Which SES should we use in our lemmatizers?

Evaluating Shortest Edit Script Methods for Contextual Lemmatization

Olia Toporkov, Rodrigo Agerri

HiTZ Center - Ixa, University of the Basque Country UPV/EHU

LREC-COLING 2024

Motivation

- Contextual lemmatizers often rely on Shortest Edit Scripts (SES);
- Different methods of computing SES;
- We investigate the direct impact of SES in the final lemmatization performance.

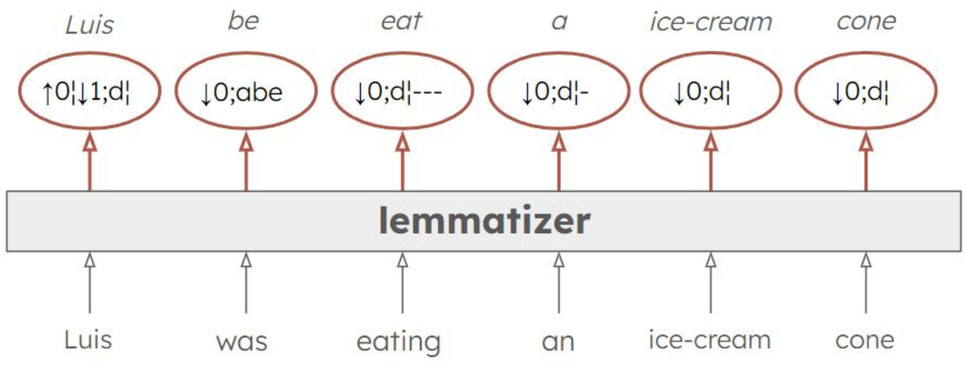
We compare 3 SES types

word→*lemma*	ses-udpipe	ses-ixapipes	ses-moprheus
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

Approach & Data

	corpus		number of unique labels		
	ind	ood	ses-udpipe	ses-ixapipes	ses-morpheus
es	GSD	AnCora	444	670	1,213
ru	GSD	SynTagRus	1,157	2,390	3,208
en	EWT	GUM	286	445	891
eu	BDT	Armiarma	2,247	5,324	3,710
tr	IMST	PUD	236	4,147	799
cz	CAC	PUD	1,020	2,345	3,033
pl	LFG	CZ	947	1,920	2,692

Focusing on lemmatization as a token classification task:



Results

Word accuracy*

	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

*trained with XLM-RoBERTa large, best configuration

Results

Sentence accuracy

	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

Discussion

Generalization on Out-of-Vocabulary words

		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

Error analysis

- Indexing:

  folklorearen ⟶ *folklore*

  1) folklorearen D5rD4eD3aD0n - folklo**rea**re**n** - folklore
  2) folklorearen D4eD3aD2rD0n - folklo**rea**re**n** - folklore
- Dealing with non-Latin alphabet/language-specific letters (Russian, Turkish);
- Encoding the casing script is beneficial;
- Large number of generated SES classes is more difficult to learn.

Why is ses-udpipe the best option?

- best results by computing casing and edit operations separately;
- do not rely on positional indexing (especially for agglutinative languages such as Basque and Turkish);
- ses-udpipe reduces the variability in edit strings.