Cairo University

Faculty of Engineering

SUMSIM

There is beauty in simplicity

Names:
Amira Ahmed Amer
Dina Adib Fouad
Rana Amr Afifi
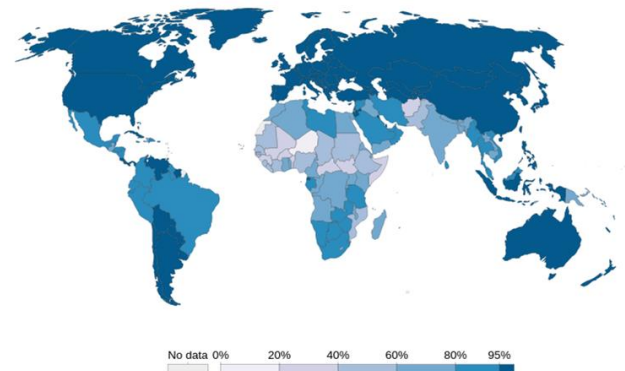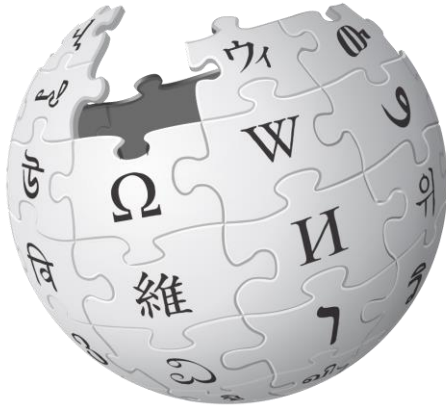Salma Hanafy

Introduction

5-10%

750 Million
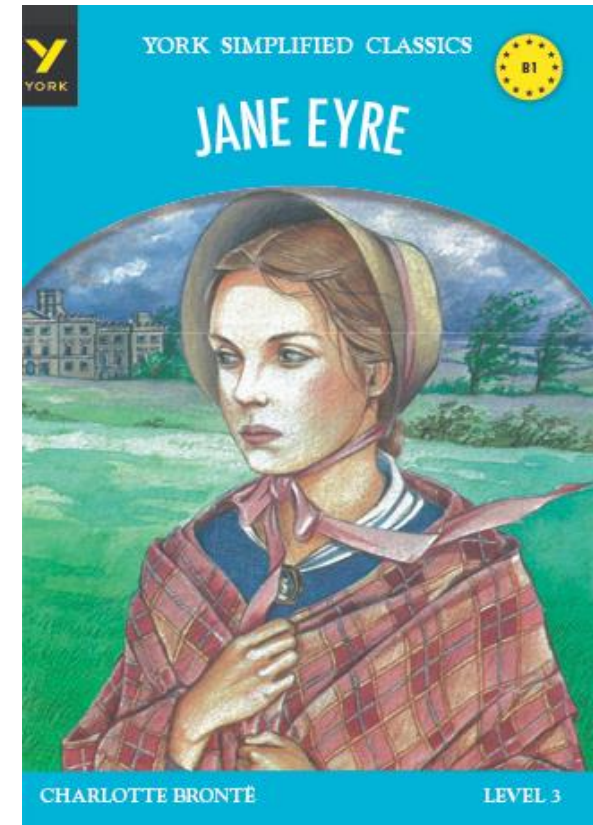
200 Million

Simple English WIKIPEDIA

Zero Project — For a world without barriers

inclusion europe

Supported by the European Union

YORK SIMPLIFIED CLASSICS
JANE EYRE
CHARLOTTE BRONTË
LEVEL 3

SUMSIM

There is beauty in simplicity

A *Summarization* and *Simplification* System

Lexical Simplification

When the leg is released, the popcorn moves like a gymnast somersaulting.

⬇

When the leg is released , the popcorn moves like **an athlete flip**.

Syntactic Simplification

A succession of steadily more powerful and flexible computing devices were constructed in the 1930s and 1940s , gradually adding the key features that are seen in modern computers .

More powerful and simpler computers were made in the 1930s and 1940s .

Text Summarization

'Alcohol' is taken in almost all cool and cold climates, and to a very much less extent in hot ones. Thus, it is taken by people who live in the Himalaya Mountains, but not nearly so much by those who live in the plains of India. Alcohol is not necessary in any way to anybody. The regular use of alcohol, even in small quantities, tends to cause mischief in many ways to various organs of the body. It affects the liver, it weakens the mental powers, and lessens the general energy of the body. In addition, damage to the central nervous system and peripheral nervous system can occur from chronic alcohol abuse..
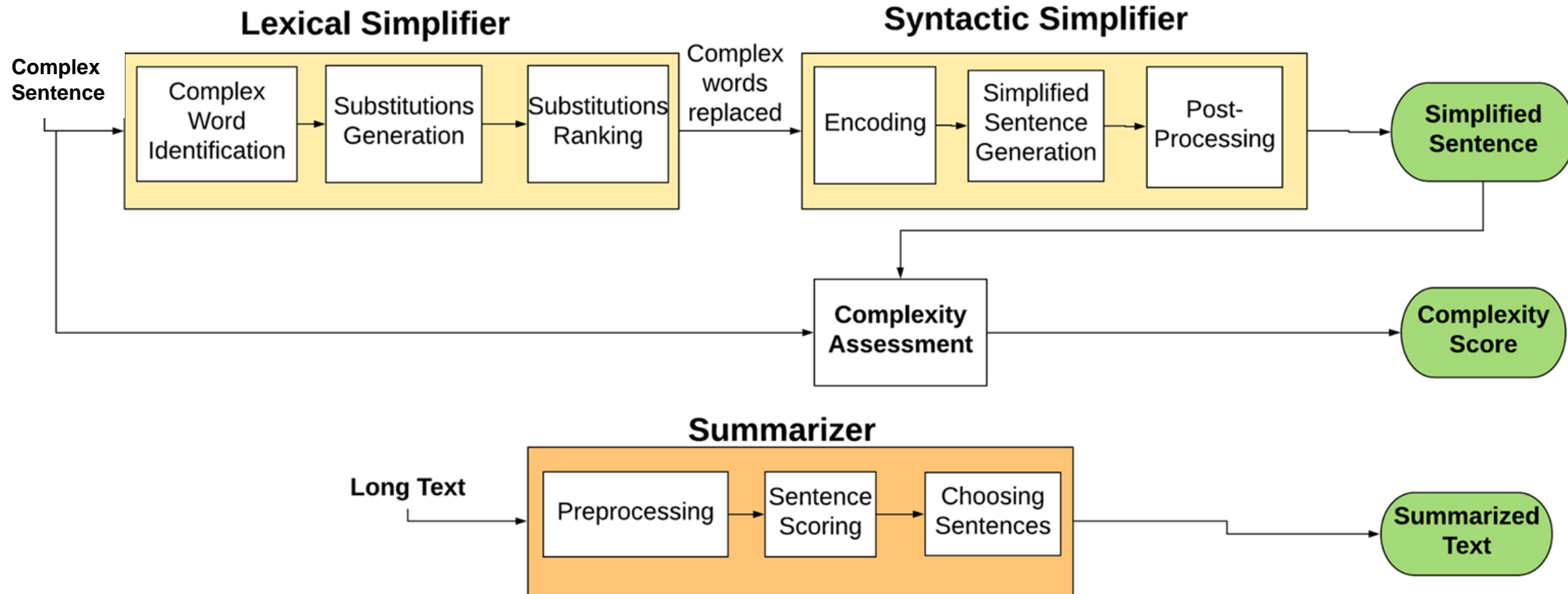
Alcohol' is taken in almost all cool and cold climates, and to a very much less extent in hot ones. Alcohol is not necessary in any way to anybody. The regular use of alcohol, even in small quantities, tends to cause mischief in many ways to various organs of the body.
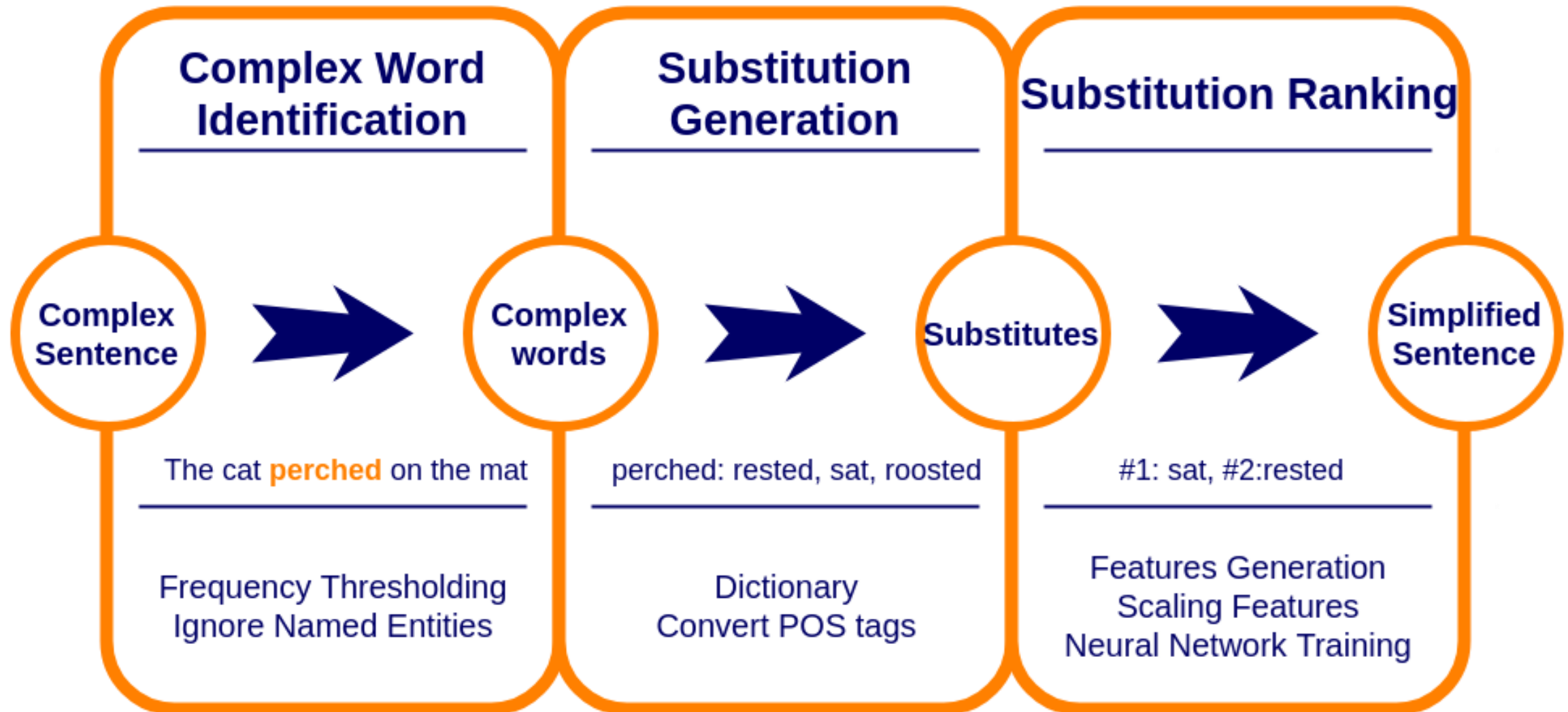
8

