The NYU-CUBoulder Systems for SIGMORPHON 2020 Task 0 and Task 2

Assaf Singer
New York University
as12152@nyu.edu

Katharina Kann
University of Colorado Boulder
katharina.kann@colorado.edu

1. Overview

- Morphological Inflection:
  - Lemma + Morphological tag -> Inflected form
- Task 0: Typologically Diverse Morphological Inflection
  - 45 development languages - model development
  - 45 surprise languages - unseen during model development
- Task 2: Unsupervised Morphological Paradigm Completion
  - Input: tokenized Bible, list of lemmas
  - Output: complete paradigms

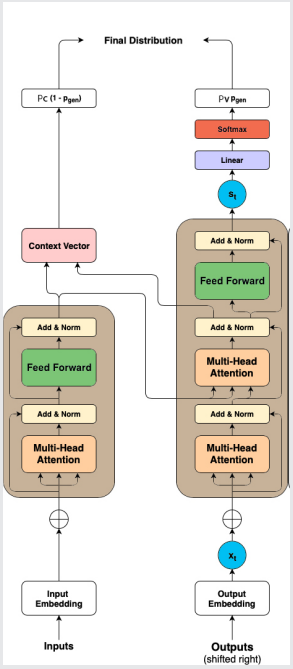
Lemma	Features	Inflected form
hug	V;PST	hugged
seel	V;3;SG;PRS	seels

2. Methods

- Transformer (Vaswani et al., 2017)
- Pointer-Generator Transformer
- Multitask Training:
  Morphological inflection and morphological reinflection

raw	grip	grips	V;SG;3;PRS
	grip	gripped	V;PST
generated	grips	grip	V;LEMMA
	grips	gripped	V;PST
	gripped	grip	V;LEMMA

- Hallucination Pretraining
  - Pretrain low-resource languages on hallucinated training set
- Task 2: Model Description
  - Official baseline (Jin et al., 2020) and Task 0 systems



3. Experiments

- Task 0:
  - Best: NYU-CUBoulder-3. 6th out of 23 systems

System:	Sub-1	Sub-2	Sub-3	Sub-4	Base
Development Set					
Low	**88.71**	88.02	84.90	84.07	-
Other	90.46	90.63	90.20	**90.94**	-
All	**90.06**	90.02	88.96	89.34	-
Test Set					
Low	84.8	84.8	85.5	83.9	**89.77**
Other	89.7	89.8	89.8	90.2	**92.43**
All	88.6	88.7	88.8	88.8	**91.81**

Table 2: Macro-averaged results over all languages on the official development and test sets for Task 0. Low=languages with less than 1000 train instances, Other=all other languages, All=all languages.

- Task 2:
  - NYU-CUBoulder-3: 0.0007% on Basque, highest in shared task
  - Best: NYU-CUBoulder-2. 2nd out of 7 systems

System	Baseline 1		Baseline 2		Sub-1		Sub-2		Sub-3	
Test Set										
	slots	macro	slots	macro	slots	macro	slots	macro	slots	macro
Basque	30	0.0006	27	0.0006	30	0.0005	30	0.0005	30	**0.0007**
Bulgarian	35	0.283	34	**0.3169**	35	0.2769	35	0.2894	35	0.2789
English	4	0.656	4	**0.662**	4	0.502	4	0.528	4	0.512
Finnish	21	0.0533	21	**0.055**	21	0.0536	21	0.0547	21	0.0535
German	9	0.2835	9	**0.29**	9	0.273	9	0.2735	9	0.2735
Kannada	172	0.1549	172	**0.1512**	172	0.111	172	0.1116	172	0.111
Navajo	3	0.0323	3	**0.0327**	3	0.004	3	0.0043	3	0.0043
Spanish	29	0.2296	29	**0.2367**	29	0.2039	29	0.2056	29	0.203
Turkish	104	0.1421	104	**0.1553**	104	0.1488	104	0.1539	104	0.1513
All		0.2039		**0.2112**		0.1749		0.1802		0.1765

4. Low-resource experiment

- Overview
  - Investigate effect of copy mechanism
  - Sampling 100 instances from low-resource languages
  - Pointer-generator transformer vs vanilla transformer
- Results
  - Pointer-generator transformer 4.46% higher average accuracy than vanilla transformer
  - Only 2.45% lower than state of the art (Makarov et al., 2017)

System	Trm	Trm-PG	Baseline
All	63.06	67.61	**70.06**

Table 4: Results on the official development data for our low-resource experiment. Trm=Vanilla transformer, Trm-PG=Pointer-generator transformer, Baseline=neural transducer by Makarov and Clematide (2018).

5. Ablation Studies

- Copy Mechanism
  - Original dataset
    +0.68% accuracy on low resource
    -0.11% accuracy on high resource
  - With multitask:
    +0.06-0.16% accuracy
- Multitask Training
  - Pointer-generator transformer:
    -1.8-2.03% accuracy
  - Transformer:
    -1.67-2.32% accuracy
  - Relatively large train sets
- Hallucination Pretraining
  - +1.85% accuracy in low-resource

Model:	1	2	3	4	5
Copy	✓	✓			✓
Multitask Train	✓		✓		✓
Hallucination	✓	✓	✓	✓	

Model:	1	2	3	4	5
Development Set					
Low	88.20	**90.00**	87.52	89.84	86.35
Other	90.63	**92.66**	90.93	92.60	90.63
All	90.02	**92.04**	90.13	91.96	89.63