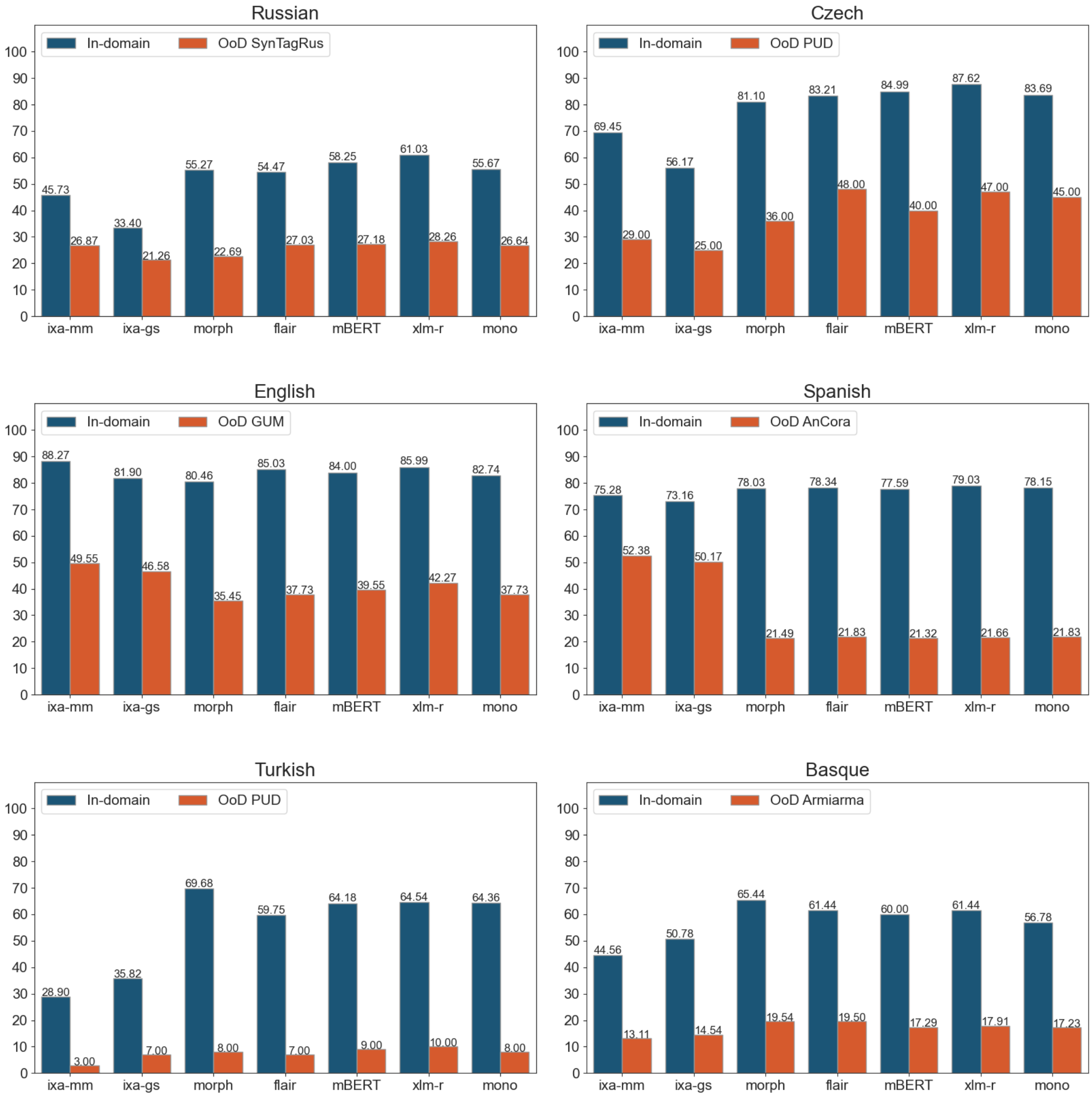


Figure 1.5: Sentence accuracy results for in- and out-of-domain settings.

	SES	**Casing**	**Edit script**	**W.acc**	**%**	**Examples**
en	↓0;d¦+	all low	do nothing	99.29	**76.87**%	positive→*positive*
	↑0¦↓1;d¦+	1st up	do nothing	96.29	6.97%	Martin→*Martin*
	↓0;d¦-+	all low	remove last ch	98.58	5.52%	things→*thing*
	↓0;abe	all low	ignore form, use *be*	99.81	2.02%	is→*be*
	↓0;d¦−+	all low	remove 2 last ch	97.42	1.52%	does→*do*
	↓0;d¦—+	all low	remove 3 last ch	96.45	1.10%	trying→*try*
	↑0¦↓-1;d¦+	all up	do nothing	94.22	0.68%	NASA→*NASA*
	↓0;d−+b¦+	all low	first 2 char to *b*	99.33	0.59%	are→*be*
	↓0;d¦-+v+e+	all low	last ch to *ve*	100.00	0.51%	has→*have*
	↓0;d¦—+e+	all low	3 last ch to *e*	96.23	0.42%	driving→*drive*
es	↓0;d¦+	all low	do nothing	99.36	**72.40**%	acuerdo→*acuerdo*
	↓0;d¦-+	all low	del last ch	97.22	5.29%	estrellass→*estrella*
	↓0;d+e¦-+	all low	add *e*, del last ch	96.73	3.36%	la→*el*
	↓0;d¦-+o+	all low	del last ch, add *o*	96.21	2.37%	una→*uno*
	↓0;d+e¦−+	all low	add *e*, del 2 last ch	99.78	2.13%	los→*el*
	↓0;d¦−+	all low	del 2 last ch	97.36	1.40%	flores→*flor*
	↓0;aél	all low	ignore form, use *él*	99.83	1.32%	se→*él*
	↓0;d¦+r+	all low	add *r*	100.00	0.91%	hace→*hacer*
	↓0;d¦+o+	all low	add *o*	97.73	0.91%	primer→*primero*
	↓0;d¦-+a+r+	all low	del last ch, add *ar*	98.07	0.83%	desarrolló→*desarollar*
ru	↓0;d¦+	all low	do nothing	99.16	**57.80**%	Петербург→*Петербург*
	↓0;d¦-+	all low	del last ch	97.67	6.97%	церковью→*церковь*
	↓0;d¦-+а+	all low	del last ch, add *а*	96.65	3.32%	экономику→*экономика*
	↓0;d¦-+й+	all low	del last ch, add *й*	96.08	3.10%	городское→*городской*
	↓0;d¦−+	all low	del 2 last ch	99.03	2.10%	странами→*страна*
	↓0;d¦-+е+	all low	del last ch, add *е*	98.04	2.07%	моря→*море*
	↓0;d¦-+я+	all low	del last ch, add *я*	97.83	1.86%	историю→*история*
	↓0;d¦-+т+ь+	all low	del last ch, add *ть*	98.88	1.81%	получил→*получить*
	↓0;d¦-+ь+	all low	del last ch, add *ь*	93.94	1.67%	сентября→*сентябрь*
	↓0;d¦−+т+ь+	all low	del 2 last, add *ть*	98.10	1.60%	были→*быть*
eu	↓0;d¦+	all low	do nothing	99.05	**49.63**%	sartu→*sartu*
	↓0;d¦−+	all low	remove 2 last ch	97.72	9.93%	librean→*libre*
	↓0;d¦-+	all low	remove last ch	96.27	6.54%	korrikan→*korrika*
	↓0;d¦—+	all low	remove 3 last ch	93.24	3.60%	aldaketarik→*aldaketa*
	↑0¦↓1;d¦+	1st up	do nothing	98.54	3.46%	MAPEI→*Mapei*
	↓0;d¦—-+	all low	del 4 last ch	93.00	2.52%	lagunaren→*lagun*
	↑0¦↓1;d¦−+	1st up	del 2 last ch	95.54	1.88%	Egiptora→*Egipto*
	↓0;d-+i+z ¦+n+	all low	del 1st ch, add *iz,n*	100.00	1.38%	da→*izan*
	↑0¦↓1;d¦-+	1st up	del last ch	90.08	1.10%	Frantziak→*Frantzia*

Table 1.9: 10 most frequent SES, brief description, corresponding word accuracy, weight (in %) in the corpus and examples of words and their lemmas for English, Spanish, Russian and Basque; SES are computed following UDPipe's method (Straka et al., 2019).

.

Russian. Thus, Table 1.9 presents examples and results for the 10 most frequent SES for each of these 4 languages development sets.

As we can see in Table 1.9, the most common lemma transformation to be learned is based on the edit script "do nothing", namely, the lemmatizer needs to learn that the lemma and the word have the same form. It is also interesting to see how the ratio of such lemma type changes across languages, from English, where such cases are observed in almost 77% of the cases to Basque, where only half of the lemmas correspond to this rule. However, in terms of word accuracy, the results are remarkably similar for all 4 languages, in the range of 99-99.30%. This demonstrates that the traditional evaluation method greatly overestimates the lemmatizers' performance.

By looking at other specific cases, we can see that in English problematic examples to learn are those related to the casing of some characters (e.g. Martin → *Martin*, NASA → *NASA*). Other noticeable issue refers to the verbs in gerund form (e.g. try**ing** → *try*, driv**ing** → *drive*).

With respect to Spanish interesting difficult lemmas are observed with articles in feminine form (e.g. la → *el*, una → *uno*), where the masculine form is considered the canonical form or lemma, and feminine articles and adjectives should be lemmatized by changing the gender of the word from female to male.

In Russian the most challenging case corresponds to the lemmatization of the nouns that end with a soft sign **ь** with the word accuracy for this SES as low as 93.94%. The possible reason of such low accuracy could be the absence of a specific grammar rule that defines the gender of such nouns and, therefore, the termination these nouns have in different cases. The second lowest accuracy among the 10 most popular SES in Russian is for adjectives, cases in which to obtain the lemma one should delete the last character of the word and add a letter **й** (pronounced as *iy kratkoe*, short y), that in Russian determines the suffix for some masculine nouns and adjectives in singular and nominative case. The words could be in different cases and genders, so it is necessary to know such information for correct lemmatization (e.g. городск**ое** → городско**й** (neutral gender, nominative case), семейным → семейны**й** (masculine gender, instrumental case)).

Finally, for Basque the most problematic cases with a rather low word accuracy of only 90.08% can be found among the nouns in ergative (e.g. Frantzia**k** → *Frantzia*) or locative cases (e.g. Mosku**n** (in Moscow) → Mosku, Katalunia**n** (in Catalonia) → Katalunia). The other two most difficult SES occur when the word forms are in possessive case (e.g. lagun**aren** → lagun) and for nouns in indefinite form (e.g., aldaketa**rik** (change) → aldaketa).

It should be noted that an extra obstacle to improving some of these difficult cases is the low number of samples available. Nonetheless, this analysis shows that lemmatizers still do not properly learn to lemmatize relatively common word forms.

