Text Normalization Challenge - English Language

Presented by
Dahou Abdelhalim Hafedh
Pavithra POORNACHANDRAN
Tahani FENNIR

UNIVERSITÉ DE LORRAINE

Institut des sciences du Digital Management & Cognition

November 24th, 2020

Table of Content

Introduction

- Definition of text normalization
- Examples

Data Analysis

Previous Work

Our state

Introduction:

- Text normalization is the process of transforming a text into a standard form.

Examples:

→ Transform word numerals into numbers (eg.: ‘twenty three’→‘23’)
→ Acronym normalization (e.g.: ‘US’→‘United States’/‘U.S.A’)
→ Removal or substitution of special characters(e.g.: remove hashtags).
→ Removal of duplicate whitespaces and punctuation.

Data Analysis:

We have an overview of training data set and testing data set.

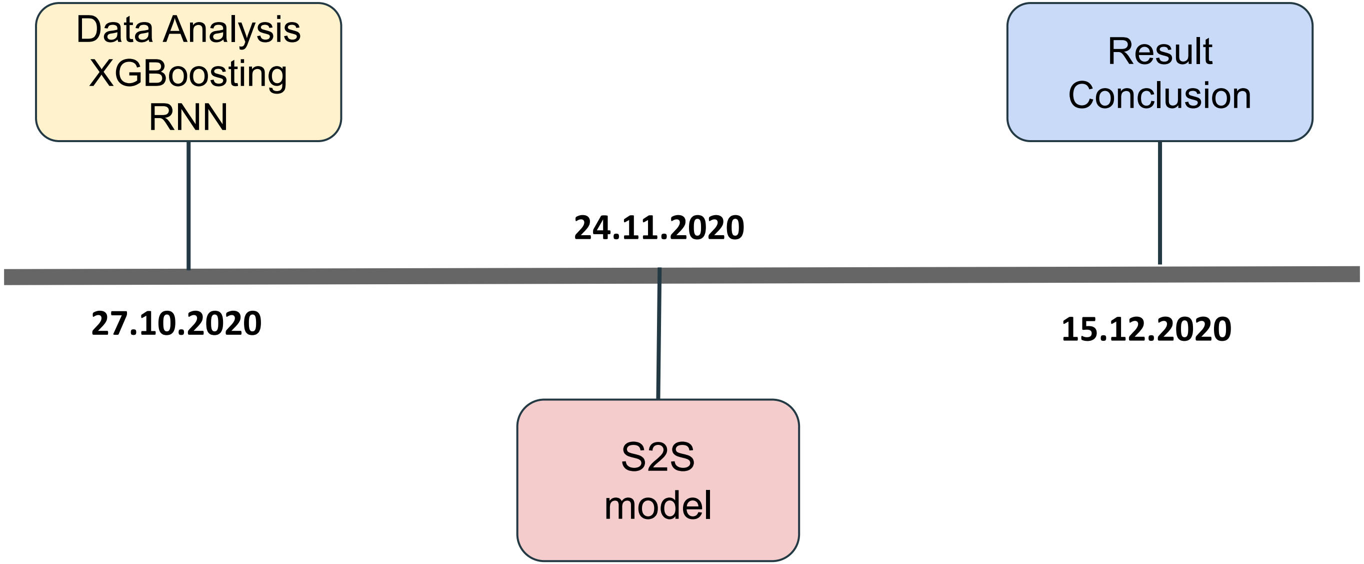
	sentence_id	token_id	class	before	after
0	0	0	PLAIN	Brillantaisia	Brillantaisia
1	0	1	PLAIN	is	is
2	0	2	PLAIN	a	a
3	0	3	PLAIN	genus	genus
4	0	4	PLAIN	of	of

(Training dataset)

	sentence_id	token_id	before
0	0	0	Another
1	0	1	religious
2	0	2	family
3	0	3	is
4	0	4	of

(Testing dataset)

Previous work:



Sequence-to-Sequence Model

- ➔ It can encode the entire sequence data with hidden neurons that would naturally capture any useful context information in a sentence to improve text normalization performance.

How does it work?

- ➔ It reads the informal text sequences in encoder and transforms them in a continuous-space representation that is passed on the decoder to generate a target normalized sequence.

Challenges :

Such an approach faces a major challenge of high percentage of OOV tokens , the model need :

- Enough training data to handle complex normalization
- Hybrid

- After exploring and analyzing different existing works in that task , we chose a good architecture to work on it and try to adapted on our data.

- Architecture is built based on **the encoder-decoder framework** both of which are parameterized by **attention-based recurrent neural networks (RNN)**.

- Proposed in : “Adapting Sequence to Sequence models for Text Normalization in Social Media” paper, 2019 by AAAI.

Tested data : LexNorm dataset from the 2015 ACL-IJCNLP Workshop on Noisy User-generated Text (W-NUT).

Dataset	Tweets	Tokens	Noisy	1:1	1:N	N:1	Our Vocab
Train	2950	44385	3942	2875	1043	10	10,084
Test	1967	29421	2776	2024	704	10	7389

Result:

Model name	Precision	Recall	F1	Method highlights
HS2S	90.66	78.14	83.94	Hybrid word-char Seq2Seq
S2S	93.39	75.75	83.65	Word-level Seq2Seq

Train and test our dataset :

We tried to train the same model on our data , but we faced a challenge in the format of input so we developed a script that takes as input the .csv data and produce like output .json data with another organization of independent and dependent features.

We used this model to predict the outputs of the our data , but the results are very low compared with the expected values.

Results :

Number of instance for train : 90000

Precision	Recall	F1 score
0.9598	0.8542	0.9039

Number of instance for train : 90000

Precision	Recall	F1 score
0.4054	0.070	0.1208

Further work :

After those results , we decide to :

- Re-modified the model to deal correctly with this task.
- Add a rule-based system or a dictionary-based
- Find the good hyper-parameters to achieve high accuracy.
- Train it and test it for the whole data.

Thank you!