Target Audience

Target Audience
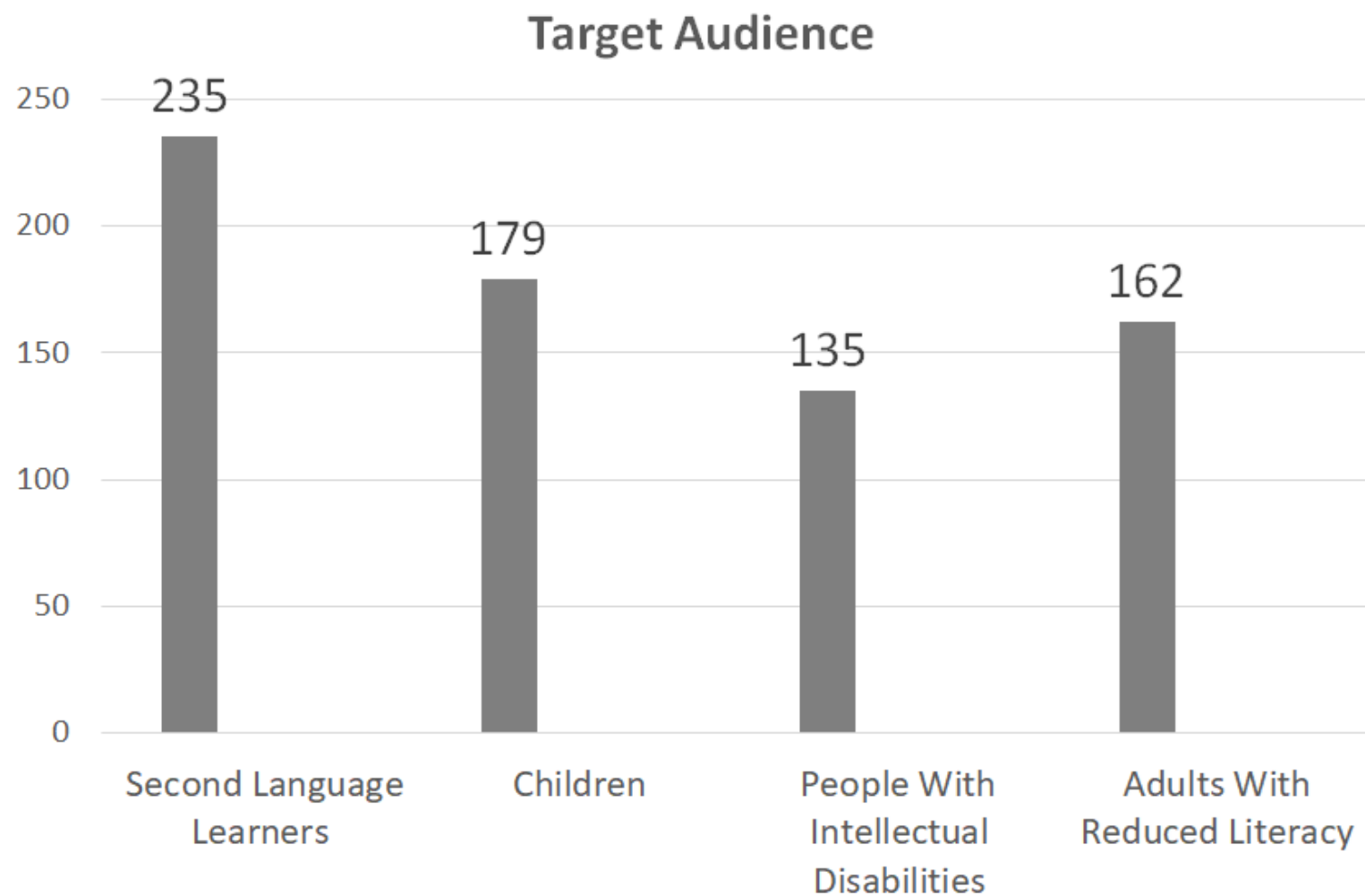
System Design Overview

System Architecture

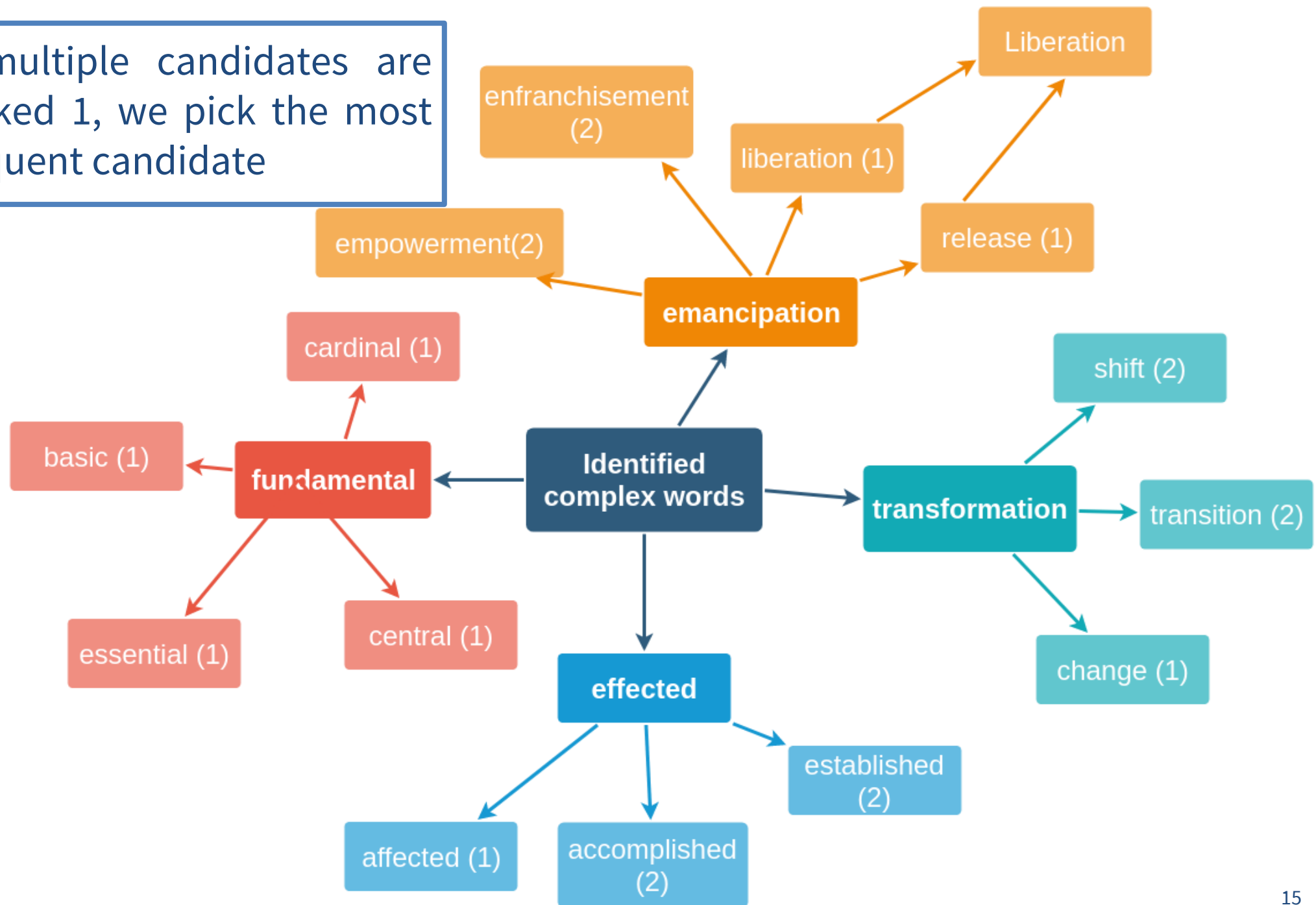# Lexical Simplification

Lexical Simplification

Example
Step 1: Complex Word Identification

The **emancipation** of women can only be completed when a **fundamental transformation** of living is **effected**.

If multiple candidates are ranked 1, we pick the most frequent candidate

enfranchisement (2)

Liberation

liberation (1)

empowerment(2)

release (1)

emancipation

cardinal (1)

shift (2)

basic (1)

fundamental

Identified complex words

transformation

transition (2)

essential (1)

central (1)

change (1)

effected

affected (1)
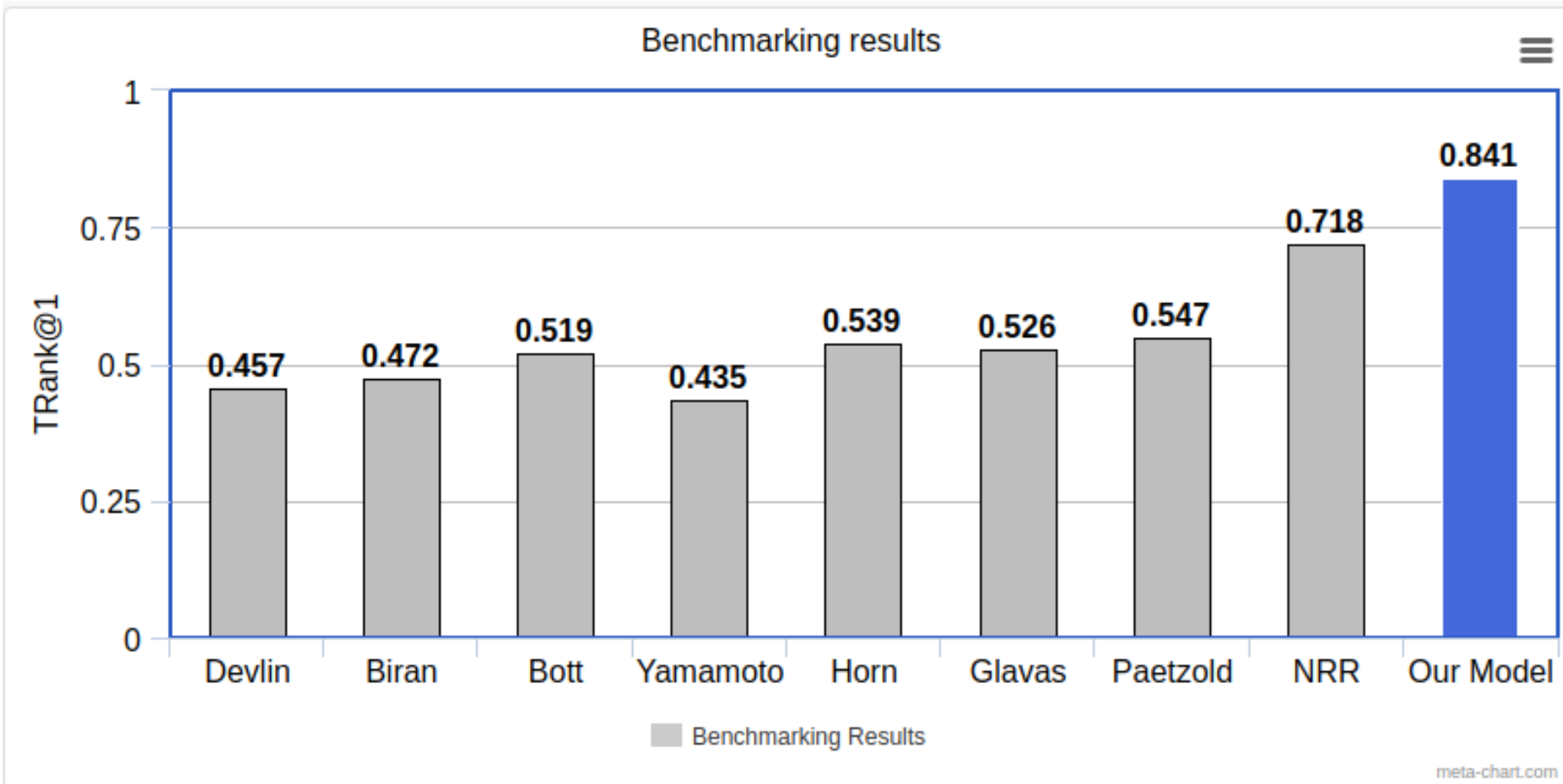
accomplished (2)

established (2)

Output

The **emancipation** of women can only be completed when a **fundamental transformation** of living is **effected**.

⬇

The **liberation** of women can only be completed when an **essential change** of living is **affected**

Benchmarking Results

Benchmarking results

Devlin	Biran	Bott	Yamamoto	Horn	Glavas	Paetzold	NRR	Our Model
0.457	0.472	0.519	0.435	0.539	0.526	0.547	0.718	0.841

- TRank@1: the proportion of times in which the highest ranked candidate in BenchLS dataset was ranked first by the ranker

(Gustavo H. Paetzold and Lucia Specia, 2016)

Limitations

| Input	System output
Police used tear gas , dogs and clubs on the unarmed protesters .	Police used tear gas , dogs and clubs on the **armed supporters** .
Unhappy with being called to task for declining scores , political educators complained that the tests were unfair to some , so they <u>watered them down</u> (a phrasal verb)	Unhappy with being called to task for reduction scores , political actors complaints that the tests were unfair to some , so they **flooded** them down

18

Syntactic Simplification

Syntactic Simplification

3 main operations: **Copying** , **Deletion** , **Substitution**

Complex	**A succession of steadily more powerful and flexible computing devices were constructed in the 1930s and 1940s , gradually adding the key features that are seen in modern computers .**
Simple	**More powerful and simpler computers were made in the 1930s and 1940s .**

