A DEEP LEARNING APPROACH TO AUTHOR OBFUSCATION*

By

Özgür Özdemir

Advisor

Assist. Prof. Dr. Tuğba Yıldız

İstanbul
Bilgi Üniversitesi
LAUREATE INTERNATIONAL UNIVERSITIES

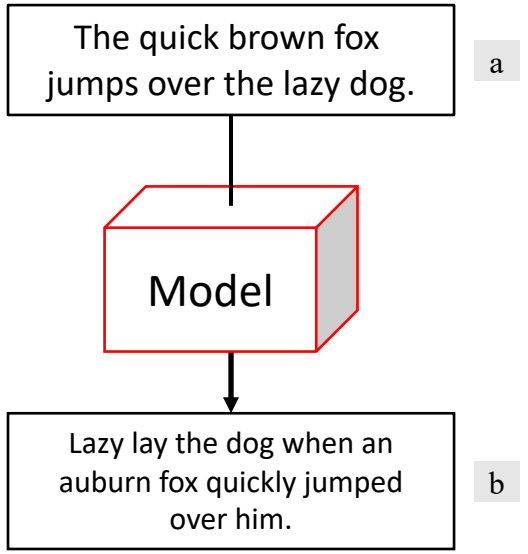
**Masking*

i. Introduction
ii. Related works
iii. Methodology
iv. Summary

INTRODUCTION

Author Obfuscation refers to that *paraphrasing* a text so that the author cannot be identified.

The obfuscation can benefit as many various cases, e.g., *witness protection* or *anonymity program* for social media users.



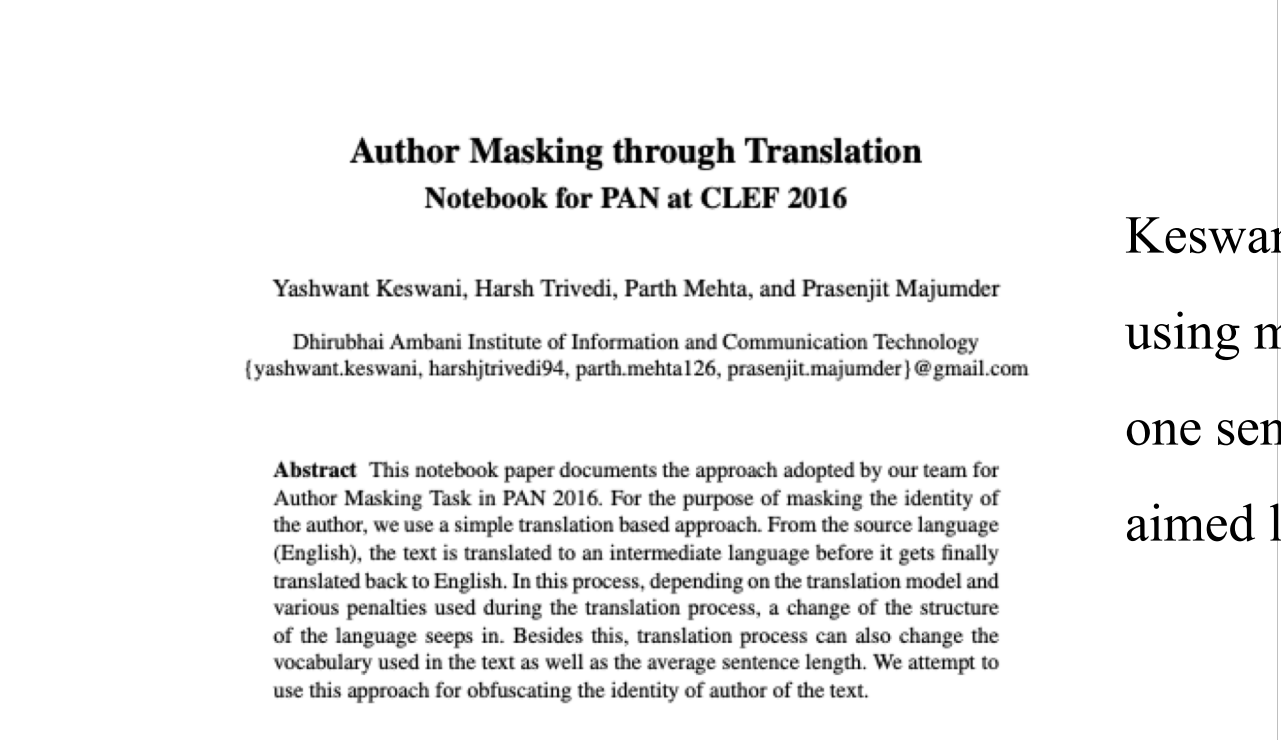
(a) Input text
(b) Output text

Author Obfuscation purposed as lab track of PAN Conference which is sub-conference of CLEF Conference.

