Generating Erroneous Text for Natural Language Corrections

Teodor-Mihai Cotet

Thesis advisor:

Prof. dr. ing. Mihai Dascălu

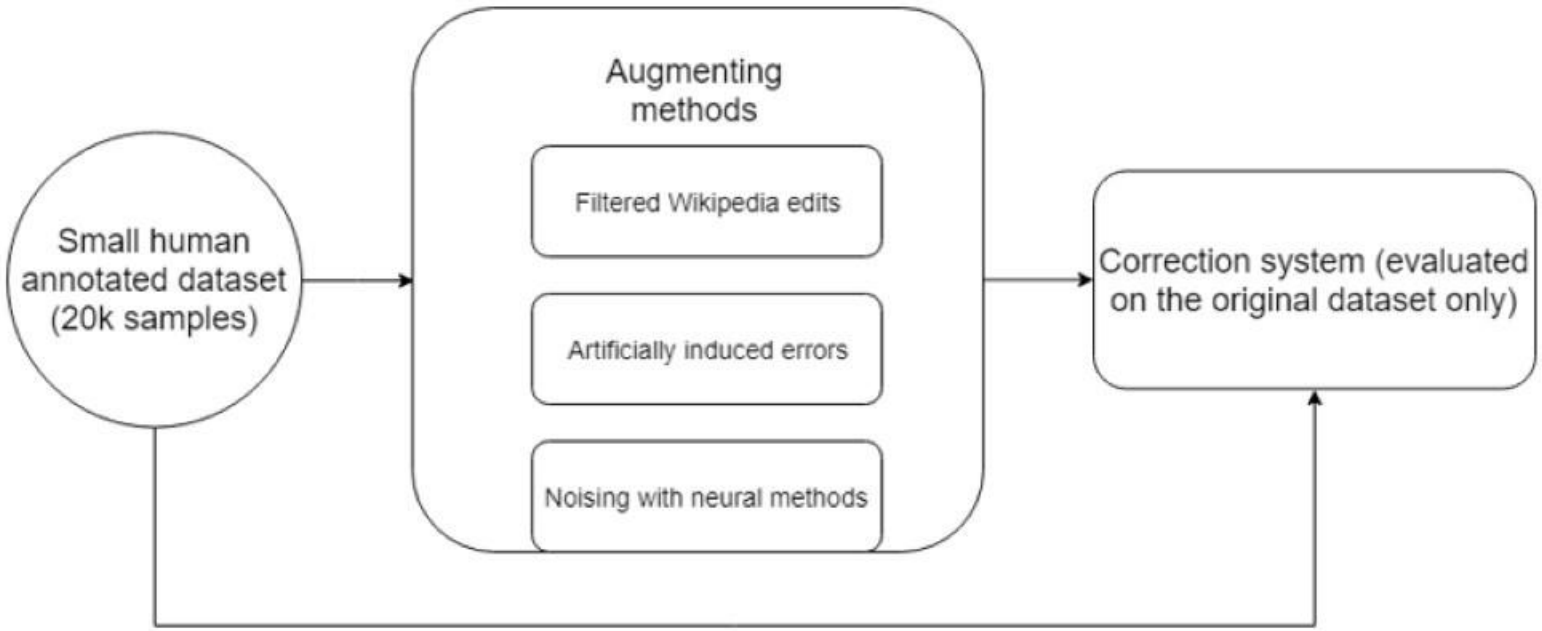
Problem

- Scarce annotated resources for the task of grammatical error correction for most of the languages
- Neural networks requires lots of samples to train -> problem even for English

Problem


New Message
Recipients
Subject
Miniștri vroiau să îi prindă defapt cei care dintro dată gresisera pe fața.

General pipeline



Wikipedia edits

Before editing	After editing
Este considerat a fi cel mai mare dramaturg român și unul dintre cei mai importanți scriitori români.	George Călinescu *îl considera* a fi cel mai mare dramaturg român și unul dintre cei mai importanți scriitori români.
Înainte de a-și publica pamfletul în broșură, Caragiale *a trimis* [...]	Înainte de a-și publica pamfletul în broșură, Caragiale *trimisese* primul capitol [...]
[...] iar în 1892 și-a exprimat intenția de a se *expatria la* [...]	[...] și-a exprimat intenția de a se *exila la* [...]
A publicat în revista literară *bimensuală* Convorbiri [...]	A publicat în revista literară *bilunară* Convorbiri [...]
A fost numit, *cu* decret regal [...]	A fost numit, *prin* decret regal [...]
[...] *a avut loc prima reprezentație* a piesei [...]	[...] *premiera* piesei [...]
În *aceste* împrejurări și-a manifestat calitățile *sale* de [...]	În *acele* împrejurări și-a manifestat calitățile de [...]
[...] *în presa vremii* [...]	[...] in presa *contemporană* [...]

