

# Text Simplification

Srđan Ljepojević

2021

# Outline

## 1 Introduction

## 2 State of the art

- Text simplification approaches
- Applications for NLP Tasks

## 3 Web Application for Alignment of Sentence Pairs from Parallel Corpora

# Outline

## 1 Introduction

## 2 State of the art

- Text simplification approaches
- Applications for NLP Tasks

## 3 Web Application for Alignment of Sentence Pairs from Parallel Corpora

# Definition

## Text simplification

process of making some text easier to understand while still making it grammatically correct and without leaving out any important information

## Early Days

- Parsers
- Systems:
  - Machine translation
  - Summarization
  - Information retrieval

## Present Day

### Benefit:

- Children
- Non-native speakers
- Low literacy readers
- Bridge the gap between layman and expert
- People with disabilities (autism, dyslexia, aphasia)

# Approaches

## Historic

- Rule based
- Syntactic simplification
- Lexical simplification

## Modern

- Data-driven
- Focus on sentences

# Outline

1 Introduction

**2** State of the art

- Text simplification approaches
- Applications for NLP Tasks

3 Web Application for Alignment of Sentence Pairs from Parallel Corpora



# Datasets

## PWKP

- Normal and Simple Wikipedia
- 108 thousand sentence pairs
- 1:1 and 1:N sentence mappings

## TurkCorpus

- Taken from Normal Wikipedia <sup>1</sup>
- Amazon Mechanical Turk
- 2350 sentences
- 8 simplifications

---

<sup>1</sup>subset of PWKP

# Metrics

- SARI
  - System Output Against References And Input Sentence
  - Zhao, Meng, He, Saptono and Parmanto (2018)
  
- BLEU
  - BiLingual Evaluation Understudy
  - Papineni, Roukos, Ward and Zhu (2002)

## Best Scores on PWKP

- PBMT-R
  - Phrase Based Machine Translation
  - Wubben, Bosch and Kraemer (2012)
  
- Hybrid approach
  - Deep semantics (DRS) + MT
  - Narayan and Gardent (2014)

## Introduce Knowledge About Paraphrasing

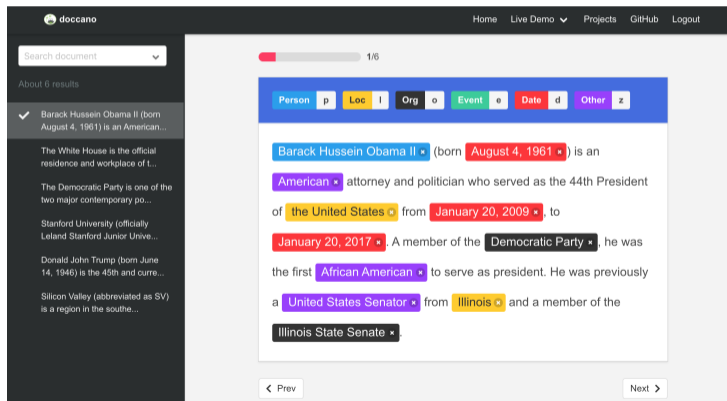
- Paraphrase database (PPDB)
  - Ganitkevitch, Van Durme and Callison-Burch (2013)
  - 220 million paraphrase pairs
- Simple PPDB (SPPDB)
  - Pavlick and Callison-Burch (2016)
  - 4.5 million paraphrase pairs

## Best Scores on TurkCorpus

- D<sub>MASS</sub>-D<sub>CSS</sub>
  - Deep Memory Augmented Sentence Simplification model
  - Deep Critical Sentence Simplification model
  - Zhao, Meng, He, Saptono and Parmanto (2018)
  
- SBSMT(PPDB+SARI)
  - Syntactic based machine translation
  - Xu, Napoles, Pavlick, Chen and Callison-Burch (2016)

# Applications for NLP Tasks

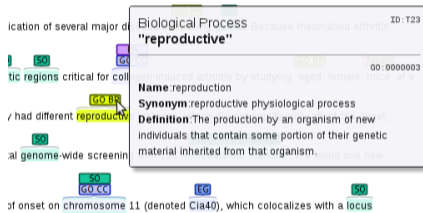
## Doccano



The screenshot displays the Doccano web application interface. On the left, a dark sidebar contains a search bar and a list of document results. The main content area shows a document snippet with various entities highlighted in colored boxes and labeled with a legend above. The legend includes: Person (p), Loc (l), Org (o), Event (e), Date (d), and Other (z). The document text is: "Barack Hussein Obama II (born August 4, 1961) is an American attorney and politician who served as the 44th President of the United States from January 20, 2009, to January 20, 2017. A member of the Democratic Party, he was the first African American to serve as president. He was previously a United States Senator from Illinois and a member of the Illinois State Senate." The interface also features a progress bar at the top of the main area and navigation buttons for "Prev" and "Next" at the bottom.