1.7.3 Generalization Capabilities of Language Models

In this subsection we aim to analyze the generalization capabilities of a MLM such as XLM-RoBERTa-base in the lemmatization task. More specifically, we will discuss two issues: (i) whether MLMs simply memorize the SES lemma classes during fine-tuning and (ii) whether the good performance of MLMs in this task might be due to some test data contamination.[7]

In order to address the first point, we evaluate the performance of XLM-RoBERTa-base, fine-tuned without morphological features, for those words seen during fine-tuning (in-vocabulary words) with respect to out-of-vocabulary occurrences.

	in-domain		**out-of-domain**	
	in-vocabulary	**out-of-vocabulary**	**in-vocabulary**	**out-of-vocabulary**
en	99.20	90.60	95.25	81.11
es	99.23	93.71	92.69	59.36
eu	98.29	83.18	91.65	74.96
ru	99.29	90.39	95.04	79.07
cz	99.31	93.33	98.80	82.31
tr	98.96	84.55	94.37	68.59

Table 1.10: Word accuracy for in-vocabulary and out-of-vocabulary words for XLM-RoBERTa-base model (original setting). Corpora: English - EWT (in-domain), GUM (out-of-domain); Spanish - GSD (in-domain), AnCora (out-of-domain); Basque - BDT (in-domain), Armiarma (out-of-domain); Russian - GSD (in-domain), SynTagRus (out-of-domain); Czech - CAC (in-domain), PUD (out-of-domain); Turkish - IMST (in-domain), PUD (out-of-domain).

	in-domain		**out-of-domain**	
	in-vocabulary	**out-of-vocabulary**	**in-vocabulary**	**out-of-vocabulary**
en	98.57	89.15	93.95	76.48
es	99.34	93.28	92.12	49.21
ru	99.25	92.28	93.78	66.22
cz	98.17	83.58	96.31	81.28
tr	93.68	70.68	95.41	59.23

Table 1.11: Word accuracy for in-vocabulary and out-of-vocabulary words for XLM-RoBERTa-base model (reversed setting).

Tables 1.10 and 1.11 report the results for both original and reversed settings and in- and out-of-domain evaluations. It is noticeable that the model performs very well

[7] https://hitz-zentroa.github.io/lm-contamination/

on out-of-vocabulary words, also in the out-of-domain evaluation, which would seem to indicate that XLM-RoBERTa is generalizing beyond the words seen during training. This seems to be confirmed also by looking at the Spanish and Russian results. It should be remembered that, while in the reversed setting the training data for Spanish (AnCora, 500K tokens) and Russian (SynTagRus, 900K words) is much larger than in the original setting (both GSD), the obtained results reflect roughly the same trend.

Finally, we should consider whether a MLM such as XLM-RoBERTa has already seen the datasets we are experimenting with during pre-training, namely, whether XLM-RoBERTa has been contaminated.[8] First, it should be noted that CC-100, the corpus used to generate XLM-RoBERTa, was constructed by processing the CommonCrawl snapshots from between January and December 2018. Second, the SIGMORPHON data we are using was released in 2019[9] with the test data including gold standard lemma and UniMorph annotations being released in April 2019. Third and most importantly, XLM-RoBERTa does not see the lemmas themselves during training or inference, but the SES classes we automatically generate in an ad-hoc manner for the experimentation. The datasets containing both the words and the SES classes used have not been yet made publicly available.

Based on this, it is possible to say that XLM-RoBERTa seems to generalize over unseen words and that its performance is not justified by any form of language model contamination.

1.7.4 Analyzing Spanish out-of-domain results

In Section 1.6.2 we saw that out-of-domain performance of Transformer-based models for Spanish was not following the pattern of the rest of the languages. Instead, they were 6-7% worse than the results obtained by the IXA pipes statistical lemmatizers (*ixa-pipe-mm* and *ixa-pipe-gs*). By checking the most common error patterns of XLM-RoBERTa-base, we found out that most of the performance loss was caused by inconsistencies in the manual annotation of lemmas between the data used for in-domain and out-of-domain evaluation. More specifically, the GSD Spanish corpus included in UniMorph wrongly annotates lemmas for proper names such as Madrid, London or Paris entirely in lowercase, namely, *madrid*, *london* and *paris*. However, the AnCora Spanish corpus used for out-of-domain evaluation correctly annotates these cases specifying their corresponding lemmas with the first character in uppercase. This inconsistency results in 3781 examples of proper names in the AnCora test set which are all lemmatized following the pattern seen during training with the GSD training set. Consequently, the word accuracy obtained by the model for this type of examples in the AnCora test set is 0%. In order to confirm this issue, we corrected the wrongly annotated proper names in the GSD training data, fine-tuned again the model and saw the out-of-domain

[8] https://hitz-zentroa.github.io/lm-contamination/blog/

[9] First GitHub commit December 19, 2018.

performance of XLM-RoBERTa-base go up from 90.26% to 96.75%, a more consistent result with respect to the out-of-domain scores for the other 5 languages.

This issue manifests the importance of consistent manual annotation across corpora from different domains in order to fairly evaluate out-of-domain performance of contextual lemmatizers.

1.8 Concluding Remarks

Lemmatization remains an important Natural Language Processing task, especially for languages with high-inflected morphology. In this paper we provide an in-depth study on the role of morphological information to learn contextual lemmatizers. By taking a language sample of varied morphological complexity, we have analyzed whether a fine-grained morphological signal is indeed beneficial for contextual lemmatization. Furthermore, and in contrast to previous work, we have also evaluated lemmatizers in an out-of-domain setting, which constitutes, after all, their most common application use. Our results empirically demonstrate that informing lemmatizers with fine-grained morphological features during training is not that beneficial, not even for agglutinative languages. In fact, modern contextual word representations seem to implicitly encode enough morphological information to obtain good contextual lemmatizers without seeing any explicit morphological signal. Finally, good out-of-domain performance can be achieved using simple UPOS tags or without any explicit morphological signal.

Therefore, our results suggest that an optimal solution among all the options considered would be to develop lemmatizers by fine-tuning a large MLM such as XLM-RoBERTa-large without any explicit morphological signal. Addressing lemmatization as a token classification task results in highly competitive and robust lemmatizers with results over or close to the state-of-the-art obtained with much more complex methods (Straka et al., 2019).

Furthermore, we have discussed current evaluation practices for lemmatization, showing that using simple word accuracy is not adequate to clearly discriminate between models, as it provides a deceptive view regarding the performance of lemmatizers. An additional analysis looking at specific lemma classes (SES) has shown that many common word forms are still not properly predicted. The conclusion is that lemmatization remains a challenging task. Future work is therefore needed to improve out-of-domain results. Furthermore, it is perhaps a good time to propose an alternative word-level metric to evaluate lemmatization that, complemented with sentence accuracy, may provide a more realistic view of the performance of contextual lemmatizers.

CHAPTER 2

Shortest Edit Scripts Methods for Contextual Lemmatization

This chapter is based on the following publication:

Olia Toporkov and **Rodrigo Agerri** (2024). Evaluating Shortest Edit Script Methods for Contextual Lemmatization. In *Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024).*

Abstract: Modern contextual lemmatizers often rely on automatically induced Shortest Edit Scripts (SES), namely, the number of edit operations to transform a word form into its lemma. In fact, different methods of computing SES have been proposed as an integral component in the architecture of several state-of-the-art contextual lemmatizers currently available. However, previous work has not investigated the direct impact of SES in the final lemmatization performance. In this paper we address this issue by focusing on lemmatization as a token classification task where the only input that the model receives is the word-label pairs in context, where the labels correspond to previously induced SES. Thus, by modifying in our lemmatization system only the SES labels that the model needs to learn, we may then objectively conclude which SES representation produces the best lemmatization results. We experiment with seven languages of different morphological complexity, namely, English, Spanish, Basque, Russian, Czech, Turkish and Polish, using multilingual and language-specific pre-trained masked language encoder-only models as a backbone to build our lemmatizers. Comprehensive experimental results, both in- and out-of-domain, indicate that computing the casing and edit operations separately is beneficial overall, but much more clearly for languages with high-inflected morphology. Notably, multilingual pre-trained language models consistently outperform their language-specific counterparts in every evaluation setting.

2.1 Introduction

Lemmatization is one of the most common basic Natural Language Processing (NLP) tasks, commonly understood as transforming an inflected wordform (e.g., *feeling, felt*) into its initial form known as lemma (e.g., *feel*), as defined by the contextual lemmatization SIGMORPHON 2019 shared task (Aiken et al., 2019).

Lemmatization remains important for morphologically-rich languages as it usually plays a crucial role for information extraction systems, sentiment analysis and helps to deal with inflected named entities during named entity recognition task, especially for high-inflected languages.

Nowadays, state-of-the-art approaches to lemmatization are based on supervised contextual methods, a technique first proposed by Chrupala et al. (2008). Treating lemmatization as a supervised classification task relies on automatically inducing a set of patterns from textual corpora, encoding the minimum amount of edits needed to map the surface word to its lemma, namely, the Shortest Edit Script (SES). Ideally these SES would capture morphological patterns about word inflection making lemmatization feasible as a classification task. Thus, in Chrupala's approach, classifiers would learn previously induced SES which, at inference time, would be decoded back into their lemmas.

Modern contextual lemmatizers often rely on automatically induced Shortest Edit Scripts (SES) for optimal performance. In fact, different methods of computing SES have been proposed as an integral component in the architecture of several state-of-the-art contextual lemmatizers currently available (Malaviya et al., 2019; Straka et al., 2019; Yildiz and Tantuğ, 2019). However, previous work has not investigated the direct impact of SES in the final lemmatization performance. In order to address this issue, in this paper we compare three popular approaches to automatically induce SES (Straka et al., 2019; Yildiz and Tantuğ, 2019; Agerri et al., 2014; Agerri and Rigau, 2016) and empirically investigate which of them (if any) is the most beneficial.

In order to do so, we follow previous work by Toporkov and Agerri (2024) which demonstrates that Masked Language Models (MLMs) can competitively perform contextual lemmatization without receiving any explicit morphological signal during training, using just the word form and its corresponding SES. This allows us to focus on lemmatization as a token classification task where the only input that the model receives is the word-label pairs in context, in other words, the labels corresponding to previously induced SES. Thus, by modifying in our lemmatization systems only the SES labels that the model needs to learn, we may then be able to objectively conclude which SES representation helps to produce the best lemmatization results.

For our experiments we pick seven languages of different morphological complexity, namely, English, Spanish, Basque, Russian, Czech, Turkish and Polish. Moreover, we use a number of multilingual and language-specific pre-trained MLMs as backbone to build our lemmatizers. To the best of our knowledge, this is the first systematic evaluation of the impact of the SES representations for contextual lemmatization.