PAN Conference will be hosted, besides CLEF Conference, by Università della Svizzera italiana, *Switzerland*, 09-12 September 2019.

Images taken from: https://pan.webis.de/clef19/pan19-web/index.html

RELATED WORKS



Author Masking through Translation

Notebook for PAN at CLEF 2016

Yashwant Keswani, Harsh Trivedi, Parth Mehta, and Prasenjit Majumder

Dhirubhai Ambani Institute of Information and Communication Technology
{yashwant.keswani, harshjtrivedi94, parth.mehta126, prasenjit.majumder}@gmail.com

Abstract This notebook paper documents the approach adopted by our team for Author Masking Task in PAN 2016. For the purpose of masking the identity of the author, we use a simple translation based approach. From the source language (English), the text is translated to an intermediate language before it gets finally translated back to English. In this process, depending on the translation model and various penalties used during the translation process, a change of the structure of the language seeps in. Besides this, translation process can also change the vocabulary used in the text as well as the average sentence length. We attempt to use this approach for obfuscating the identity of author of the text.

Keswani et al. tried to approach the problem by using machine translation method, translating one sentence to another language then back to aimed language.

Source: http://ceur-ws.org/Vol-1609/16090890.pdf

KESWANI ET AL., AUTHOR MASKING THROUGH TRANSLATION, 2016

İstanbul Bilgi Üniversitesi
LAUREATE INTERNATIONAL UNIVERSITIES

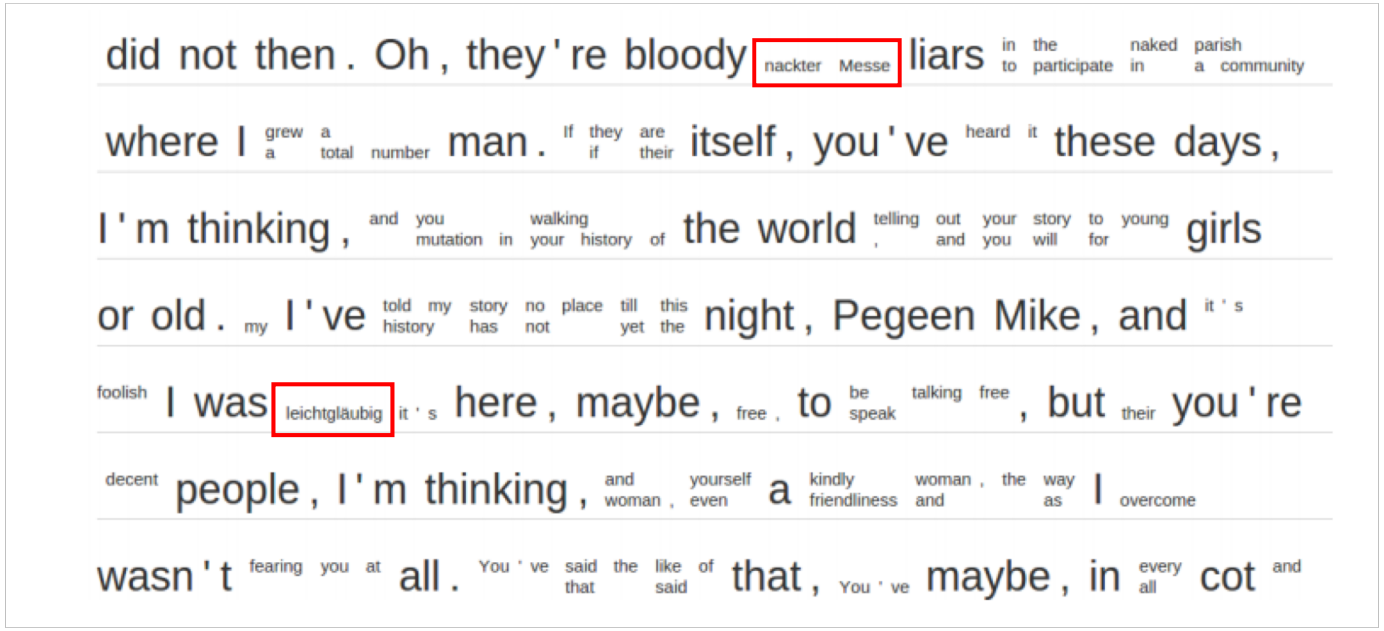


Table. Difference visualization of the original text and the output of Keswani et al.'s obfuscation approach (for the differences there are two lines, original on top).

