Lexical Normalization for Code-switched Data and its Effect on POS Tagging

IT UNIVERSITY OF CPH

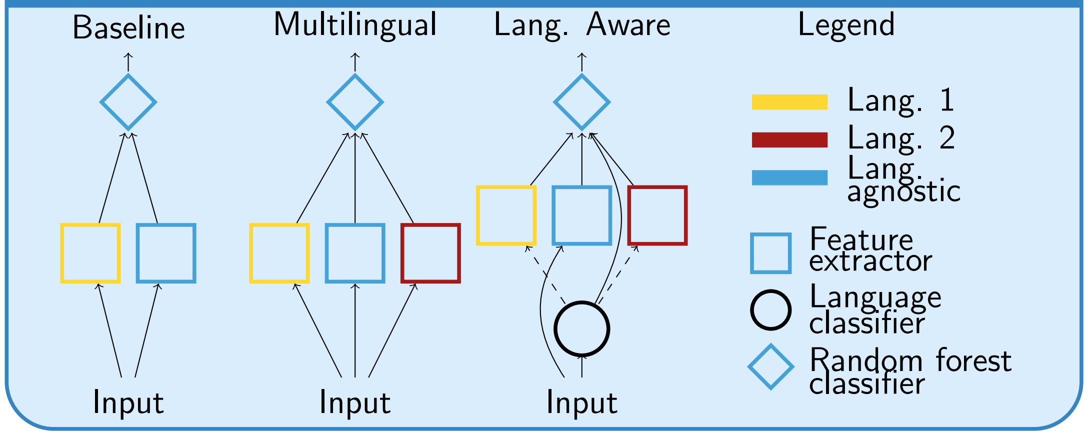
Rob van der Goot and Özlem Çetinoğlu

benimde saprachdiplom vardi ama yinede gittim kursa
????
Benim de Sprachdiplom vardı ama yine de gittim kursa
Ahh! 👍

Data

Raw:	@Erkan1903 nerdee 3 **semester**dayim dha.					
Tok+Anon:	@username	nerdee	3	**semester**dayim	dha	.
Norm	@username	Nerde	3.	**Semester**dayım	daha	.
	OTHER	TR	OTHER	MIXED	TR	OTHER
Seg+CS:	@username Nerde 3. **Semester**§da -yım daha .					

Models



Contributions

- We introduce a publicly available dataset for Tr-De with normalization, language ID and POS layers
- Publicly available normalization models for multiple languages without language-specific heuristics
- Reach new SOTA for normalization on code-switched data
- Show that normalization is beneficial for POS tagging

Code: `https://bitbucket.org/robvanderg/csmonoise`
Data: `https://github.com/ozlemcek/TrDeNormData`

Results

	Normalization		POS
Model	Id-En	Tr-De	Tr-De
LAI	74.03	67.02	60.77
Monolingual (Id/De)	*94.62	76.33	*63.47
Multilingual	94.27	*78.28	*64.06
Language-aware	94.32	77.83	*63.92
Gold	*100.00	*100.00	*67.75