Comprehensive experimental results, both in- and out-of-domain, indicate that computing the casing and edit operations separately, as proposed by UDPipe, is the best method to obtain SES overall, particularly for the languages with more complex morphology. Chrupala's approach as implemented by Agerri et al. (2014) performs as a close second, while the Morpheus method (Yildiz and Tantuğ, 2019) is the less optimal one. In addition, our results show that multilingual MLMs consistently outperform their language-specific counterparts in every evaluation setting. This is consistent with previous research comparing monolingual and multilingual encoder-only models (Agerri and Agirre, 2023). Furthermore, our experimental setting shows how to easily obtain competitive lemmatization results for the languages of our choice.

Code, data and fine-tuned models are publicly available to facilitate further research on this topic and reproducibility of the results.[1]

2.2 Related Work

Attempts to resolve the lemmatization task started with systems based on dictionary lookup and/or finite set of rules (Karttunen et al., 1992; Oflazer, 1993; Alegria et al., 1996; Segalovich, 2003; Carreras et al., 2004; Stroppa and Yvon, 2005). These systems, apart from being language dependent, required a lot of effort, linguistic knowledge and manual intervention, especially for more complex languages with a high level of inflection. The creation of large annotated corpora, which included morpho-syntactic features and lemmas, led to the development of machine learning approaches to lemmatization in a variety of languages. Thus, initiatives such as the Universal Dependencies (Nivre et al., 2017) and the UniMorph project (McCarthy et al., 2020) allowed to gather annotated corpora in more than 118 languages, including low-resourced and endangered ones.

The hypothesis that context is beneficial in the case of unseen and ambiguous words incentivized the appearance of supervised contextual lemmatizers. One of the pioneer works in this field is the statistical contextual lemmatizer Morfette (Chrupala et al., 2008). It is based on a pipeline approach and uses a Maximum Entropy classifier to predict morphological tags and lemmas. Crucially, Chrupala et al. (2008) presents for the first time the idea of treating lemmatization as a classification task by predicting the Shortest Edit Script (SES), namely, the shortest sequence of instructions (insertions, deletions or replacements) needed to transform a reversed inflected word to its lemma. The work of Chrupala et al. (2008) inspired the development of many methods for contextual lemmatization, which most of the time included the idea of using minimum edit scripts. Among others, the IXA pipes system (Agerri et al., 2014; Agerri and Rigau, 2016) and Lemming (Müller et al., 2015) apply the same principle of edit trees, combining it with the possibility of adding external lexical information. Other examples of the systems that use the concept of SES are the works of Gesmundo and Samardžić (2012), Chakrabarty et al. (2017) and the system of Malaviya et al. (2019).

[1] `https://github.com/hitz-zentroa/ses-lemma`

The development of supervised approaches involving deep learning algorithms and the appearance of the Transformer architecture (Vaswani et al., 2017) and Transformer-based MLMs such as BERT (Devlin et al., 2019) and XLM-RoBERTa (Conneau et al., 2020) allowed to significantly improve the performance of supervised lemmatizers. Thus, in the SIGMORPHON 2019 shared task on contextual lemmatization (McCarthy et al., 2019) most of the participating systems were based on MLMs. The best overall system was UDPipe (Straka et al., 2019), which ensembled various pre-trained contextualized BERT and Flair embeddings as an additional input to a Bi-LSTM network. To perform lemmatization they classify the input words according to the set of generated lemma rules or SES. The third best model, Morpheus, proposed a two-level LSTM network (Yildiz and Tantuğ, 2019) which used vector-based representations of words, morphological tags and SES as input. The output of the system results in a corresponding morphological labels and SES representing the lemma class which is later decoded into its lemma form.

However, while many of these top performing systems included different methods to compute SES as an integral component in their lemmatization models, there has not been an attempt to compare and establish which of the existing methods is the optimal one for the task. In this paper we pick three of the most popular SES approaches (according to performance on the SIGMORPHON 2019 lemmatization benchmark) and make a systematic comparison with the aim of clarifying this issue. We believe that this could benefit the development of future lemmatizers which may include SES as an integral component of their systems.

2.3 Data

To train and evaluate our models we used the datasets developed for the SIGMORPHON 2019 shared task on contextual lemmatization (McCarthy et al., 2019). These datasets are annotated according to the Unimorph schema guidelines (McCarthy et al., 2020). For in-domain evaluation we chose one corpus per language with standard train and development partitions. Additionally, we also provide out-of-domain evaluation results, as this is the setting in which lemmatizers are usually deployed. As most languages are represented in the SIGMORPHON 2019 by more than one dataset, for out-of-domain evaluation we picked the test sets of datasets different from the ones selected for in-domain evaluation. The exception was Basque, for which we selected a dataset external to the UniMorph SIGMORPHON data.

With respect to in-domain, in the case of Spanish and Russian we used GSD data, which consists of Wikipedia and news articles, texts from blogs and reviews. As the lemmas of these two corpora were originally lower-cased and giving the fact that the methods of generating the Shortest Edit Scripts (SES) are case dependent (Toporkov and Agerri, 2024), we performed a simple adjustment by changing the lemmas of the proper nouns to their upper-cased version. For the rest of the languages, there was no need of performing such adjustment, as the lemmas for the proper nouns in the corresponding corpora were correctly upper-cased by default.

For Polish we chose the LFG corpus (Przepiórkowski and Patejuk, 2018), derived from a corpus of LFG (Lexical Functional Grammar) syntactic structures, and consisting mostly of sentences from fiction, news and non-fiction genres, as well as the texts from the Internet sources. For out-of-domain we used PUD.

The datasets for English, Basque, Turkish and Czech were the same as in the previous paper, described in Section 1.3.2. Table 2.1 provides the details about the size of the datasets used for training, development and evaluation, both in- and out-of-domain.

	train	**dev**	**test**	**test(OOD)**
es	345,545	42,545	43,497	54,449
ru	79,989	9,526	9,874	109,855
en	204,857	24,470	25,527	8,189
eu	97,336	12,206	11,901	299,206
tr	46,417	5,708	5,734	1,795
cz	395,043	50,087	49,253	1,930
pl	104,730	13,161	13,076	8,511

Table 2.1: Number of tokens in the train, development, in-domain (test) and out-of-domain (test(OOD)) test sets.

2.4 Methods to Induce Shortest Edit Scripts

The general idea of computing the Shortest Edit Script (SES) in contextual lemmatization is based on finding the minimum number of edit operations necessary to convert a surface word into its corresponding lemma. By edit operations we understand any change applied to the wordform, which consists in deleting, inserting or replacing letters in the surface word, as well as leaving the word unchanged in the case the inflected form and the lemma coincide (e.g. the→*the*, road→*road*). SES methods focus on codifying such minimum edits for their further application as a set of instructions to modify the surface word. In this paper we address three different approaches based on the Shortest Edit Scripts. The methods chosen are those implemented by the first and third best systems in the SIGMORPHON 2019 shared task, namely UDPipe (Straka et al., 2019) and Morpheus (Yildiz and Tantuğ, 2019), and Chrupala's original proposal as implemented by the IXA pipes system (Agerri et al., 2014; Agerri and Rigau, 2016). [2]

[2] It may be argued that the methods of Morpheus and UDPipe systems do not strictly always generate the *shortest* edit script (SES). However, we keep the SES term as it was originally coined by Chrupala et al. (2008) as a convenient acronym.

2.4.1 UDPipe

The approach applied in the UDPipe system focuses on performing character level edits on the suffixes and prefixes of the word. They divide their script creation in two parts: (i) encoding the correct casing as a casing script and (ii) creating a sequence of character edits. Regarding the casing script, they consider both wordform and lemma as lower-cased. If the lemma contains upper-cased characters they add a rule to the casing script to uppercase such characters in the final lemma. The next step is creating a sequence of character edits by splitting the wordform into a prefix, a root (stem) and a suffix in order to process them separately. The root is obtained by finding the longest equal substring between the input word and its lemma and is kept unchanged. Then they process the prefixes and suffixes of the target word, including possible character operations such as *copy*, *add* or *delete*. The final script is produced by a concatenation of the casing and the edit scripts. The obtained SES are the complete rules which convert input words to their lemmas. When the word and lemma do not share any common parts, the word is considered irregular and is directly replaced by its lemma, skipping any possible edits.

2.4.2 Morpheus

Morpheus's approach is based on the prediction of minimum edits between a surface word and its lemma using four fundamental operations such as *same*, *delete*, *replace* and *insert*. *Same* and *delete* operations have only one version (the character may be left without changes (s) or deleted (d)), while *replace* and *insert* operations may vary, depending on the character they are tied to. As the minimum edit prediction decoder of Morpheus creates edit labels for each character in the word, it is only able to generate lemmas shorter or equal to the inflected forms. Still, in some languages lemmas may be longer that their corresponding wordforms. For such cases Yildiz and Tantuğ (2019) modify the standard Levenshtein distance algorithm by merging successive *insert* labels into one in the same position with multiple characters. They perform the same process for the *replace* label, combining it with the successive *insert* labels into one *replace* label and ensuring the correct lemma generation. They also consider the cases where the word is situated in the beginning of the sentence and should be lowercased, reflecting it in the Shortest Edit Script.

2.4.3 IXA pipes

The third method is based on the interpretation of Chrupala's technique (Chrupala et al., 2008) by Agerri et al. (2014). This approach addresses the suffixal nature of inflectional morphology where the end of the word is the most changing part and is more likely subject to modifications than the prefix or root. Chrupala et al. (2008) propose to compute the minimum edit distance between the *reversed* wordform and its lemma. They index word characters starting from the end of the string, allowing to form more coherent lemma classes and to perform lemmatization more efficiently. In the set of instructions that are generated using this technique the position of the letters that are subject to change are indicated along with the type of operation (insertion or deletion). In the adaptation of this approach it is also considered the casing of proper nouns, as well as the casing of the words that appear in the beginning of the sentence and should be lowercased for their correct lemmatization.

2.4.4 SES Comparison

In order to obtain a better understanding of the described methods and their core differences, we provide a brief comparison of the three minimum edit approaches, namely, UDPipe system's approach (*ses-udpipe*), Morpheus's approach (*ses-morpheus*) and IXA pipes approach (*ses-ixapipes*).

word→*lemma*	**ses-udpipe**	**ses-ixapipes**	**ses-morpheus**
cats→cat	↓0;d¦-	D0s	s|s|s|d
birds→*bird*	↓0;d¦-	D0s	s|s|s|s|d
did→*do*	↓0;d¦−+o	R1ioD0d	s|r_o|d
Wolak→*Wolak*	↑0¦↓1;d¦	O	s|s|s|s|s
You→*you*	↓0;d¦	1	l|s|s

Table 2.2: Examples of the three types of SES patterns: UDPipe - ses-udpipe, IXA pipes - ses-ixapipes and Morpheus - ses-morpheus.

