Contextualization of Morphological Inflection

Ekaterina Vylomova Ryan Cotterell Tim Baldwin
Trevor Cohn Jason Eisner

SIGMORPHON Shared Task 2016--2019

PLAY + PRESENT PARTICIPLE ---> playing
played + PRESENT PARTICIPLE ---> playing

Inflectional Paradigm Table:

Lemma	Tag	Form
RUN	PAST	ran
RUN	PRES;1SG	run
RUN	PRES;2SG	run
RUN	PRES;3SG	runs
RUN	PRES;PL	run
RUN	PART	running

2018:~96\% accuracy on avg. in high-resource setting

Contextualization

The tags must be inferred from the context!

Two cats are ???
TWO/NUM CAT/N+PL BE/AUX+PRES+3PL PLAY

Two cats are ???
__ __ __ PLAY

Hybrid Model (Structured-Unstructed)

$$p(\mathbf{w}, \mathbf{m} \mid \boldsymbol{\ell}) = \left(\prod_{i=1}^{n} p(w_i \mid \ell_i, m_i)\right) p(\mathbf{m} \mid \boldsymbol{\ell})$$

Hybrid (Structured-Unstructed) Model

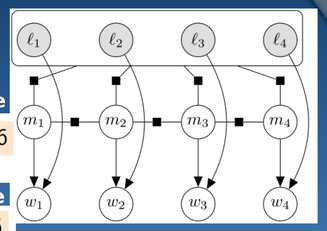
lemmatized sequence

predicted tag sequence

Lample et al., 2016
$p(\mathbf{m} \mid \boldsymbol{\ell})$ Neural CRF

predicted form sequence

Rastogi et al., 2016
Aharoni et al., 2017
$p(w_i \mid \ell_i, m_i)$ Word Emission



Q2: How Well is Agreement Captured?

Adjective-Noun (AMod) is captured quite well

Verb-Noun(Subject--Verb) is more challenging since agreement categories can vary depending on tense, e.g. Ru: Person+Number in present vs. Gender in past, singular

PL (Flex): Jen**ia** daje Masz**y** ciekaw**a** książk**ę**
NOM.SG Pres.3Sg DAT ACC.Sg.Fem ACC.Sg

PL (Flex): Jen**ia** dał Masz**y** ciekaw**a** książk**ę**
NOM.SG Past..Sg.Masc DAT ACC.Sg.Fem ACC.Sg

AMod (ADJ-NOUN): Case, Number, Gender

NSubj (VERB-NOUN): Present Tense > Person, Number
Past Tense > Gender, Number

General-purpose inference of agreement categories is still a challenging task!

Q1: Do Linguistic Features Help?

YES! They Do!
Most systems that make use of morphological tags outperform the ``Direct'' baseline on most languages

Joint prediction of tags and forms further improves the results

Q5: Does morphological complexity impact empirical performance?

Yes, it does!

Performance drops in languages with rich case systems such as Slavic and Uralic

The model needs to learn which grammatical categories should be in agreement

Language	tag	form				
	JOINT	GOLD	JOINT	DIRECT	SM	CPH
Bulgarian	81.6	91.9	**78.8**	71.5	77.1	76.9
English	89.6	95.6	**90.4**	86.8	86.5	86.7
Basque	66.6	82.2	61.1	59.7	**61.2**	60.2
Finnish	66.0	86.5	**59.3**	51.2	56.6	56.4
Gaelic	68.3	84.5	**69.5**	64.5	68.9	66.9
Hindi	85.3	88.3	81.4	85.4	86.8	**87.5**
Italian	92.3	85.1	80.4	85.2	88.7	**90.5**
Latin	82.6	89.7	**75.7**	71.4	74.2	74.9
Polish	71.9	96.1	**74.8**	61.8	72.4	70.2
Swedish	81.9	96.0	**82.5**	75.4	78.4	80.9

Q4: Which language is least affected by lemmatization?