Source: Stein et al., Author Obfuscation: Attacking the State of the Art in Authorship Verification, n.d.

The Case for Being Average: A Mediocrity Approach to Style Masking and Author Obfuscation

(Best of the Labs Track at CLEF-2017)

Georgi Karadzhov[1], Tsvetomila Mihaylova[1], Yasen Kiprov[1], Georgi Georgiev[1], Ivan Koychev[1], and Preslav Nakov[2]

[1] Faculty of Mathematics and Informatics, Sofia University "St. Kliment Ohridski", Bulgaria
{georgi.m.karadjov, tsvetomila.mihaylova}@gmail.com,
{yasen.kiprov, g.d.georgiev}@gmail.com, koychev@fmi.uni-sofia.bg
[2] Qatar Computing Research Institute, HBKU, Qatar
pnakov@qf.org.qa

Abstract Users posting online expect to remain anonymous unless they have logged in, which is often needed for them to be able to discuss freely on various topics. Preserving the anonymity of a text's writer can be also important in some other contexts, e.g., in the case of witness protection or anonymity programs. However, each person has his/her own style of writing, which can be analyzed using stylometry, and as a result, the true identity of the author of a piece of text can be revealed even if s/he has tried to hide it. Thus, it could be helpful to design automatic tools that can help a person obfuscate his/her identity when writing text. In particular, here we propose an approach that changes the text, so that it is pushed towards average values for some general stylometric characteristics, thus making the use of these characteristics less discriminative. The approach consists of three main steps: first, we calculate the values for some popular stylometric metrics that can indicate authorship; then we apply various transformations to the text, so that these metrics are adjusted towards the average level, while preserving the semantics and the soundness of the text; and finally, we add random noise. This approach turned out to be very efficient, and yielded the best performance on the Author Obfuscation task at the PAN-2016 competition.

Mihaylova et al. tried to approach the problem by extracting the *stylometric features* of test data and generates the output based on them.

Source: https://arxiv.org/abs/1707.03736v2

MIHAYLOVA ET AL., THE CASE FOR BEING AVERAGE: A MEDIOCRITY APPROACH TO STYLE MASKING AND AUTHOR OBFUSCATION, 2017

İstanbul Bilgi Üniversitesi
LAUREATE INTERNATIONAL UNIVERSITIES

Text Metric	Before (Input)	After Obfuscation PAN-2016	After Obfuscation New	Average (Target)
Punctuation to word token count ratio	0.14	0.14	0.15	0.15
Uppercase word tokens to all word tokens count ratio	0.03	0.01	0.02	0.02
Stop words to word token count ratio	0.52	0.45	0.50	0.50
Word type to token ratio	0.44	0.47	0.45	0.44
Number of nouns	0.23	0.24	0.24	0.24
Number of adjectives	0.08	0.09	0.08	0.06
Number of adverbs	0.07	0.09	0.08	0.07
Number of verbs	0.20	0.21	0.21	0.19

Table 2. Impact of the obfuscation on some text metrics. The first column shows the name of a text metric. The second column shows the value of the metrics for the input, i.e., *before* the obfuscation. Then follow the values after obuscation, when using our *PAN-2016* and our *new* method, respectively. Finally, the values in the *average* column are calculated on the training dataset and on some texts from Project Gutenberg; these are the target values we want to push the metrics towards.

However, deciding the stylometric feature is not easy task. (Further reading: *Patrick Juola and Darren Vescovi, Analyzing Stylometric Approaches to Author Obfuscation*)

Source: Mihaylova et al., The Case for Being Average: A Mediocrity Approach to Style Masking and Author Obfuscation, 2017