Table 2.2 provides some examples of the Shortest Edit Scripts used in lemmatization for the aforementioned SES methods. For an action such as removing the last letter of the surface word (as in the case of the words 'cats' and 'birds') both *ses-udpipe* and *ses-ixapipes* apply the edit instruction to the reversed wordform, removing the last letter. Additionally, *ses-udpipe* method indicates that the word has to be lowercased. As for *ses-morpheus*, it processes each letter separately, leaving those that should remain untouched as 's' (same) and deleting the last one, marking such operation with 'd' (delete). Unlike *ses-udpipe* and *ses-ixapipes*, the scripts corresponding to the same action of deleting the last word's letter generate two different label classes as the number of the letters in 'cats' and 'birds' is distinct.

The next lemmatization example (did→*do*), demonstrates how each of the SES approaches treats the cases where one or more letters should be inserted in order to obtain the lemma. Here *ses-ixapipes* and *ses-morpheus* methods apply similar order of minimum edits using delete and replace operations, while *ses-udpipe* first deletes the two ultimate letters of the word and only then makes the insertion of the letter 'o'.

Finally, the last two examples are provided in order to reflect the edit scripts that are generated in the case of proper nouns in contrast to the ordinary nouns situated in the beginning of the sentence and, therefore, starting with the capital letter. We could see that for the proper noun 'Wolak' *ses-udpipe* indicates that the first letter should remain uppercased, whereas the scripts of *ses-ixapipes* and *ses-morpheus* simply leave the word unchanged. As for the pronoun 'You' situated in the beginning of the sentence, all three SES approaches lowercase it in order to obtain the correct lemma. It is important to mention that, as with the first two examples, in the case of the longer proper nouns the UDPipe's and IXA pipes' scripts would remain the same, while the script of the Morpheus's approach would vary according to the number of letters in the surface word.

2.5 Systems

In our experiments we apply two multilingual and seven language-specific pre-trained masked language models (MLMs). With respect to multilingual models we use multilingual BERT (mBERT) (Devlin et al., 2019), a Transformer-based masked language model pre-trained on the Wikipedias of 104 languages. mBERT was pre-trained using both masking and next sentence prediction objectives, applying a batch size of 256 and 512 sequence length for 1M steps. The second multilingual model we apply is XLM-RoBERTa (Conneau et al., 2020), pre-trained on 2.5TB of filtered CommonCrawl data for 100 languages. This model is based on the RoBERTa architecture, was trained only on the MLM task, implies dynamic mask generation and was pre-trained over 1.5M steps with a batch of 8192 and sequences of 512 length. We used both base and large versions of XLM-RoBERTa.

Regarding the language-specific models, we choose one model for each of the target languages. For Spanish we use the cased version of BETO (Cañete et al., 2020). It is a BERT-base language model trained on a large Spanish corpus including all Spanish Wikipedia as well as the Spanish part of the OPUS project (Tiedemann, 2012) in a total size of around 3 billion words. BETO is an upgraded version of the initial BERT-base model with the application of the dynamic masking technique, introduced in RoBERTa. It was trained with the total of 2M steps in two stages: 900K steps with a batch size of 2048 and maximum sequence length of 128, and the rest of the steps using batch size of 256 and maximum sequence length of 512.

For the Czech language we apply slavicBERT (Arkhipov et al., 2019), developed by continuing the training of multilingual BERT on Russian news and the Wikipedias of Russian, Bulgarian, Czech and Polish. The vocabulary of subword tokens was also rebuilt with the use of the subword-nmt repository.[3]

For Russian we choose RuBERT (Kuratov and Arkhipov, 2019), which was developed similarly to slavicBERT, with the difference of having only Russian as the target language. The system was trained using the Russian Wikipedia and news. The authors obtain a new subword vocabulary with longer Russian words and subwords from subword-nmt.

In the case of English we train RoBERTa-base (Liu et al., 2019), an optimized version of the BERT model. This model was obtained using more than 160GB of uncompressed

[3] https://github.com/rsennrich/subword-nmt/

text, including, apart from the standard BERT datasets, the CC-news dataset with English news articles published between January 2017 and December 2019.

For the Polish language we apply the base version of HerBERT (Mroczkowski et al., 2021). This model is based on the original BERT architecture and achieves state-of-the-art results on several downstream tasks, obtaining the best overall scores for Polish language understanding on the KLEJ Benchmark. HerBERT was trained on two datasets merged from six corpora such as NKJP, Wikipedia, Wolne Lektury, CCNet and Open Subtitles. Its base version outperformed the base version of Polish RoBERTa despite being trained with a smaller batch size (2560 vs 8000) and for a fewer number of steps (50k vs 125k).

In the case of Turkish we use BERTurk.[4] It is a cased BERT-base model, trained on 35GB of data, including Wikipedia, various OPUS corpora (Tiedemann, 2016), data provided by Kemal Oflazer and the version of the Turkish OSCAR corpus (Ortiz Suárez et al., 2019) which was previously filtered and sentence segmented.

Finally, for Basque we use BERTeus (Agerri et al., 2020), a BERT-base model trained on the BMC Basque corpus, which consists of news articles from online newspapers and the Basque Wikipedia. The authors also perform the subword tokenization, which is closer to linguistically interpretable strings in Basque. BERTeus outperforms mBERT and XLM-RoBERTa in several NLP tasks including named entity recognition, POS tagging, sentiment analysis and topic modelling.

2.6 Experimental Setup

In order to compare the three different approaches to generate the Shortest Edit Scripts (SES) described in Section 2.4, we fine-tuned the multilingual and language-specific pre-trained masked language models for each language in a token classification task, where the labels to be predicted correspond to the automatically induced SES. The MLMs were fine-tuned by adding a single linear classification layer on top. We performed a grid search of hyperparameters to select the best batch size (8, 16), weight decay (0.01, 0.1), learning rate (2e-5, 3e-5, 5e-5) and epochs (5, 10, 15, 20). We conduct both in-domain and out-of-domain evaluation of the models. By out-of-domain evaluation we understand evaluating on a data distribution different from the one that was used for training (in the in-domain setting). For each type of SES we chose the best model on the development set among the four MLMs in terms of word accuracy and loss. For all the languages the highest accuracy was achieved using XLM-RoBERTa-large model, being the only exception the *ses-morpheus* method in the case of Russian, where the best accuracy was achieved using mBERT. Thus, every result reported in the next subsections is obtained using XLM-RoBERTa-large as a backbone. Finally, apart from calculating word and sentence accuracy scores, we also report the statistical significance across the three SES methods using the McNemar test (Dietterich, 1998).

[4] https://github.com/stefan-it/turkish-bert

2.6.1 Results

Table 2.4 reports the best overall word accuracy results for in-domain and out-of-domain settings. We can see that among the three SES types *ses-morpheus* is the least optimal. Since its functioning principle implies that the same edit operation may generate various labels depending on the word's total number of characters (as demonstrated in Table 2.2 with the examples of the words 'cats' and 'birds'), this approach creates the highest amount of unique labels for 5 out of 7 languages of our survey (as illustrated by Table 2.3). This might be one of the possible reasons that leads to the lower performance of this SES method, as in this case the range of the SES classes is wider, which could difficult the learning and generalization processes of the model.

	ses-udpipe	ses-ixapipes	ses-morpheus
es	444	670	1,213
ru	1,157	2,390	3,208
en	286	445	891
eu	2,247	5,324	3,710
tr	236	4,147	799
cz	1,020	2,345	3,033
pl	947	1,920	2,692

Table 2.3: The amount of unique labels for each SES type.

	ses-udpipe		ses-ixapipes		ses-morpheus	
	IND	OOD	IND	OOD	IND	OOD
es	0.983	0.971	**0.983**	**0.972***	0.975	0.963
ru	**0.973**	**0.945***	0.970	0.941	0.927	0.885
en	0.991	0.939	**0.991**	**0.941**	0.979	0.916
eu	**0.969***	**0.890***	0.966	0.885	0.952	0.857
tr	**0.964***	**0.853***	0.915	0.827	0.938	0.804
cz	**0.994***	0.947	0.991	**0.951**	0.987	0.924
pl	**0.982***	**0.952**	0.980	0.950	0.943	0.917

Table 2.4: Word accuracy results for the 3 SES types for in-domain (IND) and out-of-domain (OOD) settings. In **bold**: best overall results across systems and SES types. *:results, that are statistically significant at $\alpha = .05$.

With respect to the other two methods, we could observe that for 5 out of 7 languages (namely, for Russian, Basque, Turkish, Czech and Polish) the highest word accuracy in-domain is achieved using *ses-udpipe* approach (4 out of 5 of these results are statistically significant). However, in the case of Spanish and English the results are almost identical for both *ses-udpipe* and *ses-ixapipes* methods. Regarding out-of-domain,

	ses-udpipe		ses-ixapipes		ses-morpheus	
	IND	OOD	IND	OOD	IND	OOD
es	0.703	0.489	**0.708**	**0.505***	0.582	0.397
ru	**0.614**	**0.426***	0.604	0.401	0.314	0.187
en	**0.890**	0.425	0.888	**0.439**	0.773	0.305
eu	**0.684**	**0.203***	0.663	0.195	0.551	0.150
tr	**0.707***	**0.080***	0.496	0.010	0.583	0.050
cz	**0.896***	0.430	0.855	**0.500**	0.796	0.320
pl	**0.876***	**0.656**	0.861	0.657	0.675	0.519

Table 2.5: Sentence accuracy results for the 3 SES types for in-domain (IND) and out-of-domain (OOD) settings. In **bold**: best overall results across systems and SES types. *:results, that are statistically significant at $\alpha = .05$.

in 4 out of 7 cases *ses-udpipe* is the optimal choice as well (3 statistically significant), while *ses-ixapipes* benefits the Czech language and performs similar to the UDPipe's method for English and Spanish.

Still, the differences in word accuracy results for *ses-udpipe* and *ses-ixapipes* are very small, which makes it difficult to distinguish between approaches. In order to obtain a clearer picture in the methods' performance we decided to additionally compute the sentence accuracy metric as proposed for POS tagging by Manning (2011).

As demonstrated in Table 2.5, sentence accuracy allows us to better distinguish between the models' performance. First, it confirms the results regarding *ses-morpheus* approach, achieving much lower accuracy for all the languages. Second, the almost equivalent results in word accuracy for English and Spanish using both *ses-udpipe* and *ses-ixapipes* methods are now noticeably different when evaluated using sentence accuracy. While in the case of Spanish the approach of IXA pipes seems to be more beneficial both in-and out-of-domain, for English it allows to achieve 1.4 points better in sentence accuracy out-of-domain. The same phenomenon can be observed in the case of the Czech language, with 7 points better in sentence accuracy out-of-domain for *ses-ixapipes* method with respect to *ses-udpipe*. The results for the rest of the languages follow the tendency obtained with the word accuracy metric, where the *ses-udpipe* method scores the highest.