Word Order vs. Morphology

Most information on roles and dependencies is expressed non-morphologically, e.g. in word order or by prepositions:

EN (SVO): Kim gives Sandy an interesting book
Subject IObject DObject

PL (Flex): Jen**ia** daje Masz**y** ciekaw**a** książk**ę**
NOM DAT ACC.Sg.Fem ACC.Sg
== Masz**y** daje Jen**ia** ciekaw**ą** książk**ę**
!= Jen**ie** daje Masz**a** ciekaw**ą** książk**ę**

EN: Kim gives Sandy an interesting book ---->
KIM GIVE SANDY AN INTERESTING BOOK
SVO/Roles are still there

PL: Jenia daje Maszy ciekawą książkę ----->
JENIA DAWAĆ MASZA CIEKAWY KSIĄŻKA
Flexible/Roles are partially lost

Q3: Where Does Most Uncertainty Come From?

Inherent and Contextual Morphological Categories

Contextual categories participate in agreement: adjective number, case, gender, verbal gender, etc.

Inherent express the speaker's intentions: noun number, verbal tense

Most uncertainty comes from inherent categories!

E.g., Tense can be both Present and Past:

JOHN GIVE MARY AN INTERESTING BOOK --->

Jonh gives Mary an interesting book +
John gave Mary an interesting book +

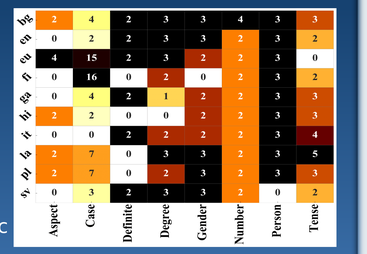
whereas Contextual comes from dependencies (adjective gender, number, case):

ciekaw**a** książk**ę**
ACC.Sg.Fem ACC.Sg +
ciekawemu książk**ę**
DAT.Sg.MA ACC.Sg -

Often such categories must be inferred!

Number of values of an inflectional feature

Bulgarian (bg), Slavic
English (en), Germanic
Basque (eu), Isolate
Finnish (fi), Uralic
Gaelic (ga), Celtic
Hindi (hi), Indic
Italian (it), Romance
Latin (la), Romance
Polish (pl), Slavic
Swedish (sv), Germanic



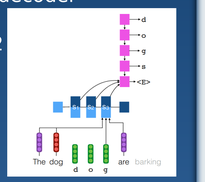
Languages differ in what is explicitly morphosyntactically marked, and how

Experiments

Data: Universal Dependencies v1.2

Baselines:

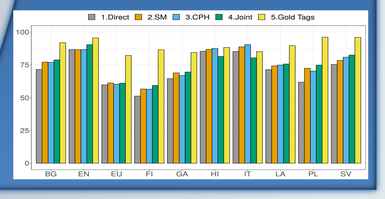
1) SM (Cotterell, 2018): biLSTM encoder--decoder with context window of size 2



2) CPH (Kementchedjhieva et al.,2018): biLSTM encoder--decoder with no context window size restrictions
-- input = concat (full context, lemma, tags, char-level center lemma)
-- also predicts target tags as an auxiliary task

3) DIRECT: more basic model that relies only on forms and lemmas

Accuracy:



Conclusion

1) we proposed a method for contextual inflection using a hybrid architecture;
2) consistent improvements over state of the art;
3) contextual inflection can be a highly challenging task;
4) inclusion of morphological features prediction is an important element;
5) two types of morphological categories, contextual and inherent, in which the former relies on agreement and the latter comes from a speaker's intention.

Future Directions

1)Evaluation of grammaticality
2) Data de-biasing (En->Ru): *smart student --> umnyj.Nom.Masc.Sg student.Nom.Sg*
augment with *umnaja.Nom.Fem.Sg studentka.Nom.Fem.Sg*

THE UNIVERSITY OF MELBOURNE

JOHNS HOPKINS UNIVERSITY