Data Augmentation for Morphological Reinflection

University of Colorado **Boulder**

Miikka Silfverberg, Adam Wiemerslage, Ling Liu and Lingshuang Jack Mao

University of Colorado Boulder, Department of Linguistics

firstname.lastname@colorado.edu

Overview

We present the University of Colorado system for data-driven morphological reinflection.

fizzle V PRS PCP	→	fizzling

Task 1: Inflection

- Our system is an RNN Encoder-Decoder with attention.
- The system is specifically geared toward a **low resource setting**.
- During training, we use **data augmentation** in order to counteract data sparsity.
- We additionally use a **copy symbol** to handle input characters which were unseen during training.
- We train an **ensemble** of 10 models and combine them using a **weighted voting scheme**.

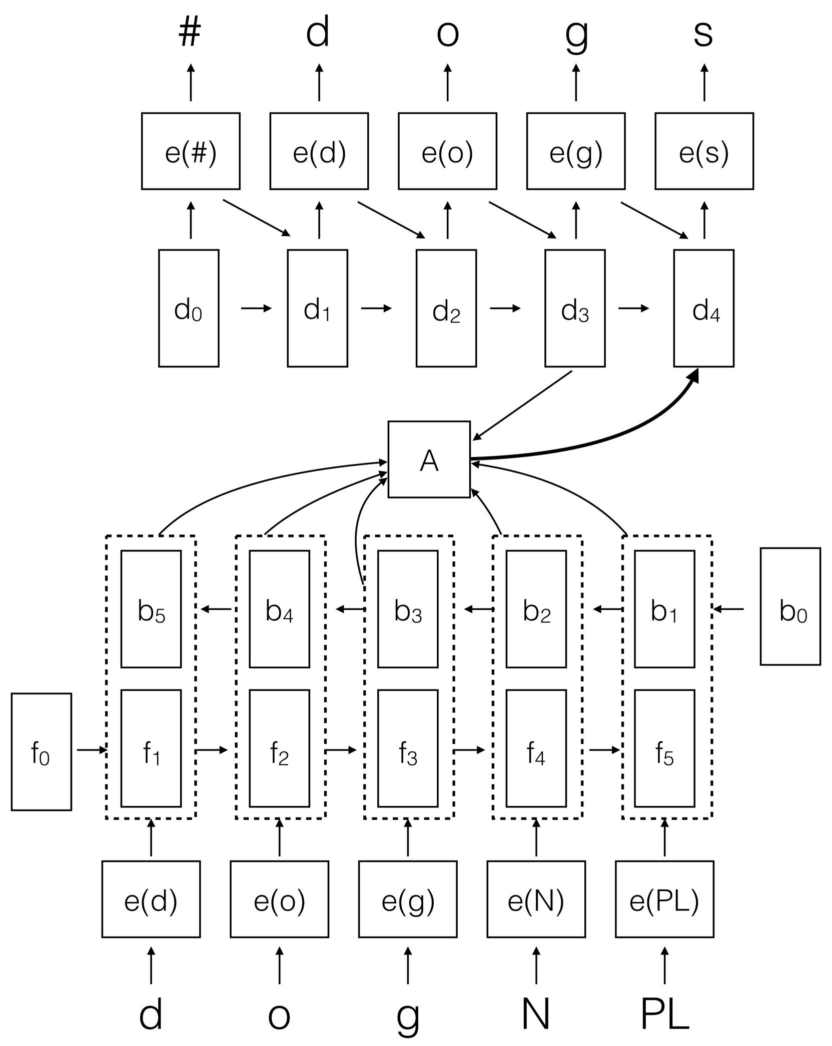
lock	?	V PRS PCP
lock	?	V 3 SG PRS
lock	?	V PST PCP
lock	lock	V NFIN
lock	?	V PST

Task 2: Paradigm completion

System Architecture

Our system is an RNN Encoder-Decoder with a bidirectional encoder and attention.

- We use character embeddings of dimension 32 or 100*.
- For the encoder and decoder, we use 2-layer LSTMs with hidden dimension 32 or 100*.