Although sentence accuracy results provide a clearer picture, we would like to establish whether the differences are in fact statistically significant. Thus, we perform the McNemar test to determine whether the scores obtained by *ses-udpipe* and *ses-ixapipes* are statistically significant or not (null hypothesis). When evaluating word accuracy the test shows that the differences in performance of the two SES approaches mentioned above are statistically significant ($\alpha = .05$) in *ses-udpipe* favor for the agglutinative languages (Basque and Turkish) both in-domain (with p-value < 0.02 for Basque and p-value < 0.001 for Turkish) and out-of-domain (with p-value < 0.001 for both Basque and Turkish); for Czech and Polish languages in-domain (p-value < 0.001 for Czech and p-value < 0.005 for Polish) and for Russian out-of-domain (p-value < 0.001). Such small p-value results indicate that the differences in performance of the models trained with different minimum edit approaches is noticeable. The test results also suggest that in the case of lemmatizing using *ses-ixapipes* method the model commits a larger percentage of the errors respect to *ses-udpipe*. As for *ses-ixapipes*, the results are statistically significant only for Spanish in the out-of-domain setting (p-value < 0.002). For sentence accuracy the McNemar test results reflect the same tendency as for word accuracy. Therefore, the McNemar test indicates that *ses-udpipe* approach is more beneficial in the generation of the Shortest Edit Scripts that the other two methods, at least in the proposed spectrum of languages.

2.7 Discussion

In order to make the comparison of the three Shortest Edit Script methods more complete we discuss the following points. First, we analyze the performance of the pre-trained masked language models on in-vocabulary and out-of vocabulary words. The aim of such analysis is to understand which SES approach contributes better to the generalization capabilities of the MLMs. Second, we conduct a brief error analysis in order to understand what makes UDPipe's method more successful that its two other counterparts. Finally, we discuss model contamination issues.

Generalization on out-of-vocabulary words: Pre-trained masked language models, in particular XLM-RoBERTa, demonstrate good generalization abilities and are capable of achieving competitive results lemmatizing the words they did not see during the training process (Toporkov and Agerri, 2024). In order to check which SES approach benefits such capabilities more, we calculate word accuracy on in-vocabulary and out-of-vocabulary words, comparing how the model performs on the words it has seen during the training respect to the words it sees for the first time. Table 2.6 reports the results.

Interestingly, all three SES approaches perform equally well on in-vocabulary words in-domain and obtain very similar results out-of-domain. Things start changing when we analyze the out-of-vocabulary performance. We can see the significant drop in the generalization capability of the models using *ses-morpheus* approach, which confirms the word and sentence accuracy results. We also could see that for Spanish, English and Czech the results are better using *ses-ixapipes* method, the point that reinforces the sentence accuracy results. There is also a strong correlation between the results where the differences between *ses-udpipe* and *ses-ixapipes* are statistically significant and how these approaches perform on unseen words.

In any case, from an overall perspective *ses-udpipe* demonstrates stronger performance, achieving the highest accuracy in-domain for 5 out of 7 languages and out-of-domain for 4 out of 7 languages both for in-vocabulary and out-of-vocabulary

		ses-udpipe		ses-ixapipes		ses-morpheus	
		INV	OOV	INV	OOV	INV	OOV
es	ind	0.989	0.906	**0.989**	**0.912**	0.989	0.816
	ood	0.976	0.904	**0.977**	**0.917***	0.975	0.807
ru	ind	**0.995**	**0.908**	0.994	0.900	0.991	0.741
	ood	**0.972**	**0.878***	0.972	0.865	0.967	0.686
en	ind	**0.995**	**0.931**	0.994	0.927	0.993	0.751
	ood	**0.954**	0.833	0.953	**0.849**	0.954	0.631
eu	ind	**0.990**	**0.852***	**0.990**	0.832	0.989	0.748
	ood	**0.926**	**0.777***	**0.926**	0.757	0.926	0.645
tr	ind	0.991	**0.882***	0.991	0.685	**0.992**	0.775
	ood	**0.946**	**0.693***	0.945	0.625	0.944	0.564
cz	ind	**0.998**	**0.955***	**0.998**	0.923	**0.998**	0.876
	ood	0.987	0.807	**0.988**	**0.821**	0.987	0.703
pl	ind	**0.998**	**0.919***	0.997	0.909	0.992	0.742
	ood	**0.981**	**0.816**	0.981	0.808	0.974	0.650

Table 2.6: Word accuracy for in-vocabulary (INV) and out-of-vocabulary (OOV) words for in-domain (ind) and out-of-domain (ood) results. In **bold**: best results per SES and per language; *:results, that are statistically significant at $\alpha = .05$.

words. Table 7 in Appendix A provides more detailed results on out-of-vocabulary statistics with respect to lemmas and SES. Thus, the overall better performance of *ses-udpipe* is reinforced by having the lowest percentage rate of SES that have not been seen during training. This data indicates that the *ses-udpipe* approach has better generalization capabilities.

In conclusion, the results of our experiments show that the *ses-udpipe* method is more beneficial for the lemmatization task, especially in the case of the languages with more complex morphology. To analyze what makes this method better than its close counterpart *ses-ixapipes*, we conduct a brief error analysis in an attempt to identify the most important factors that may influence performance.

Error Analysis: The first noticeable advantage that is perceived in the structure of the *ses-udpipe* patterns is the absence of indexing. While *ses-ixapipes* misplaces some indexes, wrongly annotating them to the letters that should be deleted or replaced, *ses-udpipe* approach simplifies this process by only indicating the positions of the letters that should be modified without having to map it with the corresponding index. Such misplacements usually affect the complex words that need a lot of edit operations in order to be lemmatized.

Another important difference is how to deal with non-Latin alphabet and some language-specific letters. In the cases of such languages as Russian and Turkish these letters may cause a certain confusion during minimum edit generations as it happens to *ses-ixapipes*, which sometimes assigns to the final SES pattern the letters that do not appear neither in the surface word, nor in the lemma.

The third interesting observation is encountered mostly in the lemmatization of agglutinative languages (Basque and Turkish) and has to do with their suffixal nature. Whereas the *ses-udpipe* method processes the parts of the words separately, *ses-ixapipes* does not take into account this issue. Thus, *ses-ixapipes* focuses on indexing the correct letters without considering if its the part of the suffix or of the root. As a result, this approach may create an alternative minimum edit script, which may map correctly, but that does not coincide with the gold standard SES. For example, when lemmatizing the Basque word *folklorearen* ('of folklore', lemma *folklore*), the gold standard SES would be D5rD4eD3aD0n, while in one of the predictions *ses-ixapipes* offered an alternative version of SES, which is D4eD3aD2rD0n. Applying both sets of edits will deliver the same result, but as the goal of the classification task is to correctly assign the corresponding SES to its surface word, such cases are considered incorrect. In order to check whether this could be crucial in evaluating the overall SES performance, we calculate the total number of occurrences where the SES distinct from the gold standard delivers the correct lemma for the Basque language. Our results show that for *ses-udpipe* approach there are 9 out of 11901 cases where an alternative SES leads to the same lemma (in-domain), while in the case of *ses-ixapipes* the number of such occurrences is 17 out of 11901 respectively. This data indicates, that although such cases could appear, their influence on the overall result is insignificant.

Finally, it also seems beneficial to encode the casing script as implemented in the *ses-udpipe* method and which is only partially implemented in both *ses-ixapipes* and *ses-morpheus* approaches.

Regarding the other two minimum edit approaches, namely, *ses-ixapipes* and *ses-morpheus*, a brief error inspection shows that in the case of *ses-ixapipes* most of the errors are of suffixal and root nature, more precisely, in the incorrect indexing or letter misplacement. Furthermore, the performance of *ses-morpheus* is mainly affected by the large number of generated SES classes, which makes the classification task much more difficult. The cases where lemma is longer than wordform, and, therefore the edit operations are applied jointly, constitute between 5 and 15 of the total error rate across the inspected languages, and is another source of possible low performance of this method with respect to the other two.

2.8 Conclusion

In this paper, we present the first detailed systematic comparison of three popular methods to compute Shortest Edit Scripts (SES), widely used in modern contextual lemmatization models. After a comprehensive battery of experiments with various evaluation metrics and statistical tests, results indicate that *ses-udpipe* is the optimal method for contextual lemmatization among the Shortest Edit Script approaches. Its main advantages consist in: (i) computing casing and edit operations separately; (ii) processing the wordform by morphemes and the absence of indexing, which allows to avoid the cases where there are the same letters in the suffix and the root (especially for agglutinative languages such as Basque and Turkish) and to create less ambiguous SES; (iii) better generalization capabilities, that result in obtaining less out-of-vocabulary SES and creating fewer SES labels, which benefits the models by having to learn a smaller amount of SES classes. Furthermore, our results indicate the following: (i) more metrics should be implemented in the analysis of the MLMs performance along with the word accuracy; (ii) out-of-domain evaluation should be considered as an important step as it allows to obtain a clearer picture of how far the task is solved.

We believe that the results of our study could be useful for the future development of contextual lemmatizers which may include SES as an integral component of their systems.

CHAPTER 3

Future Work

In addition to the issues already raised in this paper, contextual lemmatization an additional challenge, namely, the lack of annotation data for any domain and language in order to avoid deploying contextual lemmatizer in out-of-domain settings.

As demonstrated by the empirical results obtained, contextual lemmatizers substantially degrade when evaluated out-of-domain, their most common application use-case. This means that obtaining optimal results would require to manually generate annotated data for each application domain and language, an unfeasible task in terms of monetary cost and human effort.

Cross-lingual Transfer for Cross-Domain Contextual Lemmatization: Therefore, devising an alternative strategy to obtain good quality contextual lemmatizers when no manually annotated data is available for a given specific domain and language is paramount. Previous work has addressed this problem for other Natural Language Processing tasks by proposing various crosslingual transfer techniques (García-Ferrero et al., 2022, 2023). The basic idea is to leverage available knowledge (annotations) for a given language and domain (usually English) to automatically generate taggers for that domain in different target languages (Agerri et al., 2018).

The emergence of multilingual language models (Devlin et al., 2019; Conneau et al., 2020) allows for *model-based* crosslingual transfer. In the ***model-transfer*** approach a language model fine-tuned in a given source language, typically English, can directly be applied to other target languages. However, good results can also be obtained by either machine translating the training data from English into the target languages or, conversely, translating the test data from the target language into English (Hu et al., 2020; Artetxe et al., 2023). In what can be called the *data-transfer* approach.