MIHAYLOVA ET AL., THE CASE FOR BEING AVERAGE: A MEDIOCRITY APPROACH TO STYLE MASKING AND AUTHOR OBFUSCATION, 2017

İstanbul Bilgi Üniversitesi
LAUREATE INTERNATIONAL UNIVERSITIES

Original Text	Machine Translation [12]	Word Substitution [14]	Our PAN-2016 Obfuscation [17]	Our New Obfuscation
I am proud. Though I carry my love with me to the tomb, he shall never, never know it.	I believe expensive Though continue to never, tomb, it. ever be learned	Though I carry my love with me to the tomb , he shall never , never know it .	myself 'm proud in them, and though myself carry my beloved with me to the tomb he shall ever ever know it.	I 'm proud of them; and though I carry my beloved with me to the tomb he shall ever ever know it.
4) Religion discriminates. Sure, it unifies (...). On the other hand	4) religion discriminates. some people Sure, unifies (...) second , it condems	4) Religion discriminates . Sure , it unifies (...) On the other hand	Four) Religion discriminate; as sure, it unified (..), and on the other hand	Four) Religion discriminate, sure, it unified (...); on the other side
Consequently, they see a connection between development in spiritual life and professional economic development	they Consequently, a link between the spiritual development and economic professional development.	Consequently , they see a connection between growth in spiritual life and professional economic development .	Definitely, Consequently, they see a connection between development inside spiritual life also professional economic development;	Consequently, they see a link between development in spiritual life and professional economic development,

Table 3. Obfuscation examples. Shown are examples of how the different systems that participated in the PAN-2016 Author Obfusction task transform the original text; the last column shows the output of our new obfuscation method, which we introduced in this paper.

Source: Mihaylova et al., The Case for Being Average: A Mediocrity Approach to Style Masking and Author Obfuscation, 2017

İstanbul
Bilgi Üniversitesi
LAUREATE INTERNATIONAL UNIVERSITIES

METHODOLOGY

i. Machine Learning with TensorFlow
ii. Sequence to Sequence Learning
iii. Attention Model
iv. Approach to Problem

TERMINOLOGY

İstanbul Bilgi Üniversitesi
LAUREATE INTERNATIONAL UNIVERSITIES

RNN (Recurrent Neural Network): A machine learning algorithm works with sequences like texts.

LSTM (Long Short Term Memory): A specialized RNN structure that decides what dependencies are need to memorize or forget. Think about whole paragraph instead of one sentence.

Embeddings – Word Embeddings: Vectorized version of the words in space. Embeddings use for operating mathematical operation that necessary for machine learning.

For further information about RNN structure and usage: *Sutskever et al., An Empirical Exploration of Recurrent Network Architectures, 2015*

MACHINE LEARNING WITH TENSORFLOW

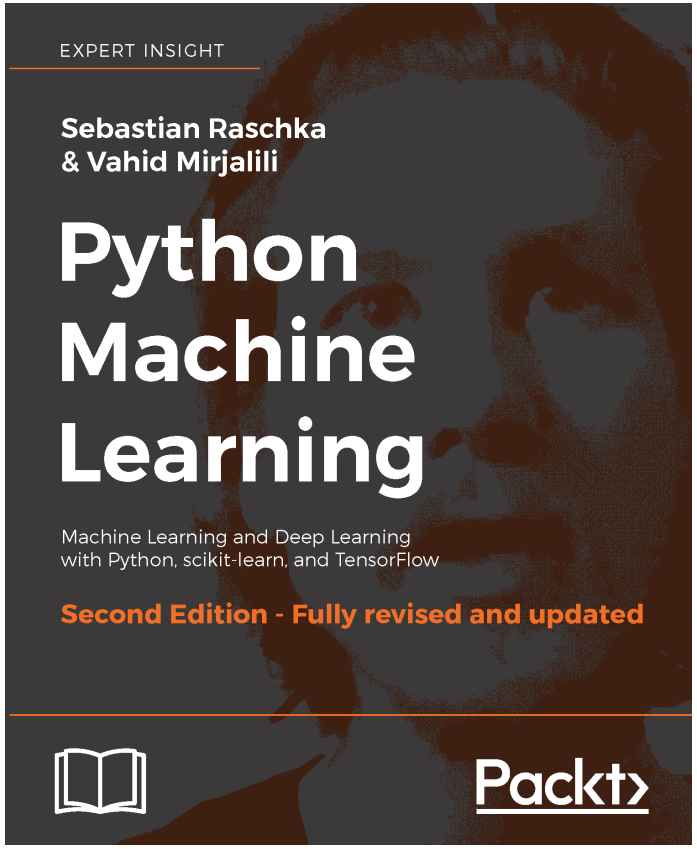
İstanbul Bilgi Üniversitesi
LAUREATE INTERNATIONAL UNIVERSITIES