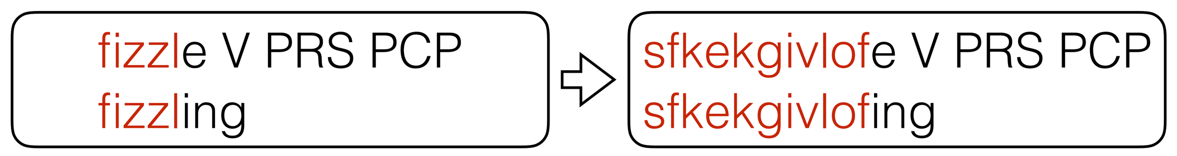


- The system is trained using stochastic gradient descent.
- For each language and data setting, we train ten models combined using a weighted voting scheme.
- We implement the system using DyNet.
- Our code is freely available: `https://github.com/mpsilfve/conll2017`

(*) For ten languages, we use dimension 100 because 32 gave poor results.

Data Augmentation

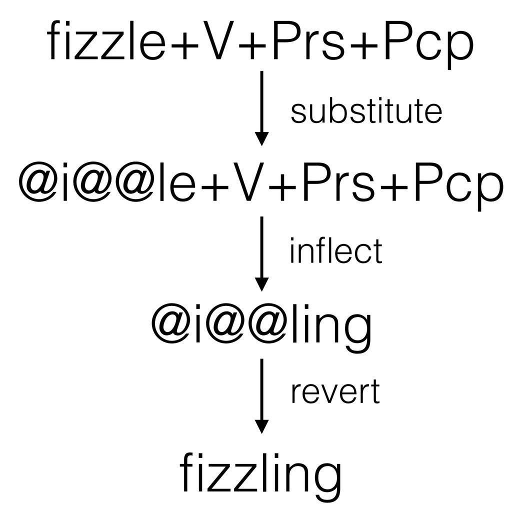
In order to counteract overfitting caused by data sparsity in the low and medium data settings of the shared task, we use data augmentation. That is, we generate new training examples from existing training examples. Without data augmentation, accuracy in the low setting is 0-1% for all languages.



- We replace the stem of an existing training example with a string of random symbols.
- As an approximation of word stems, we use the longest common substring of the lemma and word form.
- $\geq 90\%$ of our training data consists of generated examples.
- This naïve data augmentation gives better results than generating examples from a character n-gram model.

Copy Symbol

The decoder of can only emit characters that were observed in the training data. In a low resource setting, this problem can have a surprisingly large impact on overall accuracy because reinflection will typically fail when input words contain unknown characters.

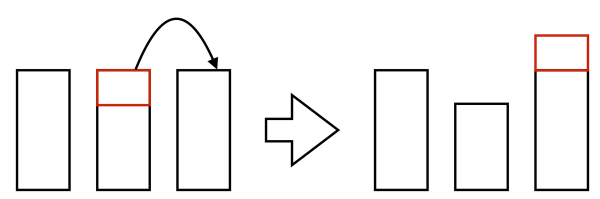


- In order to solve the problem of missing characters, we substitute unknown input characters with *copy symbols* @ during test time (only in Task 2).
- Generated stems with copy symbols are added to the training data during data augmentation. This allows the system to learn to copy the symbols from the input lemma to the output word form.

Weighted Voting

For each language and setting, we train an ensemble of ten models. We apply a *weighted voting scheme* to the model ensemble.

- In weighted voting, each model receives a weight $w_i \in [0, 1]$.
- It then uses this weight to vote for the output candidate that it generated.



- We tune model weights on the development set using a sampling based approach.
- Model weights are initialized evenly. Weight is then iteratively transferred from one model to another depending on the effect on development set accuracy.