Data-transfer methods aim to automatically generate labelled data for a target language. Previous works on *data-transfer* have proposed *translation and annotation projection* as an effective technique for zero-resource cross-lingual sequence labelling (Jain

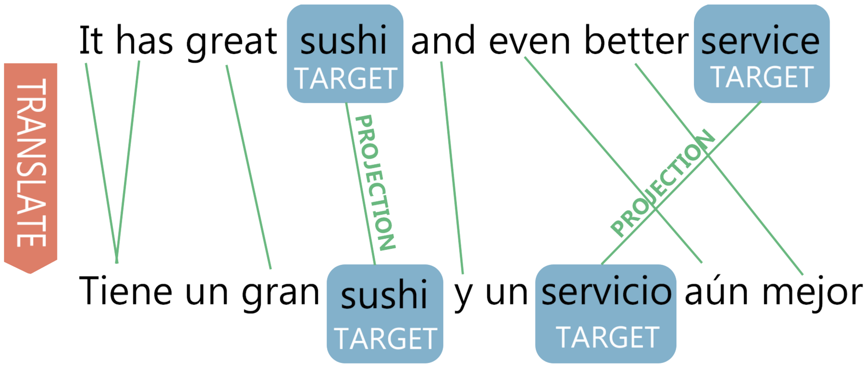

It has great sushi and even better service
TARGET
TARGET
TRANSLATE
PROJECTION
PROJECTION
Tiene un gran sushi y un servicio aún mejor
TARGET
TARGET

Figure 3.1: Illustration of the *data-transfer* method for Opinion Target Extraction (OTE). Original figure from García-Ferrero et al. (2022).

et al., 2019; Fei et al., 2020). In this setting, as illustrated by Figure 3.1 for the sequence labelling task of Opinion Target Extraction, the idea is to translate gold-labelled text into a target language and then, using automatic word alignments, project the labels from the source into the target language. The result is an automatically generated dataset in the target language that can be used for training a sequence labelling model (Yarowsky et al., 2001; Ehrmann et al., 2011; Agerri et al., 2018; García-Ferrero et al., 2022).

Despite crosslingual transfer (both model- and data-transfer) being successfully applied in many sequence labelling tasks to mitigate the lack of annotated data for a given language in a specific domain, this technique has not yet been applied to lemmatization. As future work, we propose to study crosslingual transfer methods for contextual lemmatization. Thus, instead of applying contextual lemmatizers trained on Universal Dependencies or Unimorph across domains as it is now customary, the idea would be to leverage lemma annotations in some specific language and domain to apply crosslingual transfer into a target language for which we do not have any manually annotated data. Therefore, by doing so we would be lemmatizing with a model fine-tuned on domain-specific data which will hopefully lead to a increase in performance with respect to the out-of-domain evaluation results reported in Section 1.6.2.

In any case, it should be noticed that it is not possible to directly apply *model-transfer* and *data-transfer* techniques as formulated by previous work. Even though contextual lemmatization can, as we have seen in this paper so far, be cast as a sequence labelling or token classification task, previous work has focused on tasks such as Named Entity Recognition, Opinion Target Extraction, Semantic Role Labelling or POS tagging. In all these tasks, MLMs typically contextually learn to assign a label to a given token. Furthermore, the relation between a label such as *B-LOC* or *NOUN* or *TARGET* and the words they are assigned to is completely arbitrary in the sense that the label is not generated based on the word form itself and it does not change across languages. Thus, what models learn is that a given sequence of characters should be assigned a specific label based on the context it occurs. Now, going back to Figure 3.1, in *model-transfer* MLMs are fine-tuned in English and learn that the word 'service' is a *TARGET* and that, when applied to predict in Spanish, the word to which to attach the label learned for 'service' corresponds, in Spanish, to 'servicio'.

If we consider contextual lemmatization as presented in this paper it can easily be seen that *model-transfer* will quickly run into difficulties as it would be learning Shortest Edit Scripts (SES) induced from the word forms and lemmas themselves in English which are then applied to a different language in which the SES induced from the corresponding word forms and lemmas are going to be completely different. To cut a long story short, lemma classes, unlike in NER, OTE or POS tagging, do change across languages because they are induced by calculating the edit distance between a word form and its lemma.

With respect to *data-transfer*, the majority of previous published work on this line of research explores the application of word-alignments (Ehrmann et al., 2011), projection methods based on word-alignments have achieved mixed results as they often produce partial, incorrect or missing annotation projections (García-Ferrero et al., 2022). This is due to the fact that word alignments are computed on a word-by-word basis leveraging word co-occurrences or similarity between word vector representations. That is, without taking into consideration the labeled spans or categories to be projected. Other techniques have also been proposed, such as fine-tuning language models in the span projection task (Li et al., 2021), translating the labeled spans independently from the sentence (Zhou et al., 2022) or including markers during the machine translation step (Chen et al., 2023). However, automatic annotation projection remains both an effective technique on specialized domains (Yeginbergen and Agerri, 2024) and an open and difficult research challenge.

Regarding *data-transfer* for contextual lemmatization, projecting the labels (SES) via word alignments would be pairing a SES induced for a word form in the source language with a completely different word form in the target language which, at decoding time, will give us the completely wrong lemma. In order to solve this issue a possible strategy could consist of: (i) projecting the lemma itself via word alignments; (ii) translating the projected lemma to its corresponding lemma in the target language, and (iii) inducing the SES between the translated lemma and the word form in the target language.

Step (ii) may be hindered by out-of-vocabulary issues if using dictionary-based approaches to translate the lemmas, especially for non-standard language. Thus, this step may need careful consideration and could involve experimenting with Machine Translation (Costa-jussà et al., 2022) and/or decoder-based Large Language Models (LLMs) to maximize the recall in the lemma translation process (Touvron et al., 2023; Jiang et al., 2023).

Acknowledgements

This work has been partially supported by the HiTZ center and the Basque Government (Research group funding IT-1805-22). We are also thankful to several MCIN/AEI/10.13039/501100011033 projects: (i) DeepKnowledge (PID2021-127777OB-C21) and by FEDER, EU; (ii) Disargue (TED2021-130810B-C21) and European Union NextGenerationEU/PRTR; (iii) DeepMinor (CNS2023-144375) and European Union NextGenerationEU/PRTR. Olia Toporkov is supported by a doctoral grant from the UPV/EHU “Formación de Personal Investigador” (PIF 20/186). Rodrigo Agerri was additionally funded by the RYC-2017-23647 fellowship (MCIN/AEI/10.13039/501100011033 and by ESF Investing in your future).

Bibliography

Agerri, R. and Agirre, E. (2023). Lessons learned from the evaluation of Spanish Language Models. *Proces. del Leng. Natural*, 70:157–170.

Agerri, R., Bermudez, J., and Rigau, G. (2014). IXA pipeline: Efficient and ready to use multilingual NLP tools. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 3823–3828, Reykjavik, Iceland. European Language Resources Association (ELRA).

Agerri, R., Chung, Y., Aldabe, I., Aranberri, N., Labaka, G., and Rigau, G. (2018). Building named entity recognition taggers via parallel corpora. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Agerri, R. and Rigau, G. (2016). Robust multilingual named entity recognition with shallow semi-supervised features. *Artificial Intelligence*, 238(2):63–82.

Agerri, R., San Vicente, I., Campos, J. A., Barrena, A., Saralegi, X., Soroa, A., and Agirre, E. (2020). Give your text representation models some love: the case for Basque. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4781–4788, Marseille, France. European Language Resources Association.

Agić, Ž. and Schluter, N. (2017). How (not) to train a dependency parser: The curious case of jackknifing part-of-speech taggers. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 679–684, Vancouver, Canada. Association for Computational Linguistics.

Aiken, B., Kelly, J., Palmer, A., Polat, S. O., Rama, T., and Nielsen, R. (2019). Sigmorphon 2019 task 2 system description paper: Morphological analysis in context for many languages, with supervision from only a few. In *Proceedings of the 16th Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 87–94, Florence, Italy. Association for Computational Linguistics.

Akbik, A., Blythe, D., and Vollgraf, R. (2018). Contextual string embeddings for sequence labeling. In *Proceedings of the 27th International Conference on*

Computational Linguistics, pages 1638–1649, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Aldezabal, I., Aranzabe, M., Diaz de Ilarraza, A., and Fernández, K. (2008). From dependencies to constituents in the reference corpus for the processing of Basque. *Procesamiento del Lenguaje Natural*, (41):147–154.

Alegria, I., Artola, X., Sarasola, K., and Urkia, M. (1996). Automatic morphological analysis of basque. *Literary and Linguistic Computing*, 11:193–203.

Arkhipov, M., Trofimova, M., Kuratov, Y., and Sorokin, A. (2019). Tuning multilingual transformers for language-specific named entity recognition. In *Proceedings of the 7th Workshop on Balto-Slavic Natural Language Processing*, pages 89–93, Florence, Italy. Association for Computational Linguistics.

Artetxe, M., Goswami, V., Bhosale, S., Fan, A., and Zettlemoyer, L. (2023). Revisiting machine translation for cross-lingual classification. In Bouamor, H., Pino, J., and Bali, K., editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6489–6499. Association for Computational Linguistics.

Belinkov, Y., Durrani, N., Dalvi, F., Sajjad, H., and Glass, J. R. (2017). What do neural machine translation models learn about morphology? In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 861–872, Vancouver, Canada. Association for Computational Linguistics.

Bergmanis, T. and Goldwater, S. (2018). Context sensitive neural lemmatization with Lematus. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1391–1400, New Orleans, Louisiana. Association for Computational Linguistics.

Carreras, X., Chao, I., Padró, L., and Padró, M. (2004). Freeling: An open-source suite of language analyzers. In *LREC*.

Cañete, J., Chaperon, G., Fuentes, R., Ho, J.-H., Kang, H., and Pérez, J. (2020). Spanish pre-trained BERT model and evaluation data. In *PML4DC at ICLR 2020*.

Chakrabarty, A., Pandit, O. A., and Garain, U. (2017). Context sensitive lemmatization using two successive bidirectional gated recurrent networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1481–1491, Vancouver, Canada. Association for Computational Linguistics.

Chen, Y., Jiang, C., Ritter, A., and Xu, W. (2023). Frustratingly easy label projection for cross-lingual transfer. In Rogers, A., Boyd-Graber, J. L., and Okazaki, N., editors, *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 5775–5796. Association for Computational Linguistics.

Chrupala, G., Dinu, G., and van Genabith, J. (2008). Learning morphology with Morfette. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

Collins, M. (2002). Discriminative training methods for Hidden Markov Models: Theory and experiments with perceptron algorithms. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002)*, pages 1–8. Association for Computational Linguistics.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., and Stoyanov, V. (2020). Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Conneau, A., Kruszewski, G., Lample, G., Barrault, L., and Baroni, M. (2018). What you can cram into a single $&!#* vector: Probing sentence embeddings for linguistic properties. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2126–2136, Melbourne, Australia. Association for Computational Linguistics.

Costa-jussà, M. R., Cross, J., Çelebi, O., Elbayad, M., Heafield, K., Heffernan, K., Kalbassi, E., Lam, J., Licht, D., Maillard, J., Sun, A., Wang, S., Wenzek, G., Youngblood, A., Akula, B., Barrault, L., Gonzalez, G. M., Hansanti, P., Hoffman, J., Jarrett, S., Sadagopan, K. R., Rowe, D., Spruit, S., Tran, C., Andrews, P., Ayan, N. F., Bhosale, S., Edunov, S., Fan, A., Gao, C., Goswami, V., Guzmán, F., Koehn, P., Mourachko, A., Ropers, C., Saleem, S., Schwenk, H., and Wang, J. (2022). No language left behind: Scaling human-centered machine translation. *CoRR*, abs/2207.04672.