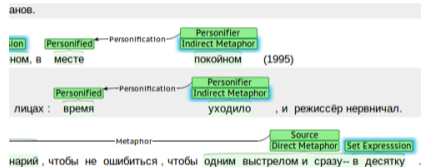
# Applications for NLP Tasks

## Brat



Biological Process ID: T23  
**"reproductive"**  
 Name: reproduction  
 Synonym: reproductive physiological process  
 Definition: The production by an organism of new individuals that contain some portion of their genetic material inherited from that organism.

The Colorado Richly Annotated Full Text Corpus (CRAFT)



анов.  
 (ion) Personified ← Personification — Personifier Indirect Metaphor  
 ном, в месте покойном (1995)  
 Personified ← Personification — Personifier Indirect Metaphor  
 лица: время ушло, и режисёр нервничал.  
 — Metaphor — Source Direct Metaphor Set Expression  
 нарий, чтобы не ошибиться, чтобы одним выстрелом и сразу— в десятку

Russian-language corpus of conceptual metaphor

# Outline

1 Introduction

2 State of the art

- Text simplification approaches
- Applications for NLP Tasks

**3** Web Application for Alignment of Sentence Pairs from Parallel Corpora

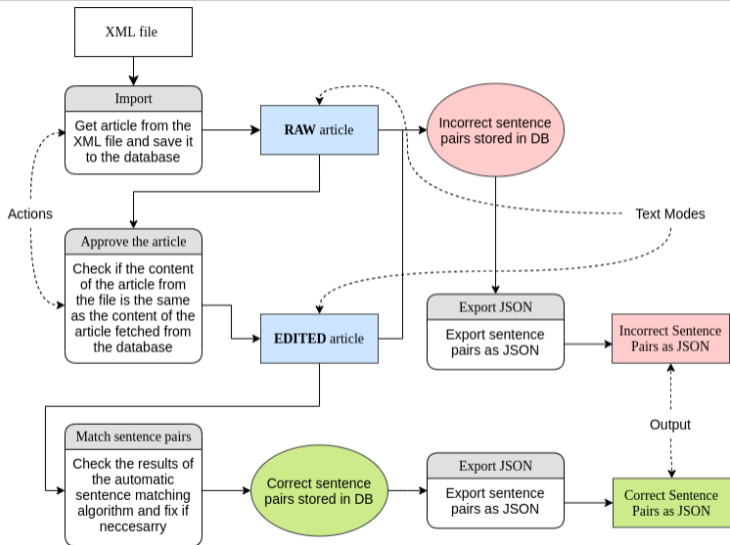


## Requirements

- Import articles from XML files
- Show side-by-side view of articles
- Create/Read/Update/Delete sentence pairs
- Export sentence pairs as JSON
- Color code matching sentence pairs

Demo

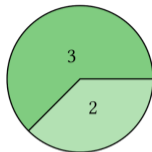
# Web Application for Alignment of Sentence Pairs from Parallel Corpora



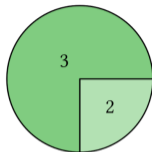
# Evaluation

## Evaluation

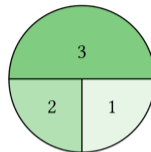
- 8 participants
- 10 tasks (ranging from simple ones to the ones representing the full workflow)
- Questionnaire
  - 7 questions
  - 7 point Likert scale



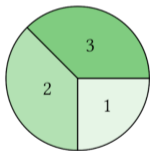
(a) How fast is the system navigation



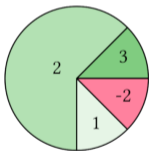
(b) How fast is the file import



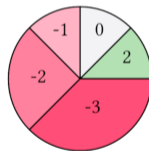
(c) How fast is sentence matching



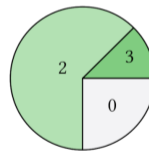
(d) How easy is to import files



(e) How easy is to match sentences



(f) How easy is to find specific article from a file



(g) How easy is to edit the content of the article

## Conclusion

- Data driven approaches
- Metrics and datasets
- Web Application for Alignment of Sentence Pairs from Parallel Corpora



# Questions?