Data Augmentation via Dependency Tree Morphing for Low-Resource Languages

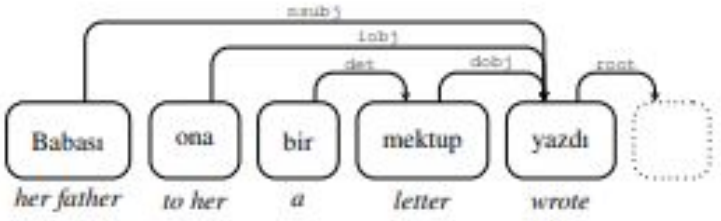
Sahin et al (2019)

NLP reading club 2021/07/15

Key Points

- Proposed augmentation method for low resource languages based on dependency trees
- Technique based on meaning preserving ‘cropping’ and ‘rotating’ of sentences
- Uses POS tagging as the NLP experiment to demonstrate the effectiveness of the method
- Results vary for different languages but in general shows improvement on the POS task as compared to not using any augmentation

Core idea-- dependency tree



(a) Dependency analysis

(1)	Babası	yazdı	*(Her father*	*he-wrote)*
(2)	Ona	yazdı	*(He-wrote*	*to her)*
(3)	Bir mektup	yazdı	*(He-wrote*	*a letter)*

(b) Sentence Cropping

(1)	Babası yazdı bir mektup ona	(SVOIO)
(2)	Yazdı babası ona bir mektup	(VSIOO)
(3)	Bir mektup yazdı babası ona	(OVSIO)
(4)	Ona bir mektup yazdı babası	(IOOVS)

Sentence Cropping

1. Find ‘root’ of sentence (e.g. ‘wrote’)
2. Make smaller sentences using parts that are linked to root (e.g. subject only, object only)

Sentence Rotation

1. Find ‘root’ of sentence (e.g. ‘wrote’)
2. Move sentence fragments around the root

**this method works better for languages where word order matter less

Experiment

- POS tagging task, using a character bidirectional LSTM model
- Decision to use POS tagging to demonstrate the augmentation method as POS tags have many downstream use cases (and also there are less parameters that can affect result as compared to more complex tasks such as classification/sentiment analysis etc)
- Method works better for languages where there are position indicators (ie word ordering less important

Results

				crop			rotate			
#Tokens	**Lang**	**Type**	**Org**	$p = 0.3$	$p = 0.7$	$p = 1$	$p = 0.3$	$p = 0.7$	$p = 1$	*Imp%*
$< 20K$	Lithuanian	*IE, Baltic*	61.51	62.17	66.28	67.64	65.28	66.56	**68.27**	*10.98*
	Belarusian	*IE, Slavic*	83.58	83.87	85.50	85.39	84.33	85.96	**86.11**	*3.03*
	Tamil	*Dravidian*	81.93	81.35	82.78	**84.34**	83.74	83.86	83.61	*2.94*
	Telugu	*Dravidian*	90.78	**90.85**	89.88	90.50	90.36	90.29	89.95	*0.07*
	Coptic	*Egyptian*	**95.17**	94.60	94.74	94.12	95.03	94.65	94.60	*-0.15*
$< 80K$	Irish	*IE, Celtic*	62.75	73.72	75.87	75.42	72.51	**76.35**	76.19	*21.68*
	North Sami	*Uralic, Sami*	86.78	86.04	87.17	87.35	87.85	**88.04**	86.65	*1.45*
	Hungarian	*Uralic, Ugric*	85.94	86.24	86.56	**86.62**	86.49	86.37	86.60	*0.80*
	Vietnamese	*Austro-Asiatic*	75.16	**75.59**	75.32	74.84	75.22	75.15	75.14	*0.57*
	Turkish	*Turkic*	93.49	93.53	93.56	93.89	93.60	93.82	**93.98**	*0.52*
	Greek	*IE, Greek*	95.18	95.32	95.46	**95.54**	95.26	95.22	95.35	*0.38*
	Gothic	*IE, Germanic*	94.38	94.42	94.35	94.44	**94.62**	94.48	94.43	*0.25*
	Old Slavic	*IE, Slavic*	95.36	95.34	95.33	**95.44**	95.17	95.35	94.93	*0.08*
	Afrikaans	*IE, Germanic*	94.91	94.52	94.86	**94.93**	94.73	94.70	94.92	*0.0*
$< 120K$	Latvian	*IE, Baltic*	91.22	91.38	91.77	**91.78**	91.69	91.62	91.76	*0.61*
	Danish	*IE, Germanic*	94.25	94.17	93.96	**94.78**	94.18	94.10	94.21	*0.56*
	Slovak	*IE, Slavic*	91.23	91.17	91.04	91.35	91.53	91.38	**91.58**	*0.38*
	Serbian	*IE, Slavic*	96.14	96.26	96.12	96.17	**96.35**	96.16	96.07	*0.22*
	Ukranian	*IE, Slavic*	94.41	94.33	94.56	94.49	**94.57**	94.38	94.47	*0.17*

Table 1: POS tagging accuracies on UDv2.1 test sets. Best scores are shown with **bold**. Org: Original. p: operation probability. *Imp%*: Improvement over original (Org) by the best model trained with the augmented data.

Resources

- Paper on arxiv: https://arxiv.org/pdf/1903.09460v1.pdf
- Library for experimenting with more NLP augmentation techniques https://github.com/makcedward/nlpaug