Cotterell, R. and Heigold, G. (2017). Cross-lingual character-level neural morphological tagging. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 748–759, Copenhagen, Denmark. Association for Computational Linguistics.

Dayanik, E., Akyürek, E., and Yuret, D. (2018). MorphNet: A sequence-to-sequence model that combines morphological analysis and disambiguation. *CoRR*, abs/1805.07946.

de Marneffe, M.-C., Dozat, T., Silveira, N., Haverinen, K., Ginter, F., Nivre, J., and Manning, C. D. (2014). Universal Stanford dependencies: A cross-linguistic typology. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 4585–4592, Reykjavik, Iceland. European Language Resources Association (ELRA).

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (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), pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Dhonnchadha, E. U. (2002). A two-level morphological analyser and generator for Irish using finite-state transducers. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC'02)*, Las Palmas, Canary Islands - Spain. European Language Resources Association (ELRA).

Dietterich, T. G. (1998). Approximate statistical test for comparing supervised classification learning algorithms. *Neural Computation*, 10(7):1895–1923.

Ehrmann, M., Turchi, M., and Steinberger, R. (2011). Building a multilingual named entity-annotated corpus using annotation projection. In *Proceedings of the International Conference Recent Advances in Natural Language Processing 2011*, pages 118–124, Hissar, Bulgaria. Association for Computational Linguistics.

Fei, H., Zhang, M., and Ji, D. (2020). Cross-lingual semantic role labeling with high-quality translated training corpus. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7014–7026, Online. Association for Computational Linguistics.

García-Ferrero, I., Agerri, R., and Rigau, G. (2022). Model and data transfer for cross-lingual sequence labelling in zero-resource settings. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 6403—-6416, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

García-Ferrero, I., Agerri, R., and Rigau, G. (2023). T-projection: High quality annotation projection for sequence labeling tasks. In Bouamor, H., Pino, J., and Bali, K., editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 15203–15217, Singapore.

Gesmundo, A. and Samardžić, T. (2012). Lemmatisation as a tagging task. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 368–372, Jeju Island, Korea. Association for Computational Linguistics.

Grave, E., Bojanowski, P., Gupta, P., Joulin, A., and Mikolov, T. (2018). Learning word vectors for 157 languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Heigold, G., Neumann, G., and van Genabith, J. (2017). An extensive empirical evaluation of character-based morphological tagging for 14 languages. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 505–513, Valencia, Spain. Association for Computational Linguistics.

Hladká, B., Hajič, J., Hana, J., Hlaváčová, J., Mírovský, J., and Raab, J. (2008). The Czech academic corpus 2.0 guide. *The Prague Bulletin of Mathematical Linguistics*, 89:41–96.

Hu, J., Ruder, S., Siddhant, A., Neubig, G., Firat, O., and Johnson, M. (2020). XTREME: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 4411–4421. PMLR.

Huang, Z., Xu, W., and Yu, K. (2015). Bidirectional LSTM-CRF models for sequence tagging. *CoRR*, abs/1508.01991.

Jain, A., Paranjape, B., and Lipton, Z. C. (2019). Entity projection via machine translation for cross-lingual NER. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1083–1092, Hong Kong, China. Association for Computational Linguistics.

Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., de Las Casas, D., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., Lavaud, L. R., Lachaux, M.-A., Stock, P., Scao, T. L., Lavril, T., Wang, T., Lacroix, T., and Sayed, W. E. (2023). Mistral 7b. *ArXiv*, abs/2310.06825.

Jongejan, B. and Dalianis, H. (2009). Automatic training of lemmatization rules that handle morphological changes in pre-, in- and suffixes alike. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 145–153, Suntec, Singapore. Association for Computational Linguistics.

Karttunen, L., Kaplan, R. M., and Zaenen, A. (1992). Two-level morphology with composition. In *COLING 1992 Volume 1: The 14th International Conference on Computational Linguistics.*

Kingma, D. P. and Ba, J. (2015). Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980.

Kirov, C., Cotterell, R., Sylak-Glassman, J., Walther, G., Vylomova, E., Xia, P., Faruqui, M., Mielke, S. J., McCarthy, A., Kübler, S., Yarowsky, D., Eisner, J., and Hulden, M. (2018). UniMorph 2.0: Universal Morphology. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

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*, pages 12–18, Florence, Italy. Association for Computational Linguistics.

Kondratyuk, D. and Straka, M. (2019). 75 languages, 1 model: Parsing Universal Dependencies universally. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2779–2795, Hong Kong, China. Association for Computational Linguistics.

Kuratov, Y. and Arkhipov, M. Y. (2019). Adaptation of deep bidirectional multilingual transformers for Russian language. *CoRR*, abs/1905.07213.

Li, B., He, Y., and Xu, W. (2021). Cross-lingual named entity recognition using parallel corpus: A new approach using xlm-roberta alignment. *ArXiv preprint*, abs/2101.11112.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). RoBERTa: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Lyashevkaya, O., Droganova, K., Zeman, D., Alexeeva, M., Gavrilova, T., Mustafina, N., and Shakurova, E. (2016). Universal Dependencies for Russian: A new syntactic dependencies tagset. *SSRN Electronic Journal*.

Ma, X. and Hovy, E. (2016). End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1064–1074, Berlin, Germany. Association for Computational Linguistics.

Malaviya, C., Wu, S., and Cotterell, R. (2019). A simple joint model for improved contextual neural lemmatization. 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)*, pages 1517–1528, Minneapolis, Minnesota. Association for Computational Linguistics.

Manning, C. D. (2011). Part-of-speech tagging from 97% to 100%: Is it time for some linguistics? In *Computational Linguistics and Intelligent Text Processing*, pages 171–189, Berlin, Heidelberg. Springer.

Manning, C. D., Clark, K., Hewitt, J., Khandelwal, U., and Levy, O. (2020). Emergent linguistic structure in artificial neural networks trained by self-supervision. *Proceedings of the National Academy of Sciences*, 117:30046–30054.

McCarthy, A. D., Kirov, C., Grella, M., Nidhi, A., Xia, P., Gorman, K., Vylomova, E., Mielke, S. J., Nicolai, G., Silfverberg, M., Arkhangelskiy, T., Krizhanovsky, N., Krizhanovsky, A., Klyachko, E., Sorokin, A., Mansfield, J., Ernštreits, V., Pinter, Y., Jacobs, C. L., Cotterell, R., Hulden, M., and Yarowsky, D. (2020). UniMorph 3.0: Universal Morphology. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 3922–3931, Marseille, France. European Language Resources Association.

McCarthy, A. D., Vylomova, E., Wu, S., Malaviya, C., Wolf-Sonkin, L., Nicolai, G., Kirov, C., Silfverberg, M., Mielke, S. J., Heinz, J., Cotterell, R., and Hulden, M. (2019). The SIGMORPHON 2019 shared task: Morphological analysis in context and cross-lingual transfer for inflection. In *Proceedings of the 16th Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 229–244, Florence, Italy. Association for Computational Linguistics.

Mroczkowski, R., Rybak, P., Wróblewska, A., and Gawlik, I. (2021). HerBERT: Efficiently pretrained transformer-based language model for Polish. In *Proceedings of the 8th Workshop on Balto-Slavic Natural Language Processing*, pages 1–10, Kiyv, Ukraine. Association for Computational Linguistics.

Müller, T., Cotterell, R., Fraser, A., and Schütze, H. (2015). Joint lemmatization and morphological tagging with lemming. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2268–2274, Lisbon, Portugal. Association for Computational Linguistics.

Nivre, J., Zeman, D., Ginter, F., and Tyers, F. (2017). Universal Dependencies. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Tutorial Abstracts*, Valencia, Spain. Association for Computational Linguistics.

Oflazer, K. (1993). Two-level description of Turkish morphology. In *Sixth Conference of the European Chapter of the Association for Computational Linguistics*, Utrecht, The Netherlands. Association for Computational Linguistics.

Ortiz Suárez, P. J., Sagot, B., and Romary, L. (2019). Asynchronous pipelines for processing huge corpora on medium to low resource infrastructures. In Bański, P., Barbaresi, A., Biber, H., Breiteneder, E., Clematide, S., Kupietz, M., Lüngen, H., and Iliadi, C., editors, *Proceedings of the Workshop on Challenges in the Management of Large Corpora (CMLC-7) 2019. Cardiff, 22nd July 2019*, pages 9–16.

Przepiórkowski, A. and Patejuk, A. (2018). From Lexical Functional Grammar to enhanced Universal Dependencies. In *Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018)*, pages 2–4, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Segalovich, I. (2003). A fast morphological algorithm with unknown word guessing induced by a dictionary for a web search engine. In *MLMTA*.

Silveira, N., Dozat, T., de Marneffe, M.-C., Bowman, S., Connor, M., Bauer, J., and Manning, C. (2014). A gold standard dependency corpus for English. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 2897–2904, Reykjavik, Iceland. European Language Resources Association (ELRA).

Straka, M. (2018). UDPipe 2.0 prototype at CoNLL 2018 UD shared task. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 197–207, Brussels, Belgium. Association for Computational Linguistics.

Straka, M., Straková, J., and Hajic, J. (2019). UDPipe at SIGMORPHON 2019: Contextualized embeddings, regularization with morphological categories, corpora merging. In *Proceedings of the 16th Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 95–103, Florence, Italy. Association for Computational Linguistics.

Stroppa, N. and Yvon, F. (2005). An analogical learner for morphological analysis. In *Proceedings of the Ninth Conference on Computational Natural Language Learning (CoNLL-2005)*, pages 120–127, Ann Arbor, Michigan. Association for Computational Linguistics.

Sulubacak, U., Gokirmak, M., Tyers, F., Çöltekin, Ç., Nivre, J., and Eryiğit, G. (2016). Universal Dependencies for Turkish. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3444–3454, Osaka, Japan. The COLING 2016 Organizing Committee.

Sylak-Glassman, J. (2016). The composition and use of the universal morphological feature schema (UniMorph schema). Technical report.

Taulé, M., Martí, M. A., and Recasens, M. (2008). AnCora: Multilevel annotated corpora for Catalan and Spanish. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

Tiedemann, J. (2012). Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218, Istanbul, Turkey. European Language Resources Association (ELRA).

Tiedemann, J. (2016). OPUS – parallel corpora for everyone. In *Proceedings of the 19th Annual Conference of the European Association for Machine Translation: Projects/Products*, Riga, Latvia. Baltic Journal of Modern Computing.