Results

	Task 1						Task 2					
	Low		Medium		High		Low		Medium		High	
	RNN	Baseline	RNN	Baseline	RNN	Baseline	RNN	Baseline	RNN	Baseline	RNN	Baseline
Albanian	**31.00**	21.10	**89.40**	66.30	**97.60**	78.90	12.19	12.69	82.17	83.87	86.36	89.46
Arabic	**29.50**	21.80	**73.60**	42.10	**90.40**	50.70	**48.78**	42.85	**63.01**	54.34	**75.54**	55.67
Armenian	**51.30**	35.80	**87.50**	72.70	**96.30**	87.20	75.47	76.18	**86.57**	80.89	**92.04**	86.11
Basque	**4.00**	2.00	**66.00**	2.00	**100.00**	5.00	**1.54**	0.46	**7.35**	4.40	-	-
Bengali	**60.00**	50.00	**95.00**	76.00	**99.00**	81.00	73.38	77.20	19.75	85.86	21.02	87.52
Bulgarian	**57.10**	30.20	**79.90**	72.80	**97.40**	88.80	**35.51**	33.50	49.25	49.58	**78.39**	74.37
Catalan	**66.40**	55.90	**89.50**	84.30	**97.60**	95.50	90.06	94.16	79.69	95.33	90.06	96.03
Czech	**41.90**	39.30	**86.30**	81.50	**92.40**	89.60	16.39	26.56	46.68	56.12	68.36	85.79
Danish	**68.90**	58.40	76.70	78.10	**90.50**	87.80	**53.11**	41.31	64.92	71.15	71.48	75.41
Dutch	51.90	53.60	**74.70**	73.20	**95.60**	87.00	45.57	50.18	60.33	67.71	73.06	78.04
English	**87.80**	80.60	**91.60**	90.90	**95.60**	94.70	**84.40**	76.40	81.60	84.00	84.00	91.60
Estonian	**32.20**	21.50	**74.00**	62.90	**97.10**	78.00	**50.17**	39.81	**61.76**	60.71	**78.29**	77.07
Faroese	**41.20**	30.00	**64.60**	60.60	**85.40**	74.10	46.79	49.78	53.21	59.19	65.62	70.10
Finnish	**15.80**	15.40	**67.20**	43.70	**93.80**	78.20	54.58	60.82	57.14	63.30	**68.78**	68.18
French	**63.00**	61.80	**77.80**	72.50	**88.20**	81.50	84.10	87.09	**85.67**	85.16	89.48	92.63
Georgian	**81.80**	70.50	**92.50**	92.00	**95.40**	93.80	78.86	78.86	81.00	82.42	89.31	90.38
German	**56.60**	54.30	**74.60**	72.10	**89.70**	82.40	68.28	69.83	68.47	70.41	75.82	76.40
Haida	24.00	32.00	**68.00**	56.00	**80.00**	67.00	45.85	47.15	59.63	64.53	-	-
Hebrew	**35.40**	24.70	**77.80**	37.50	**98.50**	54.00	28.83	33.27	**54.89**	42.70	**70.46**	54.09
Hindi	**65.30**	29.10	**93.30**	85.90	**100.00**	93.50	63.62	64.49	61.79	71.11	9.15	96.82
Hungarian	16.00	21.00	**56.20**	42.30	**86.40**	68.50	11.56	17.91	39.68	45.73	**54.95**	53.97
Icelandic	**40.80**	30.30	**67.10**	60.40	**89.10**	76.30	**51.40**	45.79	**56.57**	54.51	63.22	67.36
Irish	**31.30**	30.30	**57.00**	44.00	**88.70**	53.00	26.46	35.95	**44.34**	40.33	**53.28**	47.99
Italian	**56.40**	41.10	**85.90**	71.60	**97.00**	76.90	58.29	66.95	**77.62**	71.86	**89.86**	73.05
Khaling	**10.20**	3.10	**82.90**	17.90	**98.90**	53.70	39.16	42.30	7.53	58.20	**89.64**	79.08
Kurmanji	79.50	82.80	**91.10**	89.10	**94.40**	93.00	65.04	78.43	87.48	88.35	**93.74**	93.39
Latin	**19.30**	16.00	**46.10**	37.60	**80.50**	47.60	22.55	24.45	38.07	39.53	**50.51**	47.58
Latvian	62.60	64.20	**86.00**	85.70	**94.60**	92.10	**74.78**	68.88	**81.99**	79.97	**88.47**	86.46
Lithuanian	19.80	23.30	**58.40**	52.20	**92.90**	64.20	28.19	38.27	61.73	65.92	**64.14**	60.57
Lower Sorbian	**52.30**	33.80	**83.60**	70.80	**95.40**	86.40	27.22	38.20	**71.16**	65.92	80.15	82.27
Macedonian	**59.60**	52.10	**90.40**	83.60	**95.20**	92.10	14.74	42.49	83.12	86.41	**92.56**	89.70
