The UniMelb Submission to the SIGMORPHON 2020 Shared Task 0: Typologically Diverse Morphological Inflection

Andrei Shcherbakov

andreas@softwareengineer.pro

The University of Melbourne

Flexica01: non-neural, alignment based

Lemma-to-inflected form transformation are generated directly by the following simple process:

Step 1. Find maximal continuous matches between lemma and inflected form.

Example: `understand` → `understood`

Extracted rule: `\0an\1` → `\0oo\1`, where `\0=underst` and `\1=d` are groups.

Step 2. Starting with previously generated transformation pattern(s), generate a set of patterns more specific to a given training word by treating a limited number of characters as concrete (i.e. standing outside any group).

For the example from previous step and a limit of one character: `\0an\1` → `\0oo\1`; `u\0an\1` → `u\0oo\1`; `\0n\1an\2` → `\0n\1oo\2`, `... (3 more)`, `\0s\1an\2` → `\0s\1oo\2`, `\0tan\1` → `\0too\1`, `\0and` → `\0ood`.

When predicting a form, score matching candidate patterns using the following three components:

⇒ A (squashed) **frequency** f of transformation occurrence in a training set;

⇒ The **diversity** d of marginal (the first one and the last one) letters in groups as they occurred in different fits of a given transformation found in the training set.

⇒ **Specificity** s which here means the number of concrete characters in the pattern (without counting characters falling into groups).

We were using the following empirical formula:

$$G = \frac{1}{2}\log_2 f + 6\log_2 d + 12s$$

Near-misses (the second scored transform was correct)

deu	Kation	Kations	N;GEN;SG
eng	upswell	upswollen	V.PTCP;PST
est	põlema	olime põlenud	V;PRF;COND;PL;1;POS;PRS;ACT
isl	stelkur	stelkinn	N;NOM;DEF;SG
nob	pioner	pionerer	N;NDEF;PL
udm	*patent*	*patenntem*	N;LGSPEC_ATTR;LGSPEC1

Flexica02: Hard attention, multilingual (family-based)

This neural system is based on hard monotonic attention model proposed in [Aharoni and Goldberg(2017)], with the same loss function, but with the following differences:

⇒ We combined all the languages belonging to a given family into a single dataset, having added two extra features such as language and genus.

⇒ We used maximal continuous sub-string search to organize alignment between lemma and its inflected form in order to advance hard attention state (abolishing one-by-one alignment of mismatching characters).