Artificially induced errors

- Developed a tool to manipulate POS tags features
- Examples:
- Primul (word form):
  - RoCaseEnum: RoCaseEnum.ACC, RoCaseEnum.NOM (form of the noun)
  - RoDefiniteEnum: RoDefiniteEnum.DEF (definite form)
  - RoGenderEnum: RoGenderEnum.MASC (masculine form)
  - RoNumberEnum: RoNumberEnum.SING (number in singular form)
  - RoNumTypeEnum: RoNumTypeEnum.ORD (numeral type showing order, not comparison)
  - RoNumFormEnum: RoNumFormEnum.WORD (numeral form described through a word, not an Arab or Roman digit)

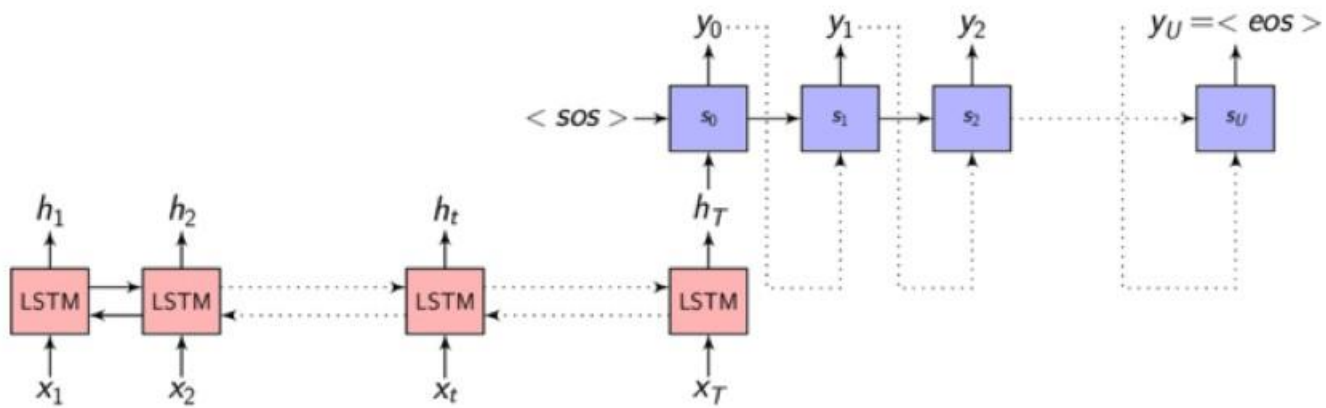
Artificially induced errors

- văzuse:
  - RoMoodEnum: RoMoodEnum.IND
  - RoNumberEnum: RoNumberEnum.SING (singular)
  - RoPersonEnum: RoPersonEnum.THIRD (third person)
  - RoTenseEnum: RoTenseEnum.PQP (past perfect tense equivalent in Romanian)
  - RoVerbFormEnum: RoVerbFormEnum.FIN (general verb form)

Artificially induced errors

Mistake type	Description	Example
Tp	Typos qwerty	Am mers la [**mahazin** -> magazine].
Diac	Missing/extra diacritics	[**Scoala** -> Școala] din zare.
Vt	Verb tense	[Merg -> Am mers] la magazin ieri].
Vf	Verb form	Masinile sunt [**prevazut** -> prevazute] cu suspensii.
SVA	Subject-verb agreement	El [**mergem** -> merge] la magazin.
Pre	Prepositions	Acest țel oarecum a început să piară [**ca și** -> ca] importanță. Spre unul [**din** -> dintre] directorii firmei.
Vir	Missing/extra commas	[**Astfel** -> Astfel,] s-a demonstrat teorema. Soră-mea cea mare [**, ->**] m-a ținut odată în casă
FC	Spelling	[**Copii** -> Copiii] merg la scoala prea mult. De mult [**vroiam -> voiam**]. 1000 de persoane care ar putea [**știi ->** sti] ceva.
Cr	Extra/missing dash	[**Dintro** -> Dintr-o] greseala am plecat.
Sp	Spacing	**deasemenea** vs de asemenea, numai vs numai
Gen	Gender	Fata [**frumos** -> frumoasă] a plecat.
Case	Noun case	Am iesit la [**pensii** → pensie]
PI	Strongness form of pronoun	Ea [**însumi** ->însăși] a câștigat.
AcP	Owner agreement	Băiatul [**al** → a] cărui carte
Cap	Capitalization errors	[**Ioana** → Ioana] merge la magazin.