System Architecture

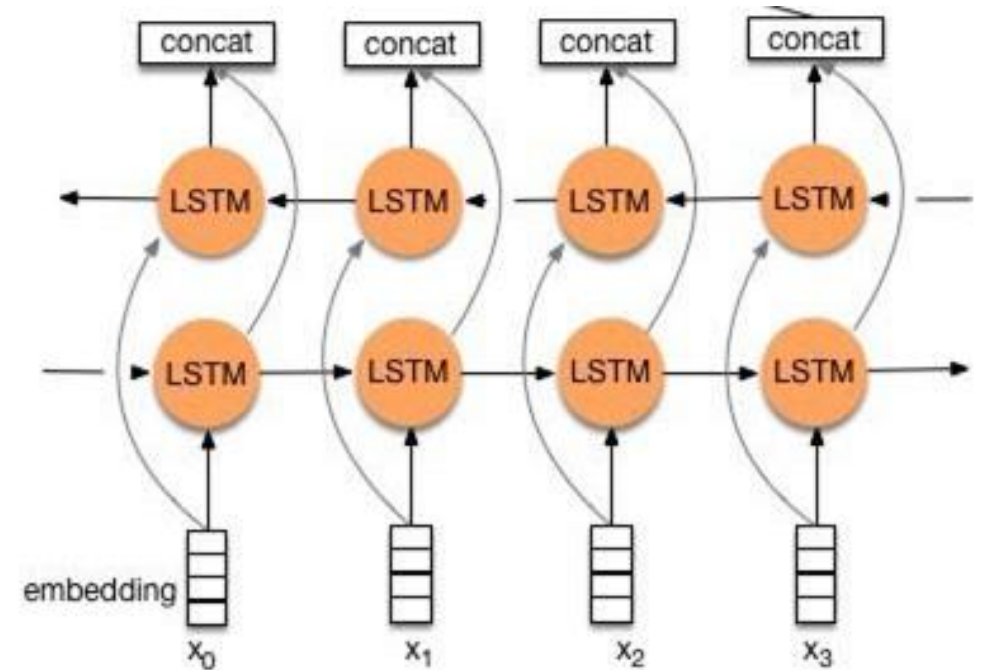Attention-based encoder decoder model with LSTMs
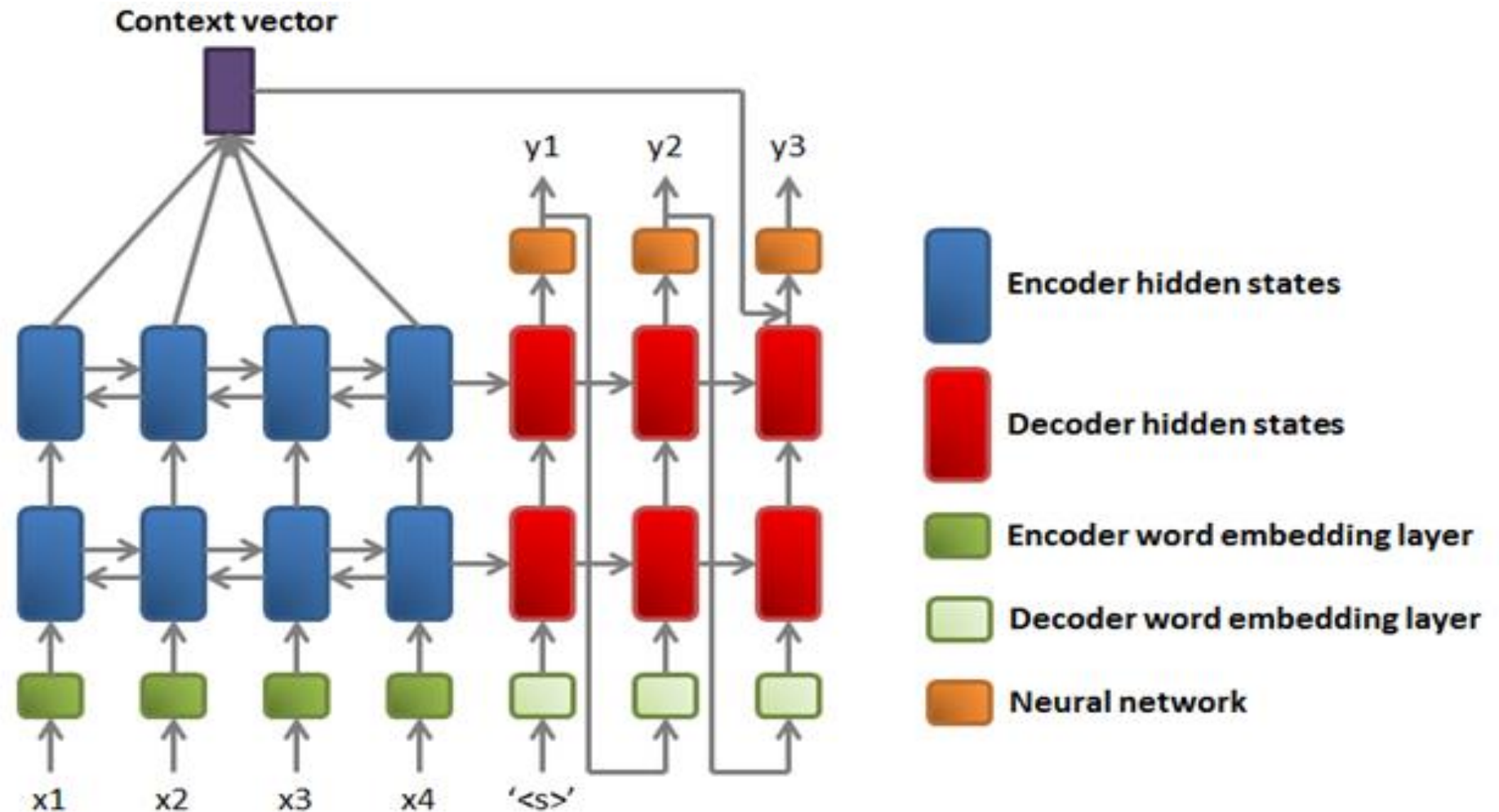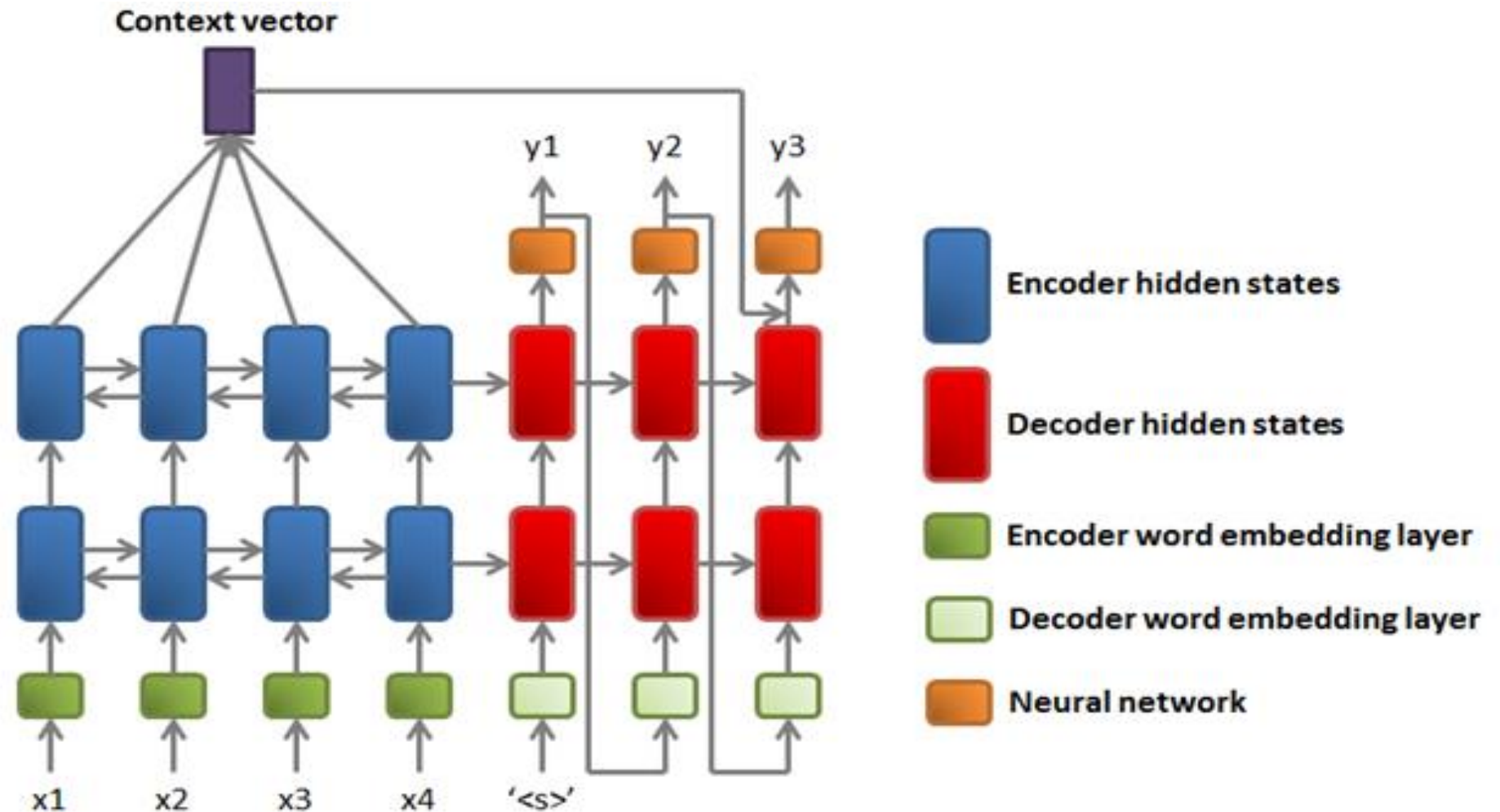
System Architecture