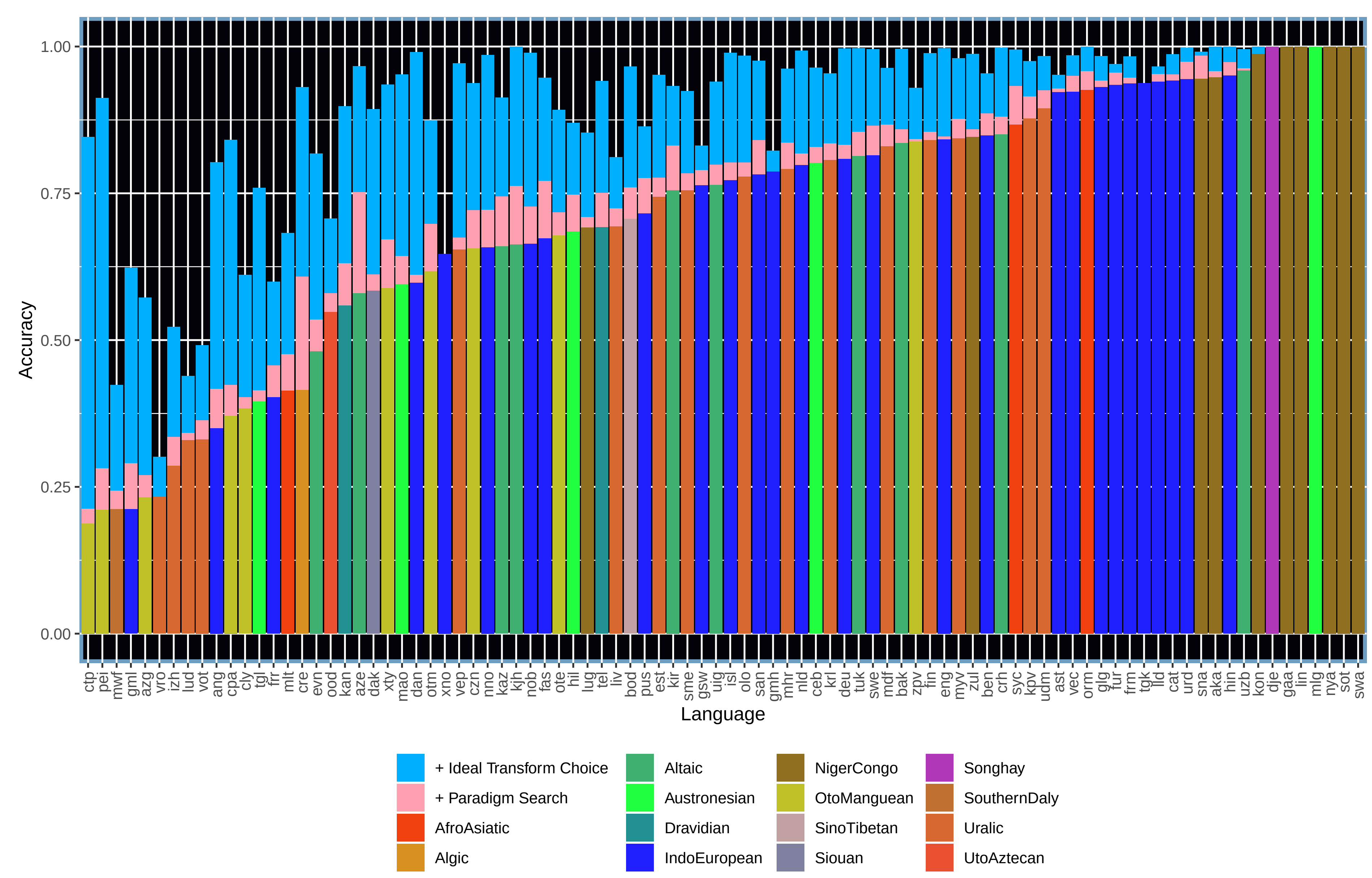
Flexica03: Adding hallucinated data

Inspired by [Anastasopoulos and Neubig(2019)]. We added 200 samples per language per part-of-speech (POS) in order to produce hallucinated inflection samples that look like real. We reused the predictor from `flexica01`. We also enriched the model with word-generator [Shcherbakov et al.(2016)Shcherbakov, Vylomova, and Thieberger], `http://regexus.com/wg.php` to produce more phonotactically plausible forms: 1) Word generator trained on inflected forms for a given POS produces samples of hallucinated inflected forms (without distinction of grammatical features); 2) The reverse `flexica01` predictor produces tag+lemma for each hallucinated inflected form. Accuracy was significantly improved in low-resource languages (such as Maori, Zarma, Tajik, Anglo-Norman, Middle High/Low German).

Conclusion

We proposed and tested (1) multilingual training, and (2) pattern-based hallucinated inflections as possible enhancements of sequence-to-sequence morphology modeling for diverse low-resource languages. We also developed a simple non-neural approach based on multi-variant search of common inflection patterns.