Inducing noise with neural methods



Modified beam search (random noising - adding $r * \beta_{random}$ to beam candidates probabilities)

RESULTS FOR DETECTING AND CORRECTING WORD LEVEL ERRORS

Corpora statistics

corpus	sentences/samples	Dictionary (# words)	size (MB)
inflected	650k	150k	150MB
Typos	550k	150k	152MB

RESULTS FOR DETECTING AND CORRECTING WORD LEVEL ERRORS

Results for correction at the word level.

Task	Model	Epochs	Val Accuracy
inflected	word	33	0.519
inflected	char-GRU	46	0.518
inflected	word + sent-GRU + char-GRU	15	**0.610**
typos	word	17	0.622
typos	char-GRU	35	0.733
typos	word + sent-GRU + char-GRU	18	**0.738**

RESULTS FOR DETECTING AND CORRECTING WORD LEVEL ERRORS

Detection statistics

Task	Model	Epochs	Val Accuracy	Precision	Recall	$F_{0.5}$
inflected	word	25	0.820	0.782	0.843	0.794
inflected	word + sent-GRU + char-GRU	15	**0.927**	**0.909**	0.943	0.916

RESULTS FOR DETECTING AND CORRECTING WORD LEVEL ERRORS

Classifying by different threshold

Threshold	Precision	Recall	$F_{0.5}$
0.50	0.909	**0.943**	0.916
0.40	0.926	0.933	0.928
0.25	0.947	0.915	0.940
0.10	**0.971**	0.879	**0.951**

Examples from the detection system (for threshold 0.5).
Text in bold marks incorrect initial text

Sentence	Type
— vrea/**vrem** oare old shatterhand să mă abată de la datorie ?	tp
vreau și eu pedepsirea/**pedepsire** ucigașilor .	tp
datinile îmi poruncesc să rămîn lîngă morții mei/**meu** pînă vor fi înmormîntați .	tp
— și cînd va/**vruseră** fi înmormîntarea ?	tp
pentru că nu pot/**putea** să - l urmăresc personal , sarcina va reveni altcuiva .	tp
Acum/**acu'**, cînd era vorba de chestiuni concrete , winnetou deveni calm ca de obicei.	tp
— acesta e numele lui/**el** adevărat ? el lui	tp
— aveti/**avuseși** cunoscuți prin apropiere , poate în vreun fort ?	tp
— voiam/**voi** să ... să ne luăm după ...	tp
- am/**avea** ghicit însă gîndul și l - am întrebat :	tp
sau aveati/**avea** de gînd să - i atacați la întoarcere și ...	tp
trecuse mult de amiază și sam mă întîmpină intrigat/**intrigată**	tp
se lăsă/**lăsa** o tăcere adîncă .	tp
apoi v - ați ascuns în **pădurice** și ne - ați observat de susul , din copaci . pădurice	fp
sper că voi răspunde cu cinsteo **încredenii** pe care fratele meu roșu mi - o acordă .	fp
gîndul **acesta** îmi grăbi pasul .	fp
n - avea izbutit din primul moment , pămîntul **fiind** aici încă prea dur .	fp
cei trei **cai** se mai afla în pădurice .	fp
trecuse mult de amiază și sam mă întîmpină intrigată :	fp
avea , nu **zic** ba .	fp
zarvă nu duce la nimic	fp
gîndul acesta îmi grăbi pasul/**pasii** .	fn
În loc de răspuns , i - am chemat pe apasi/**apaș** la mine .	fn
așadar , ce aduce tovărășia mea : viață sau moarte/**morții**	fn
Draga/**dragă** de ea , frumoasa și buna fată indiană	fn
— mai tacă/**tăcu** - ți gura cu greenhorn - ul dumitale și lasă glumele !	fn
santer poate să treacă fie peste munți/**munte** , fie printre munți , cum îi convine .	fn

Future work

- The focus will be on generating a good copora for Romanian using as few language specific features as possible

- Evaluate how good the corpus is and its methods of generation by implementing a correction system

Thank you!