➜RNN

Encoder: Bi-Directional LSTMs
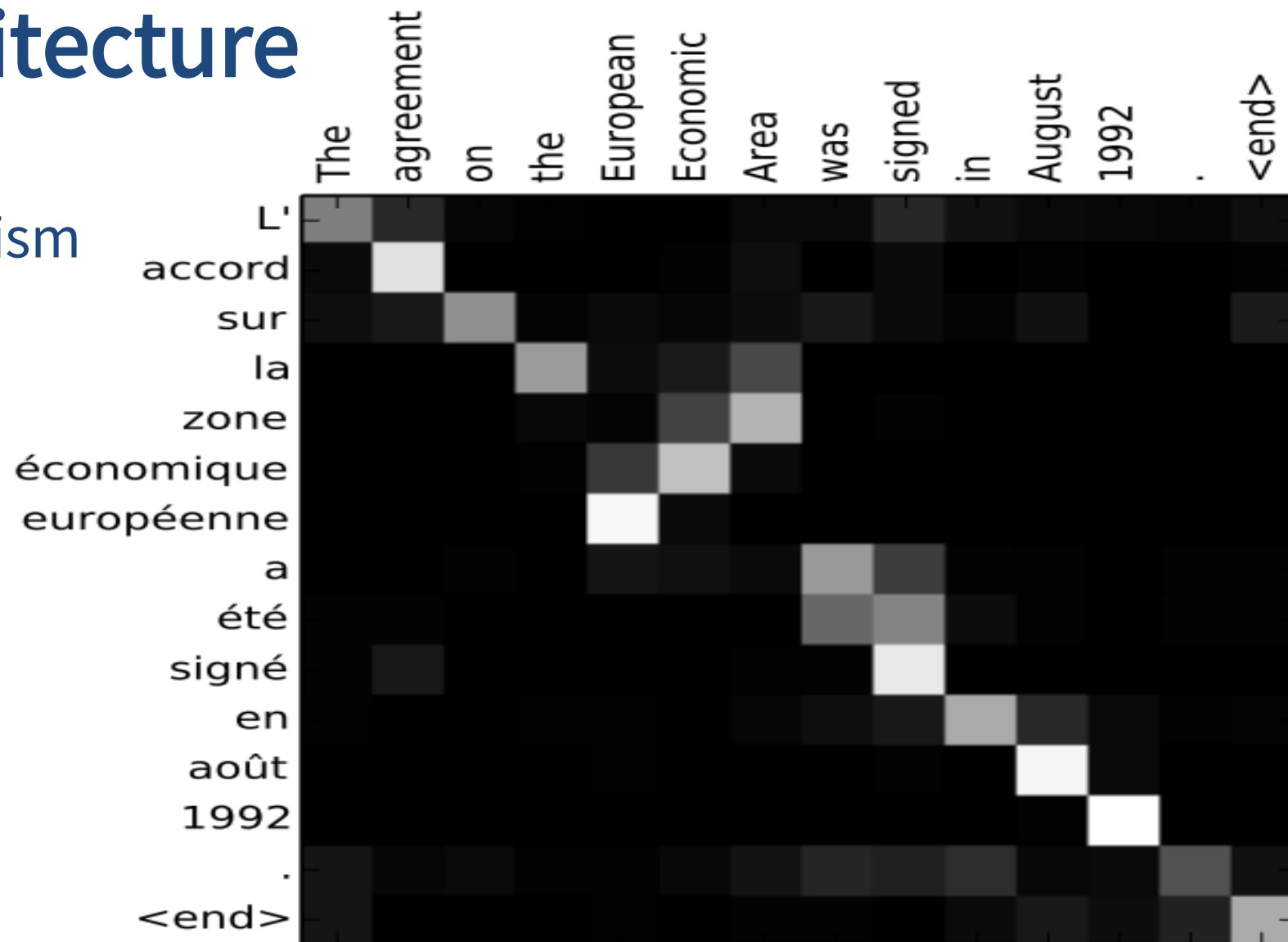
Decoder: Uni-Directional LSTMs

System Architecture

Attention-based encoder decoder model with LSTMs

System Architecture

➔ Attention Mechanism

Training Process

Supervised Learning → **Reinforcement Learning**
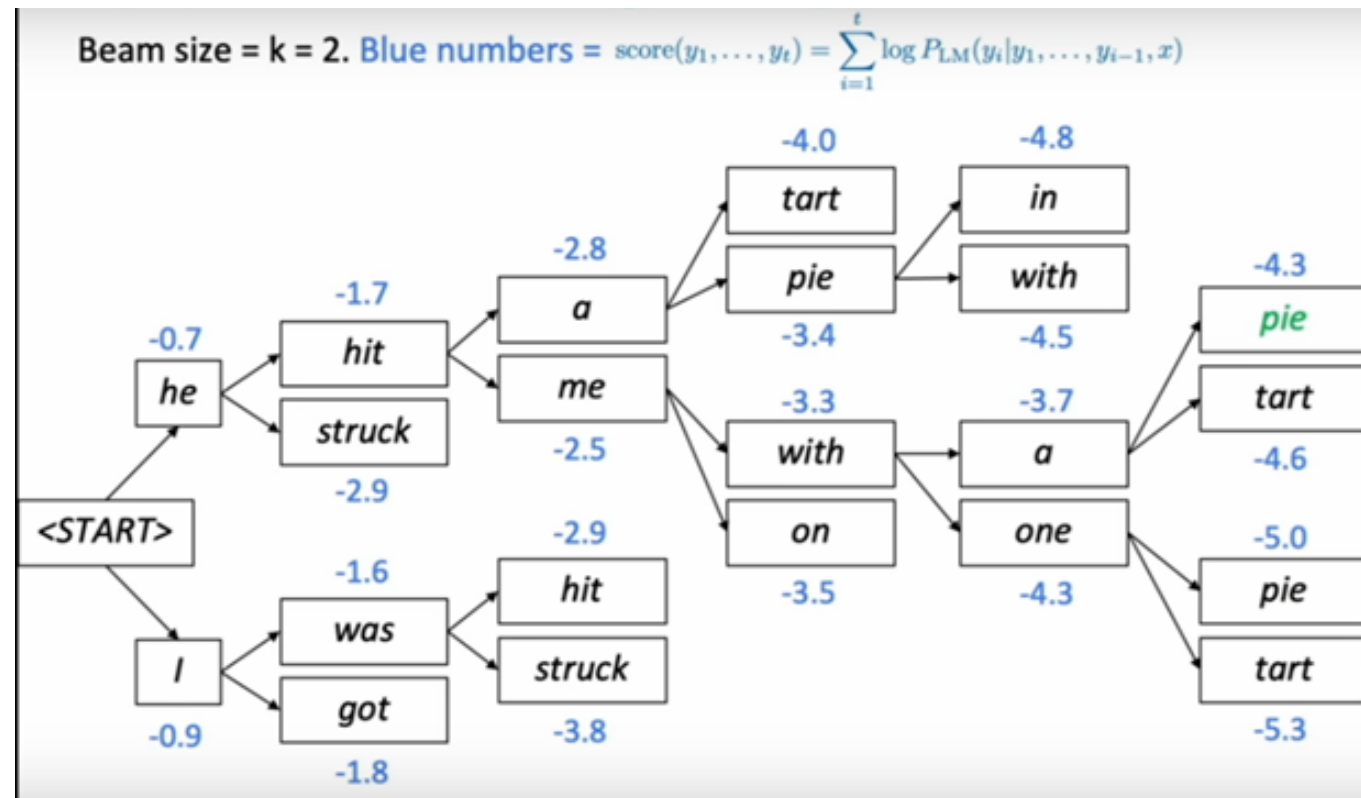
Training Process

➡️ Reinforcement Learning

$$r(\hat{Y}) = \lambda^S r^S(X, \hat{Y}, Y) + \lambda^R r^R(\hat{Y}, Y)$$

r^S : SARI score

$$r^R = \beta_1 \text{ROUGE-2}(\hat{Y}, Y) + (1 - \beta_1)\text{ROUGE-L}(\hat{Y}, Y)$$

Inference: Beam Search

Beam size = k = 2. Blue numbers = $\text{score}(y_1, \ldots, y_t) = \sum_{i=1}^{t} \log P_{\text{LM}}(y_i | y_1, \ldots, y_{i-1}, x)$

Post-Processing: Unknown Words (UNK) Replacement

Before UNK Replacement	Now \<unk> helps \<unk> with her flips .
After UNK Replacement	Now Kory helps Trinity with her flips .