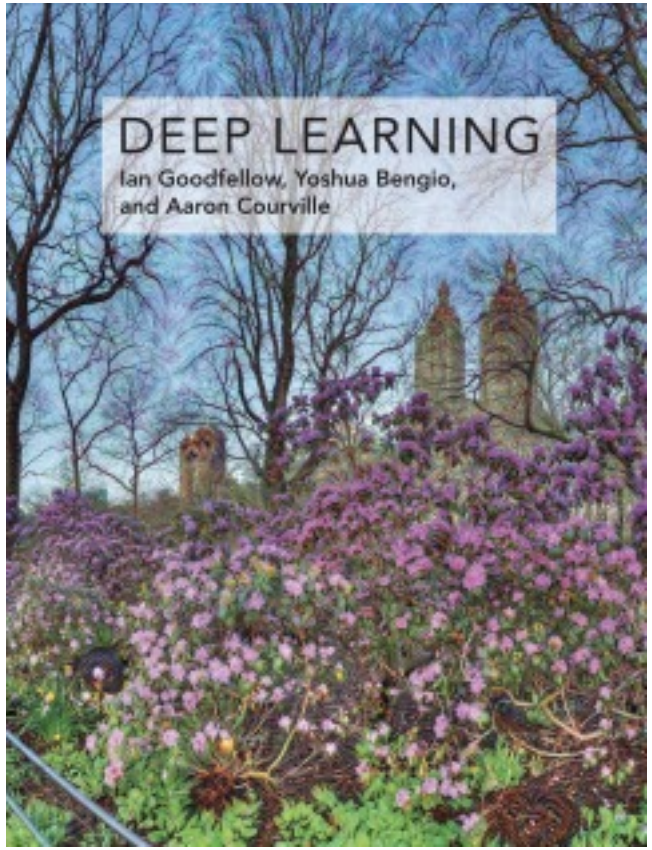
TensorFlow is an open-source machine learning library for research and production.

In order to implement deep learning approaches, TensorFlow library is the one of the strongest option out there.

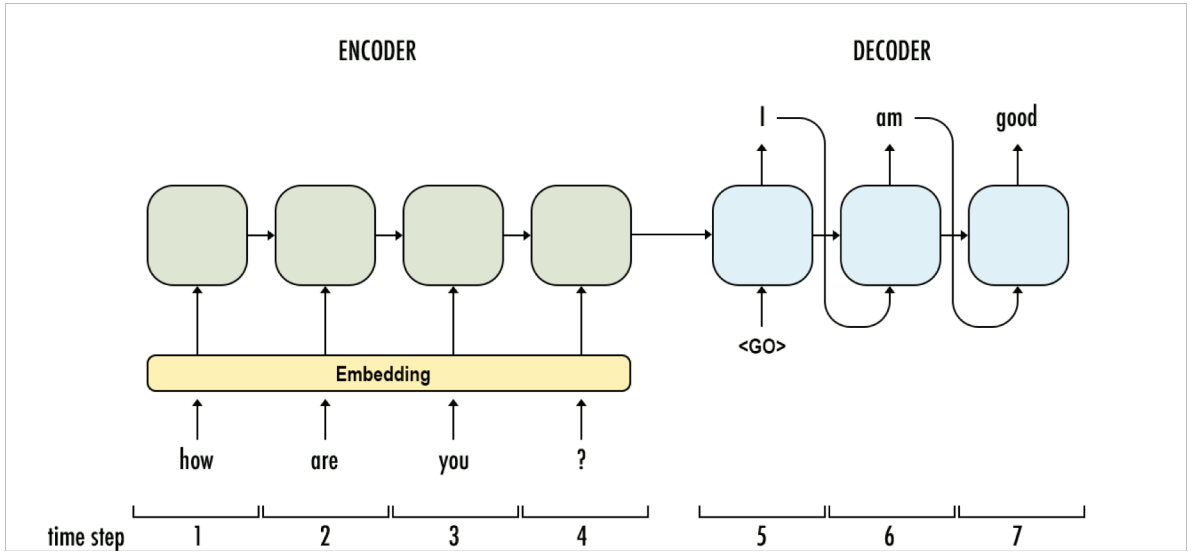
```
import tensorflow as tf
mnist = tf.keras.datasets.mnist

(x_train, y_train),(x_test, y_test) = mnist.load_data()
x_train, x_test = x_train / 255.0, x_test / 255.0

model = tf.keras.models.Sequential([
  tf.keras.layers.Flatten(),
  tf.keras.layers.Dense(512, activation=tf.nn.relu),
  tf.keras.layers.Dropout(0.2),
  tf.keras.layers.Dense(10, activation=tf.nn.softmax)
])
model.compile(optimizer='adam',
              loss='sparse_categorical_crossentropy',
              metrics=['accuracy'])

model.fit(x_train, y_train, epochs=5)
model.evaluate(x_test, y_test)
```

