Data Augmentation e NLP

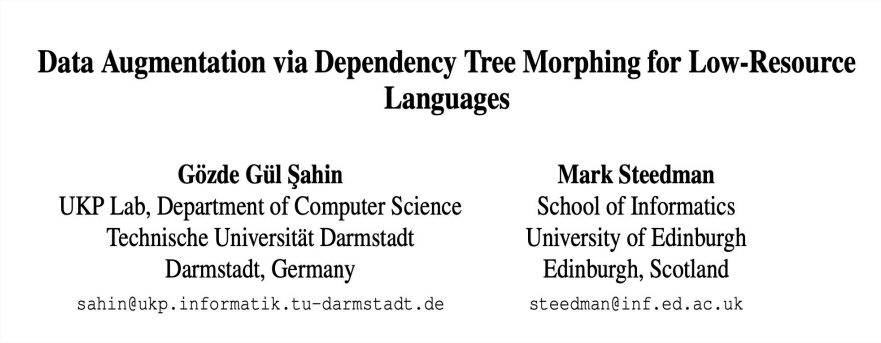
Introdução

Introdução

- Data Augmentation é uma técnica muito utilizada em Visão Computacional
- Utilizamos essa técnica como uma forma de aumentar o nosso dataset
- Podemos rotacionar as imagens de diferentes maneiras, cortar partes delas, mudar a coloração, etc… *Mas e com a linguagem*?

Augmentation com árvore de parsing

Data Augmentation via Dependency Tree Morphing for Low-Resource Languages



Data Augmentation via Dependency Tree Morphing for Low-Resource Languages

Gözde Gül Şahin
UKP Lab, Department of Computer Science
Technische Universität Darmstadt
Darmstadt, Germany
sahin@ukp.informatik.tu-darmstadt.de

Mark Steedman
School of Informatics
University of Edinburgh
Edinburgh, Scotland
steedman@inf.ed.ac.uk

- Hoje, temos resultados muito bons em NLP para línguas com muitos recursos

- O paper foca em línguas com datasets de treino muito pequenos, como línguas Urálicas, Bálticas, etc

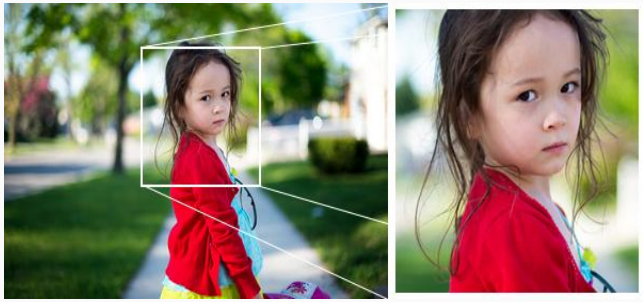
Crop

- Usado em CV para recortar uma parte específica da imagem
- *Recortar a frase*

Rotation

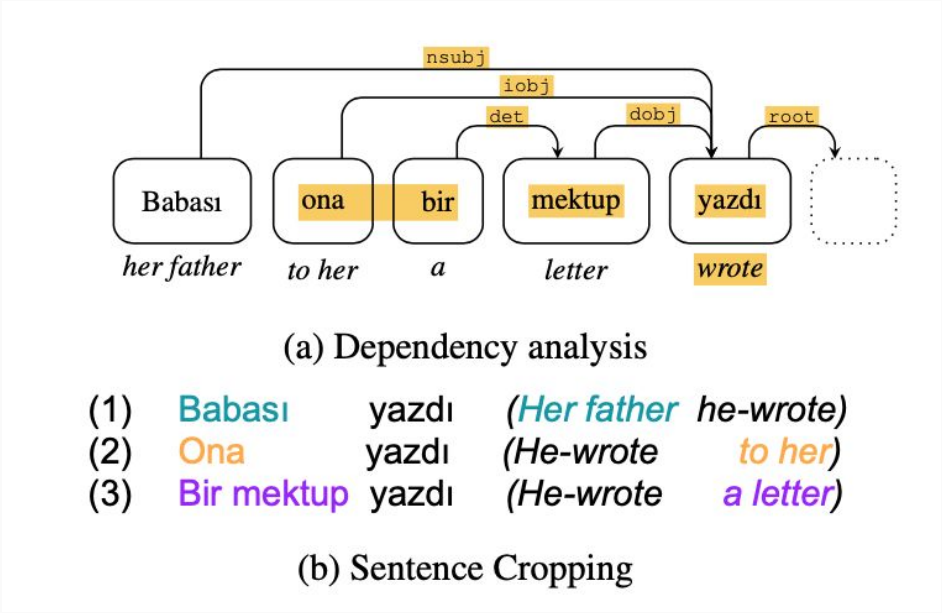
- Rotaciona a imagem em volta de um centro
- *Rotacionar os fragmentos da árvore sintática*

Crop



Usado em CV para remover áreas periféricas da imagem e para focar no sujeito/objeto

Em **NLP** vamos usar essa técnica para identificar partes da sentença mais relevantes e remover as outras partes



Conseguimos formar 3 outras frases

Rotation

Fig. a) Original image

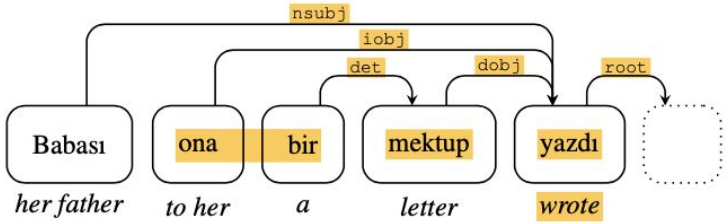

Fig. b) Image after rotation and translation applied

Imagens são rotacionadas ao redor de um centro definido

De maneira similar, selecionamos uma "raiz" da sentença e rotacionamos os fragmentos da frase ao redor dessa raiz

* Esse tipo de modelo é facilmente aplicável para línguas em que a ordem das palavras é dispensável (ordem das palavras não altera o significado), como o Turco, por exemplo. Para línguas como o Português, Inglês, etc., esse tipo de técnica não é muito prático.

Crop & Rotation



(a) Dependency analysis

(1) Babası yazdı *(Her father he-wrote)*
(2) Ona yazdı *(He-wrote to her)*
(3) Bir mektup yazdı *(He-wrote a letter)*

(b) Sentence Cropping

(1) Babası yazdı bir mektup ona (SVOIO)
(2) Yazdı babası ona bir mektup (VSIOO)
(3) Bir mektup yazdı babası ona (OVSIO)
(4) Ona bir mektup yazdı babası (IOOVS)

(c) Sentence Rotating

Seu pai escreveu uma carta para ela (SVO)

Seu pai para ela escreveu uma carta ?

Augmentation simples (eda)

EDA (Easy Data Augmentation)

O paper mostrou 4 formas simples de fazer o Data Augmentation de textos:

1. Substituição de sinônimo
2. Remoção Aleatória
3. Troca Aleatória
4. Inserção Aleatória

Augmentation Contextual

Utilizando modelos de língua

No paper: modelo de língua bi direcional, as palavras previstas como a que queremos substituir seriam as mais adequadas para substituir