Experimentation Results
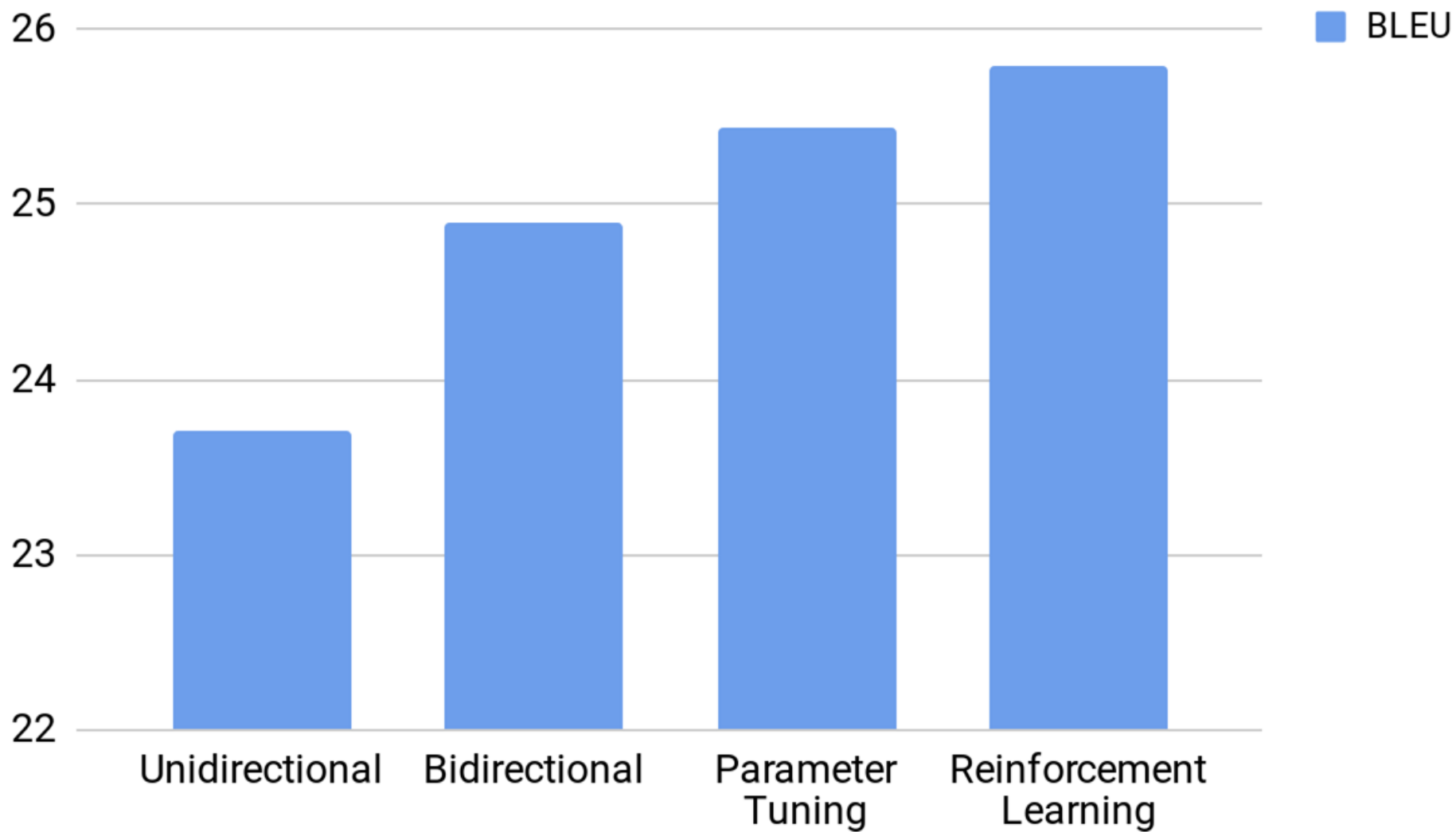
Syntatic Simplification Results
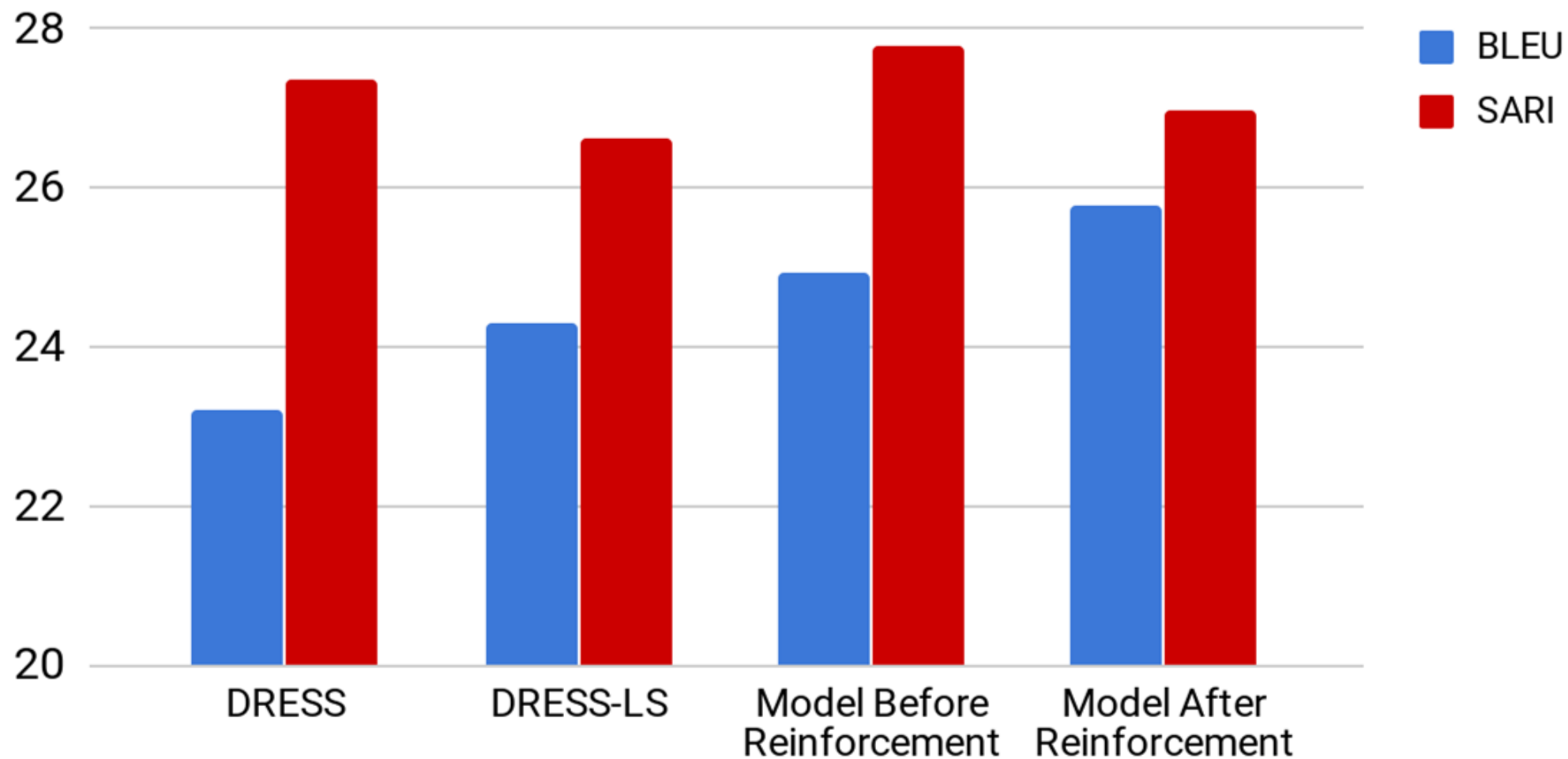
Limitations

Input	System output
Gary Goddard is the founder of Gary Goddard Entertainment .	Gary Goddard is the founder of the **British** Goddard Entertainment .
When the kernel gets very hot - above 212 degrees Fahrenheit - the water inside turns to steam , which is a gas .	When the kernel gets very hot .

Sentence Simplification

Simplification Pipeline

Source	The popcorn moves like a gymnast doing a somersault .
Lexical followed by Syntactic Simplification	The **corn** moves like an athlete doing a flip .
Syntactic followed by Lexical Simplification	The **ice** moves like an athlete .