Example code written with Python using TensorFlow library


EXPERT INSIGHT
Sebastian Raschka
& Vahid Mirjalili
Python Machine Learning
Machine Learning and Deep Learning with Python, scikit-learn, and TensorFlow
Second Edition - Fully revised and updated
Packt


DEEP LEARNING
Ian Goodfellow, Yoshua Bengio, and Aaron Courville

SEQUENCE TO SEQUENCE LEARNING



"One RNN encodes a sequence of symbols into a fixed length vector representation, and the other decodes the representation into another sequence of symbols."[2]

Image taken from: https://towardsdatascience.com/sequence-to-sequence-model-introduction-and-concepts-44d9b41cd42d

SEQUENCE TO SEQUENCE LEARNING

[1] Sutskever et al., Sequence to Sequence Learning with Neural Networks, 2014

[2] Cho et al., Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation, 2014

ATTENTION MODEL

The model aims to pay more attention on specific weights instead of equally operating the sequence.

Therefore, the model is more accurate for long sequences, because the dependencies get significant.

For example, the verb can change the whole meaning of the sentence, so it has more significant role and to be paid more attention.

An example for attention model in picture captioning

Source: Xu et al., Show, Attend and Tell: Neural Image Caption Generation with Visual Attention, 2016

ATTENTION MODEL

İstanbul Bilgi Üniversitesi
LAUREATE INTERNATIONAL UNIVERSITIES

[3] Bahdanau et al., Neutral Machine Translation by Jointly Learning to Align and Translate, 2016

[4] Xu et al., Show, Attend and Tell: Neural Image Caption Generation with Visual Attention, 2016

[5] Luong et al., Effective Approaches to Attention-based Neural Machine Translation, 2015

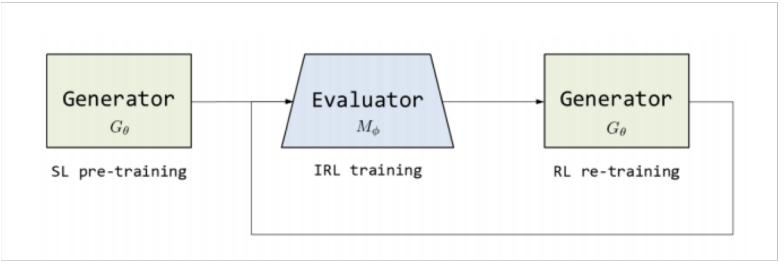
A DEEP LEARNING APPROACH TO AUTHOR OBFUSCATION

İstanbul Bilgi Üniversitesi
LAUREATE INTERNATIONAL UNIVERSITIES

I propose employing a sequence to sequence model with attention mechanism to achieve to obfuscate the text meaningfully and readable.

To do so, the encoder of the Seq2Seq model will employ to refine the stylometric identity of the text.

Furthermore, the decoder of the Seq2Seq model will employ to generate the text by supplying the BLEU* score.



Model for generating the text by adjusting with evaluator

**BLEU (Bilingual evaluation understudy): A metric for evaluating a generated sentence to a reference sentence.*

Source: Li et al., Show, Paraphrase Generation with Deep Reinforcement Learning, 2018

A DEEP LEARNING APPROACH TO AUTHOR OBFUSCATION

İstanbul Bilgi Üniversitesi
LAUREATE INTERNATIONAL UNIVERSITIES

For further information about Generator and Evaluator model:

[6] Li et al., Paraphrase Generation with Deep Reinforcement Learning, 2018

SUMMARY

Author Obfuscation is *paraphrasing* a text so that the author cannot be identified.

The obfuscation can benefit such cases as *witness protection* or *anonymity* for social media users.

The approaches before were tending to be rule-based – hard to *decide* and *implement* – or far away from the *sensibility*.

Therefore, I propose a machine learning approach to identify the features *automatically* and generate the text *far* from the identity but *meaningful*.

İstanbul
Bilgi Üniversitesi
LAUREATE INTERNATIONAL UNIVERSITIES

THANKS FOR LISTENING!

Özgür Özdemir

You can find all resources, papers, besides useful links and some codes.