Flexica01: results



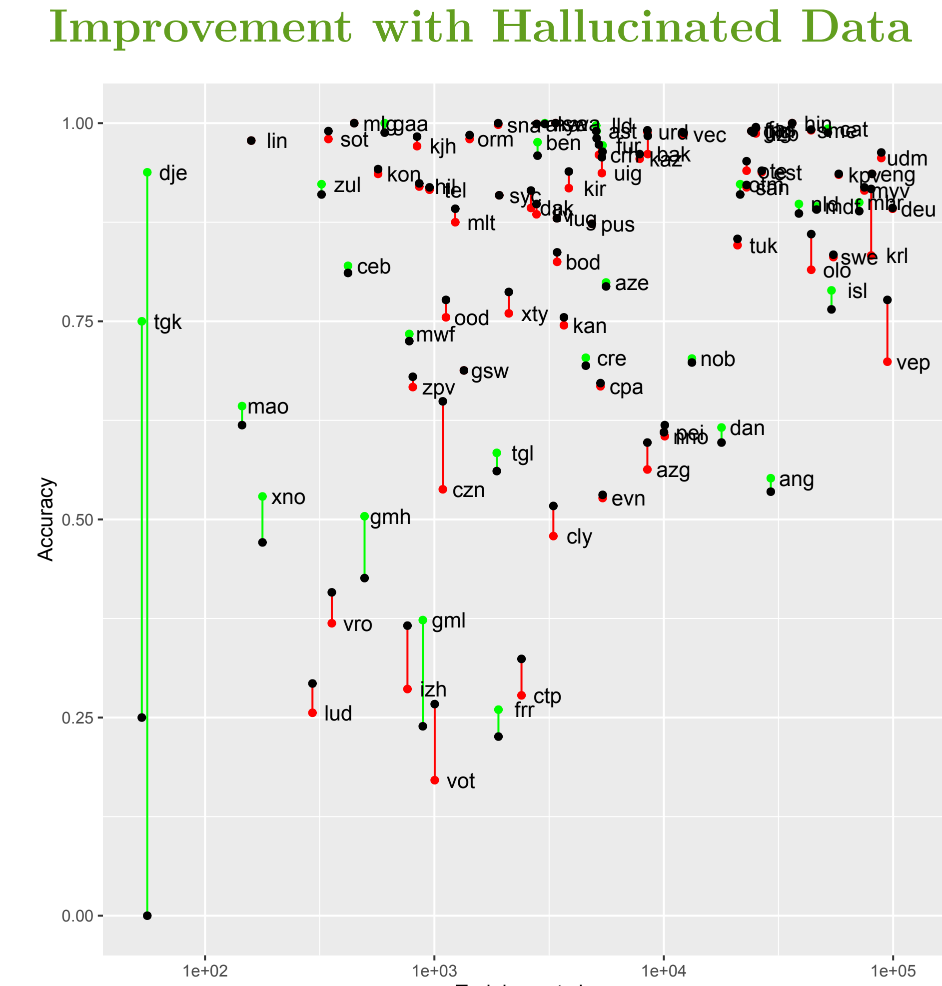
We additionally show the accuracy that would be achieved in a case of ideal selection criteria (labelled as + Ideal Transform Choice category). We also roughly measured potential improvement that may arise from considering correlations between inflection patterns for different grammatical forms of a single lemma.

Only flexica01 got it right

eng	shine	shone	V.PTCP;PST
eng	overwork	overwrought	V.PTCP;PST
eng	help	holpen	V.PTCP;PST
eng	belive	belove	V;PST
eng	arise	arose	V;PST
eng	belight	belit	V.PTCP;PST
eng	dwell	dwelt	V.PTCP;PST
eng	bespit	bespat	V;PST
eng	snatch	snaught	V.PTCP;PST
eng	stink	stank	V;PST
eng	uplight	uplit	V.PTCP;PST
dak	Dakota	uDakotapi	V;PL;1;PRS
krl	pezieie	ei pezieeta	V;IND;PL;3;NEG;PRS
isl	aðalkirkja	aðalkirknanna	N;GEN;DEF;PL
isl	hagskælingur	hagskælinginn	N;NOM;DEF;SG
mhr	*popo*	*popolam*	N;HUM;SIM;SG;PSS1S
nob	kronprinsesse	kronprinsessa	N;DEF;SG
nno	byste	bystar	N;NDEF;P
udm	*million*	*million'em*	N;LGSPEC1
olo	buabo	buaban	N;GEN;SG
swe	hålla inne	innehållande	V.PTCP;PRS
vep	pugetas	pugeiihe	V;COND;PL;3;POS;PRS

transliterated words are given in *italic*

Improvement with Hallucinated Data



Flexica03: Generating Hallucinated Data