Navajo	11.70	19.00	**40.50**	33.50	**83.10**	37.80	**19.73**	0.00	**32.60**	0.00	**46.30**	0.00
Northern Sami	**18.70**	16.20	**57.00**	37.00	**96.10**	64.00	15.32	15.62	27.93	31.43	**54.61**	45.68
Norwegian Bokmal	**73.80**	67.80	**80.80**	80.70	**91.50**	91.00	**49.06**	41.51	**57.23**	50.94	**70.44**	67.92
Norwegian Nynorsk	**50.50**	49.60	**62.50**	61.10	**87.50**	76.90	39.88	42.33	56.44	60.74	60.74	64.42
Persian	**38.30**	24.50	**86.10**	62.30	**99.50**	79.00	**84.69**	73.42	**94.47**	78.29	25.09	76.44
Polish	**43.70**	41.30	**78.00**	74.00	**90.90**	88.00	55.19	56.72	79.77	80.28	83.10	90.27
Portuguese	**68.40**	63.60	**94.70**	93.40	**99.30**	98.10	89.94	91.71	92.10	95.29	36.23	96.19
Quechua	**30.60**	16.40	**88.20**	70.30	90.30	95.40	79.84	91.33	0.04	91.34	64.45	89.13
Romanian	43.10	44.80	**77.40**	69.40	**85.50**	79.80	10.36	14.20	60.80	61.54	75.00	78.99
Russian	**45.90**	45.60	**81.90**	75.90	**90.80**	85.70	36.66	40.18	82.21	82.98	**87.42**	85.58
Scottish Gaelic	**56.00**	44.00	**74.00**	48.00	-	-	29.96	44.13	**44.53**	41.30	-	-
Serbo-Croatian	**39.20**	18.40	**83.30**	64.50	**92.10**	84.60	27.90	30.07	36.59	36.84	74.40	77.66
Slovak	**46.70**	42.40	**78.00**	72.30	**89.30**	83.30	38.86	44.39	59.00	59.89	68.27	69.16
Slovene	**60.20**	49.00	**86.30**	82.20	**95.80**	88.90	52.15	57.74	**69.15**	67.87	**77.18**	76.48
Sorani	**27.10**	19.30	**71.50**	51.70	**89.10**	63.60	43.53	54.78	67.96	68.30	8.88	72.27
Spanish	**63.60**	57.10	**89.50**	84.70	**96.80**	90.70	79.58	79.75	86.53	92.18	34.63	93.58
Swedish	**60.40**	54.20	**76.30**	75.70	**87.60**	85.40	31.47	43.53	**59.41**	57.35	70.29	78.24
Turkish	**19.70**	14.10	**66.60**	32.90	**96.40**	72.60	**34.89**	20.93	**76.05**	73.26	31.08	85.05
Ukrainian	**50.40**	43.90	**79.70**	72.80	**90.20**	85.40	32.38	43.97	65.71	67.14	72.54	73.97
Urdu	**64.60**	31.70	**96.10**	87.50	**98.30**	96.50	79.56	80.59	67.23	81.02	90.79	95.33
Welsh	**53.00**	22.00	**82.00**	56.00	**98.00**	69.00	**82.72**	51.67	79.05	82.80	81.66	85.25
AVG	**45.74**	37.90	**77.60**	64.70	**92.97**	77.81	47.90	49.63	60.94	65.20	65.11	73.11

Conclusions

- The system delivers substantial improvements in accuracy over a non-neural baseline for most of the 52 languages in Task 1.
- Due to data augmentation, it improves upon the baseline even in the extreme low resource setting of a mere 100 training examples.
- Unfortunately, the system still performs poorly in Task 2. One possible cause for this is overfitting due to insufficient variation in the training set. Additionally, time constraints prevented us from training a model ensemble for Task 2.
- The present work employs a very naïve form of data augmentation. A new training example is created from an existing one by replacing the longest common substring of the stem and word form with a sequence of random characters from the training data. We also tried to use more sophisticated language models for generating the examples. Interestingly, this failed to bring improvements.
- In conclusion, we have demonstrated that an RNN Encoder-Decoder system can be applied to morphological reinflection even in a low resource setting.