Model Scores
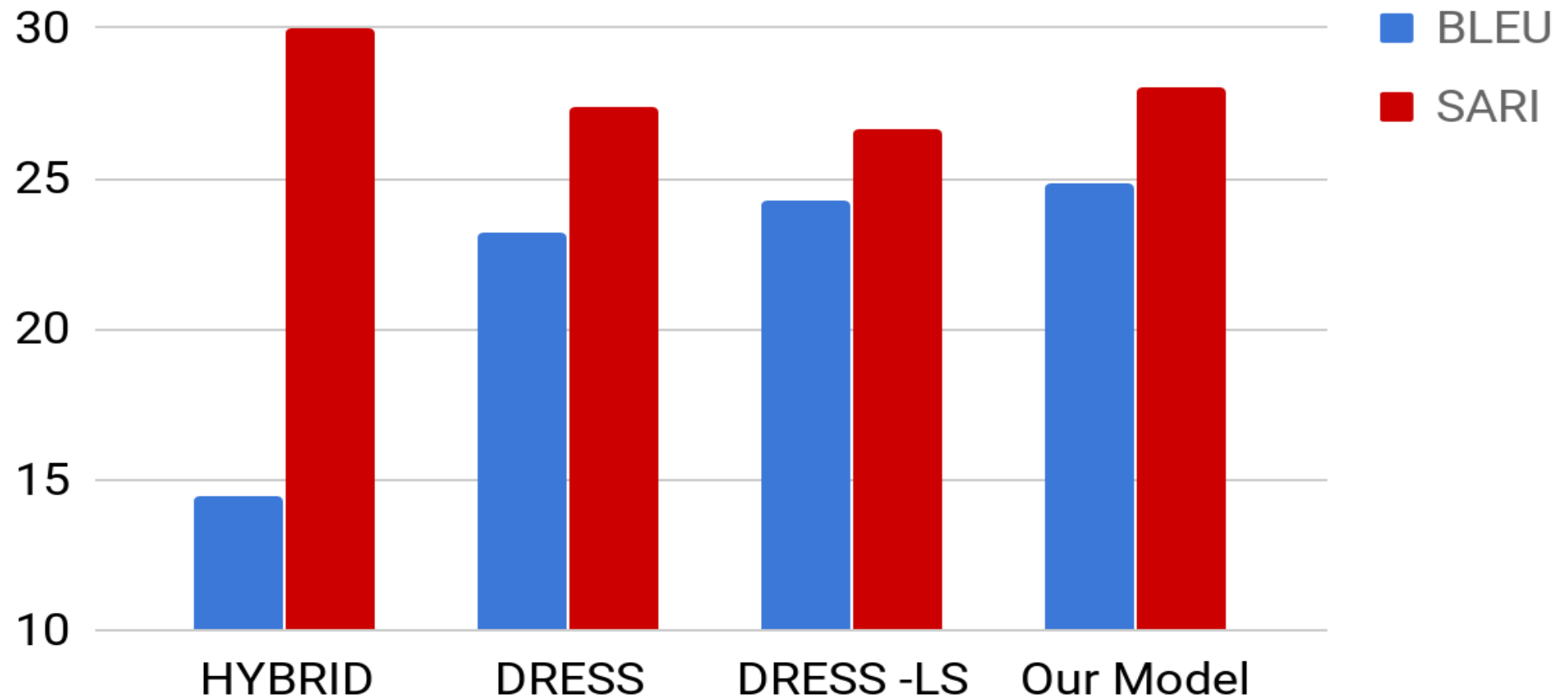
Summarization

Extractive Summarization

'Alcohol' is taken in almost all cool and cold climates, and to a very much less extent in hot ones. Thus, it is taken by people who live in the Himalaya Mountains, but not nearly so much by those who live in the plains of India. *Alcohol is not necessary in any way to anybody. The regular use of alcohol, even in small quantities, tends to cause mischief in many ways to various organs of the body.* It affects the liver, it weakens the mental powers, and lessens the general energy of the body. In addition damage to the central nervous system and peripheral nervous system can occur from chronic alcohol abuse.

Extractive Summarization

'Alcohol' is taken in almost all cool and cold climates, and to a very much less extent in hot ones. Alcohol is not necessary in any way to anybody.The regular use of alcohol, even in small quantities, tends to cause mischief in many ways to various organs of the body.

Algorithm

➜ TextRank algorithm
➜ Graph-based
➜ Rank sentences
➜ Chooses sentences

according to compression

ratio
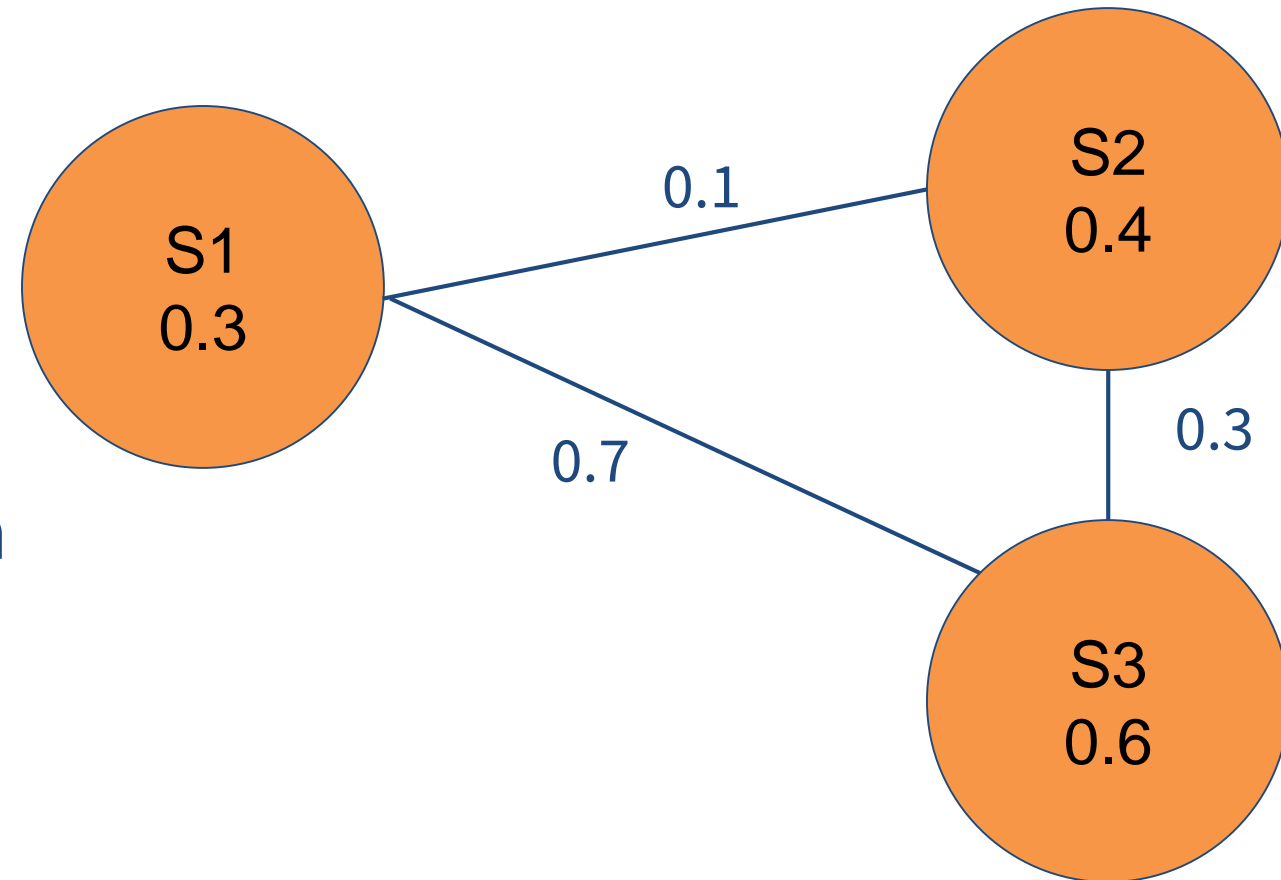
Results

➜ ROUGE score metric
➜ 500 BBC news article

Our implementation	Gensim
0.7812	0.7293

Complexity Assessment

A succession of steadily more powerful and flexible computing devices were constructed in the 1930s and 1940s , gradually adding the key features that are seen in modern computers .

(0)

More powerful and simpler computers were made in the 1930s and 1940s.