Toporkov, O. and Agerri, R. (2024). On the Role of Morphological Information for Contextual Lemmatization. *Computational Linguistics*, pages 1–35.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., Bikel, D., Blecher, L., Ferrer, C. C., Chen, M., Cucurull, G., Esiobu, D., Fernandes, J., Fu, J., Fu, W., Fuller, B., Gao, C., Goswami, V., Goyal, N., Hartshorn, A., Hosseini, S., Hou, R., Inan, H., Kardas, M., Kerkez, V., Khabsa, M., Kloumann, I., Korenev, A., Koura, P. S., Lachaux, M.-A., Lavril, T., Lee, J., Liskovich, D., Lu, Y., Mao, Y., Martinet, X., Mihaylov, T., Mishra, P., Molybog, I., Nie, Y., Poulton, A., Reizenstein, J., Rungta, R., Saladi, K., Schelten, A., Silva, R., Smith, E. M., Subramanian, R., Tan, X. E., Tang, B., Taylor, R., Williams, A., Kuan, J. X., Xu, P., Yan, Z., Zarov, I., Zhang, Y., Fan, A., Kambadur, M., Narang, S., Rodriguez, A., Stojnic, R., Edunov, S., and Scialom, T. (2023). Llama 2: Open foundation and fine-tuned chat models.

Türk, U., Atmaca, F., Özateş, Ş. B., Köksal, A., Ozturk Basaran, B., Gungor, T., and Özgür, A. (2019). Turkish treebanking: Unifying and constructing efforts. In *Proceedings of the 13th Linguistic Annotation Workshop*, pages 166–177, Florence, Italy. Association for Computational Linguistics.

van den Bosch, A. and Daelemans, W. (1999). Memory-based morphological analysis. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pages 285–292, College Park, Maryland, USA. Association for Computational Linguistics.

Vaswani, A., Shazeer, N. M., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. *ArXiv*, abs/1706.03762.

Wang, Z., Wang, Y., Wu, J., Teng, Z., and Yang, J. (2022). YATO: Yet another deep learning based text analysis open toolkit. *arXiv preprint arXiv:2209.13877*.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., and Brew, J. (2020). Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Wu, S. and Cotterell, R. (2019). Exact hard monotonic attention for character-level transduction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1530–1537, Florence, Italy. Association for Computational Linguistics.

Yarowsky, D., Grace, N., Richard, W., et al. (2001). Inducing multilingual text analysis tools via robust projection across aligned corpora. In *Proceedings of the First International Conference on Human Language Technology Research*, pages 1–8.

Yeginbergen, A. and Agerri, R. (2024). Cross-lingual argument mining in the medical domain.

Yildiz, E. and 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*, pages 25–34, Florence, Italy. Association for Computational Linguistics.

Zeldes, A. (2017). The GUM corpus: Creating multilayer resources in the classroom. *Language Resources and Evaluation*, 51(3):581–612.

Zeman, D., Popel, M., Straka, M., Hajič, J., Nivre, J., Ginter, F., Luotolahti, J., Pyysalo, S., Petrov, S., Potthast, M., Tyers, F., Badmaeva, E., Gokirmak, M., Nedoluzhko, A., Cinková, S., Hajič jr., J., Hlaváčová, J., Kettnerová, V., Urešová, Z., Kanerva, J., Ojala, S., Missilä, A., Manning, C. D., Schuster, S., Reddy, S., Taji, D., Habash, N., Leung, H., de Marneffe, M.-C., Sanguinetti, M., Simi, M., Kanayama, H., de Paiva, V., Droganova, K., Martínez Alonso, H., Çöltekin, Ç., Sulubacak, U., Uszkoreit, H., Macketanz, V., Burchardt, A., Harris, K., Marheinecke, K., Rehm, G., Kayadelen, T., Attia, M., Elkahky, A., Yu, Z., Pitler, E., Lertpradit, S., Mandl, M., Kirchner, J., Alcalde, H. F., Strnadová, J., Banerjee, E., Manurung, R., Stella, A., Shimada, A., Kwak, S., Mendonça, G., Lando, T., Nitisaroj, R., and Li, J. (2017). CoNLL 2017 shared task: Multilingual parsing from raw text to Universal Dependencies. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 1–19, Vancouver, Canada. Association for Computational Linguistics.

Zhou, R., Li, X., Bing, L., Cambria, E., Si, L., and Miao, C. (2022). Conner: Consistency training for cross-lingual named entity recognition. In Goldberg, Y., Kozareva, Z., and Zhang, Y., editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 8438–8449. Association for Computational Linguistics.

Appendix I: Morphology in Contextual Lemmatization

	ixa-mm	ixa-gs	morph	flair	mBERT	xlm-r	mono	base	UDPipe
en	**99.06**	98.95	98.10	98.58	98.56	98.76	98.49	97.68	99.01
es	98.89	98.74	99.02	99.02	99.01	**99.08**	99.04	98.42	99.31
ru	95.07	93.22	96.30	96.18	96.70	**97.08**	96.55	95.67	97.77
eu	93.41	94.06	**96.39**	96.09	95.71	95.98	95.51	96.07	97.14
cz	97.86	96.63	98.76	98.87	99.07	**99.25**	99.01	97.82	99.31
tr	85.52	86.57	**96.18**	93.98	95.15	95.38	95.20	96.41	96.84

Table 1: Overall in-domain lemmatization results for models trained with and without explicit morphological features; monolingual transformers: Russian - ruBERT, Czech - slavicBERT, Basque -BERTeus, Turkish - BERTurk, English - RoBERTa, Spanish - BETO.

	ixa-mm	ixa-gs	morph	flair	mBERT	xlm-r	mono
en	**95.16**	95.13	92.92	93.42	93.50	93.56	93.39
es	**97.59**	97.45	90.35	90.29	90.27	90.26	90.34
ru	**91.00**	88.66	87.57	89.90	90.07	90.53	89.71
eu	85.33	86.31	**89.03**	88.76	87.79	88.15	87.62
cz	92.33	91.81	91.92	95.02	94.72	**95.18**	94.40
tr	80.33	80.50	84.74	83.51	84.40	**84.90**	84.46

Table 2: Overall out-of-domain lemmatization results for models with and without explicit morphological features.

	ixa-mm	ixa-gs	morph	flair	mBERT	xlm-r	mono
en	**88.27**	81.90	80.46	85.03	84.00	85.99	82.74
es	75.28	73.16	78.03	78.34	77.59	**79.03**	78.15
ru	45.73	33.40	55.27	54.47	58.25	**61.03**	55.67
eu	44.56	50.78	**65.44**	61.44	60.00	61.44	56.78
cz	69.45	56.17	81.10	83.21	84.99	**87.62**	83.69
tr	28.90	35.82	**69.68**	59.75	64.18	64.54	64.36

Table 3: In-domain sentence accuracy results.

	ixa-mm	ixa-gs	morph	flair	mBERT	xlm-r	mono
en	**49.55**	46.58	35.45	37.73	39.55	42.27	37.73
es	**52.38**	50.17	21.49	21.83	21.32	21.66	21.83
ru	26.87	21.26	22.69	27.03	27.18	**28.26**	26.64
eu	13.11	14.54	**19.54**	19.50	17.29	17.91	17.23
cz	29.00	25.00	36.00	**48.00**	40.00	47.00	45.00
tr	3.00	7.00	8.00	7.00	9.00	**10.00**	8.00

Table 4: Out-of-domain sentence accuracy results.

	ixa-mm	ixa-gs	morph	mBERT	xlm-r	mono	base	UDPipe
en	97.56	97.12	**97.78**	97.19	97.70	96.90	97.41	98.63
es	98.70	98.53	98.98	99.14	99.19	**99.23**	98.54	99.46
ru	96.76	96.84	96.93	98.66	**98.93**	98.67	95.92	98.92
cz	89.59	88.03	93.11	93.01	93.06	**93.37**	93.58	98.13
tr	77.77	78.33	**87.02**	83.40	85.07	82.56	86.02	89.03

Table 5: Overall in-domain lemmatization results (reversed setting) for models with and without explicit morphological features.

	ixa-mm	ixa-gs	morph	mBERT	xlm-r	mono
en	**91.22**	90.55	88.97	90.80	91.21	90.94
es	**87.90**	87.47	87.50	87.51	87.65	87.33
ru	85.37	86.25	86.10	87.51	**88.43**	87.64
cz	86.09	83.94	89.13	89.73	**90.10**	89.17
tr	70.61	70.95	**81.03**	77.01	78.22	76.87

Table 6: Overall out-of-domain lemmatization results (reversed setting) for models with and without explicit morphological features.

Appendix II: SES

Appendix A. Detailed Out-of-Vocabulary Results

				ses-udpipe		ses-ixapipes		ses-morpheus	
		oov words	oov lemmas	oov ses	oov lemmas (ses in train)	oov ses	oov lemmas (ses in train)	oov ses	oov lemmas (ses in train)
es	ind	7.85	6.18	**0.02**	99.89	0.05	99.74	0.11	99.07
	ood	7.65	5.93	**0.28**	96.41	0.43	98.27	**0.28**	97.86
ru	ind	25.50	13.74	**0.27**	99.04	0.67	98.31	0.78	97.35
	ood	29.21	15.36	**1.65**	96.23	2.45	95.03	2.52	94.33
en	ind	5.71	4.19	**0.08**	99.72	0.10	99.81	0.25	97.57
	ood	11.89	11.45	**1.32**	90.41	1.56	91.58	1.36	91.15
eu	ind	15.28	5.07	**0.61**	96.52	1.45	94.86	0.92	94.69
	ood	24.26	11.99	**1.13**	95.98	2.49	94.68	1.45	94.78
tr	ind	24.83	5.67	**0.12**	99.69	4.52	95.69	0.56	97.23
	ood	36.71	20.72	**0.45**	97.85	6.52	91.40	3.40	97.85
cz	ind	8.85	3.19	**0.09**	99.11	0.20	98.66	0.24	97.90
	ood	21.97	11.76	**2.33**	99.12	2.59	99.12	2.90	98.24
pl	ind	19.53	7.80	**0.28**	99.22	0.52	98.82	0.85	97.84
	ood	17.65	8.72	**0.40**	98.65	0.62	97.57	0.67	97.57

Table 7: The proportion (in %) of out-of-vocabulary words, lemmas and SES in the in-domain (ind) and out-of-domain (ood) test sets with respect to the train set, per language. In **bold**: lowest percentage of out-of-vocabulary (oov) SES among the three SES types.

Table 7 reports the proportion of out-of-vocabulary (oov) words, lemmas and SES, both for in-domain (ind) and out-of-domain (ood) settings for the three SES types. By out-of-vocabulary we understand words, lemmas and SES in the test sets that the system did not see during the training process. The column *'oov lemmas (ses in train)'* refers to the proportion of lemmas that the model does not see during the training (out-of-vocabulary lemmas) while their corresponding SES exist in the train set. In other words, they have been seen by the system.