(3)

| 0: Most Complex   4: Most Simple |

It is a transportation , commercial , and manufacturing centre , and has numerous industrial companies , principally geared towards shipbuilding .
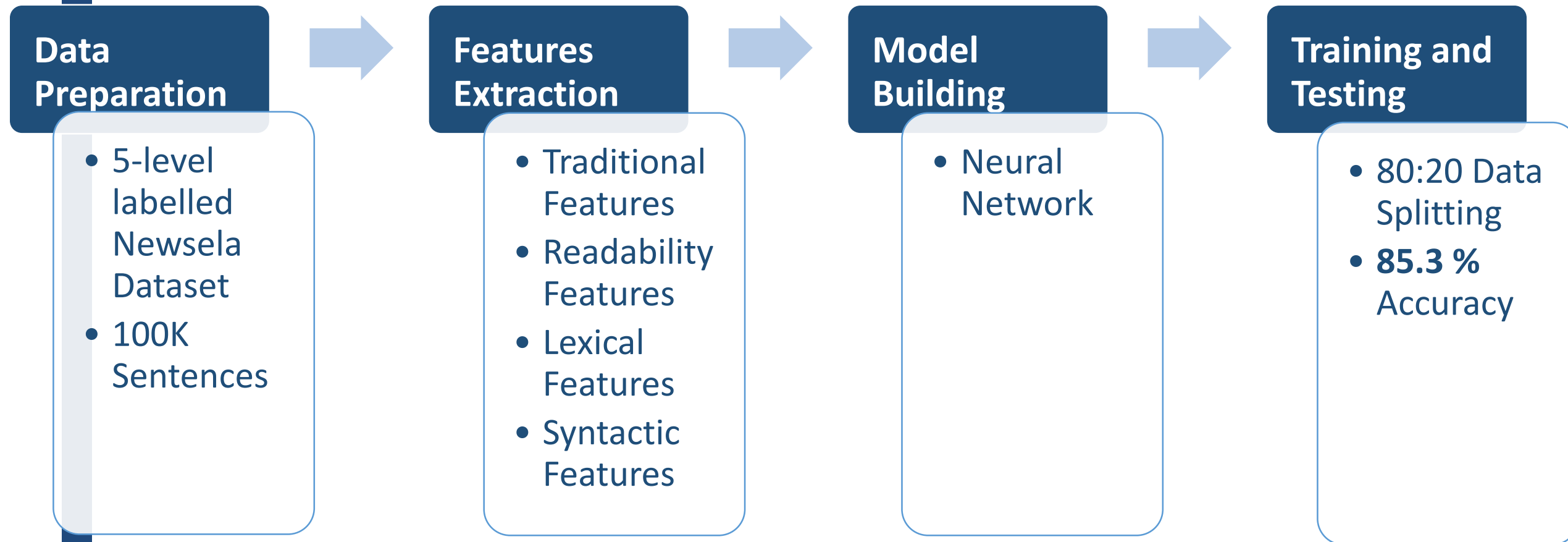
(2)

It is a transportation , business and manufacturing centre.

(4)

0: Most Complex   4: Most Simple

Pipeline

Data Preparation

- 5-level labelled Newsela Dataset
- 100K Sentences

→

Features Extraction

- Traditional Features
- Readability Features
- Lexical Features
- Syntactic Features

→

Model Building

- Neural Network

→

Training and Testing

- 80:20 Data Splitting
- **85.3 %** Accuracy

Complexity Assessment Before and After Simplification

Test Set Sentences

— Original — Simple
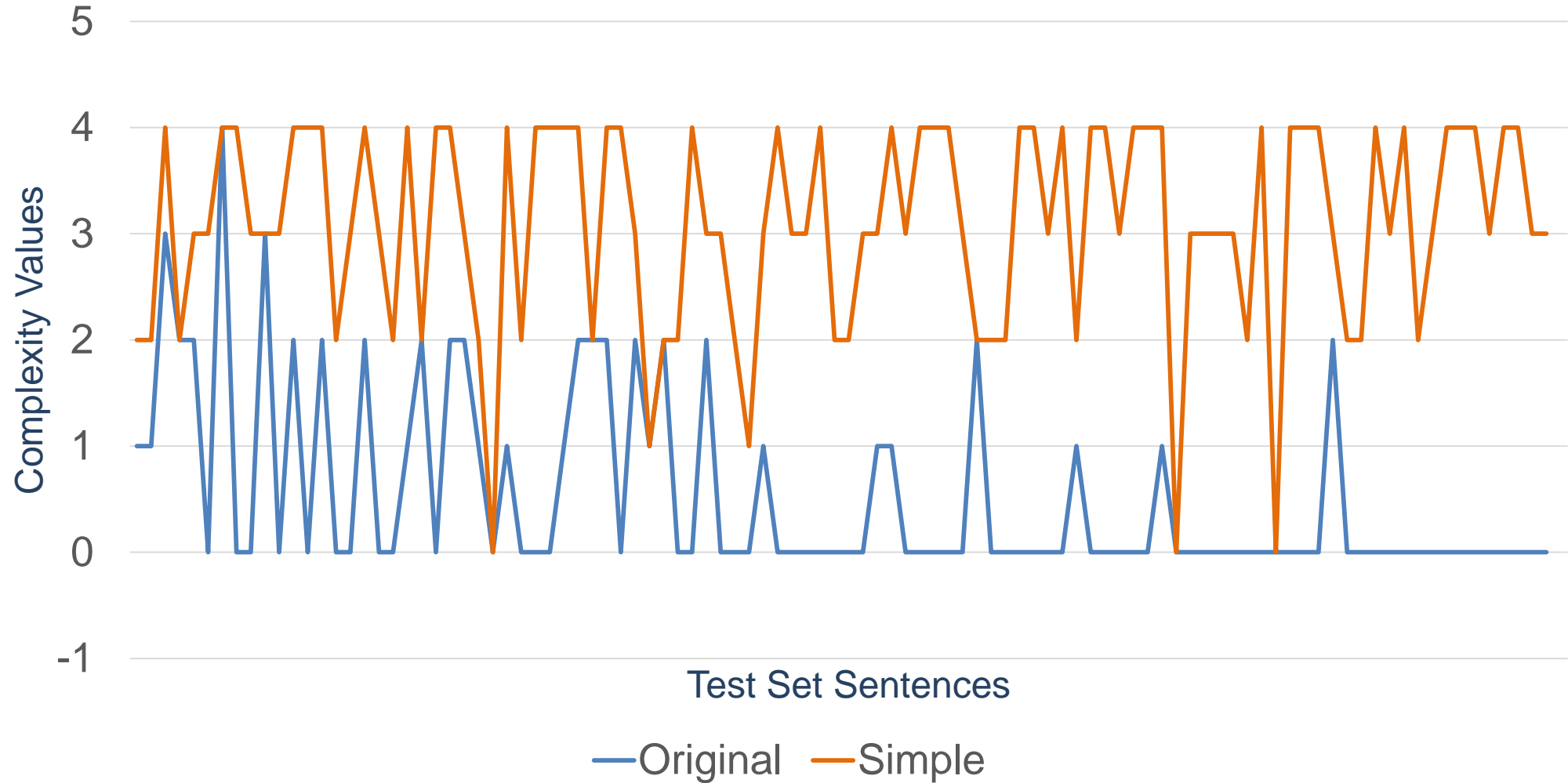
0: Most Complex 4: Most Simple

Demo

Conclusion

Lexical Simplification — Surpassed the state of the art ranker by 12.8%

Syntactic Simplification — Exceeded DRESS BLEU score

Summarization — Outperformed Genism Library by 0.07

Work-Load Division

| Amira Amer	Dina Adib	Rana Afifi	Salma Hanafy

• Summarization Module
• Experiments in Arabic
• <u>Syntactic Simplification:</u>
  • Reinforcement Learning
  • Beam Search
  • Experiments
	• GloVe
	• Interface
	• Integration
	• <u>Lexical Simplification:</u>
	  • Substitution Generation
	  • Evaluation
		• Complexity Assessment
		• ROUGE and BLEU Scores Implementation
		• <u>Lexical Simplification:</u>
		  • Feature Extraction
		  • Complex Word Identification
			• <u>Syntactic Simplification:</u>
			• Dataset Preparation
			• Encoder-Decoder Model
			• Attention Mechanism
			• Supervised Learning
			• Unknown Words Replacement
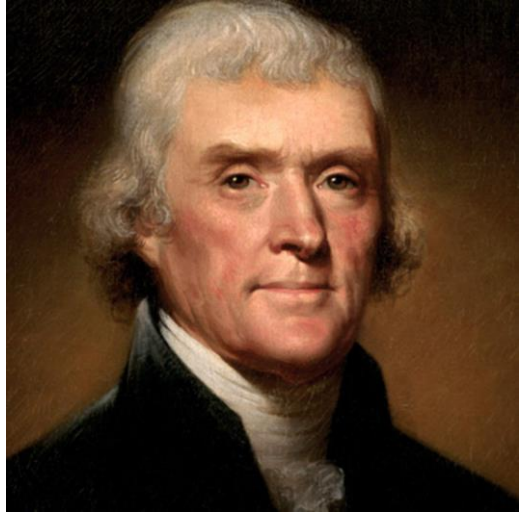
"The most valuable of all talents is that of never using two words when one will do"

-Thomas Jefferson, 3rd president of the United States