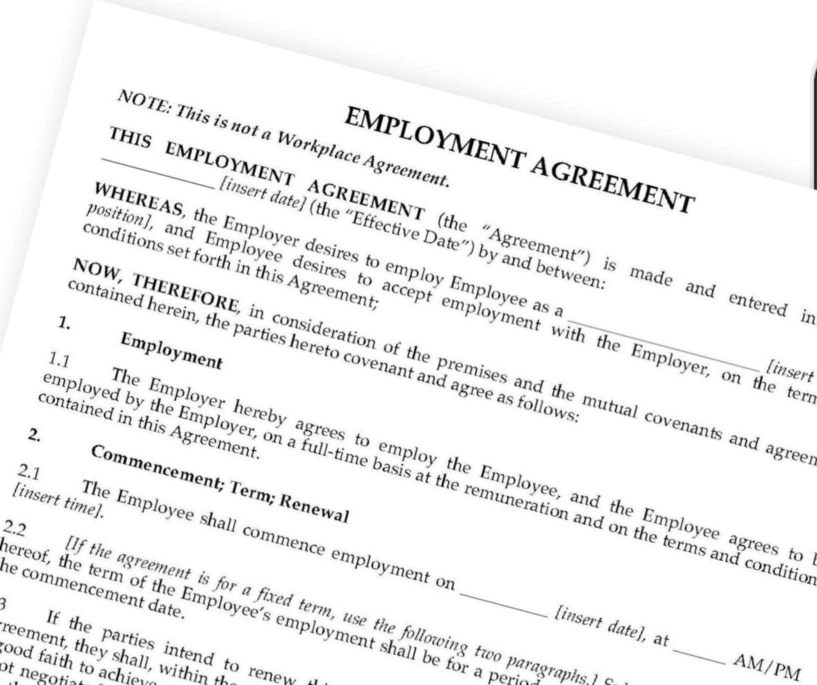
MUSS: Multilingual Unsupervised Sentence Simplification by Mining Paraphrases

Louis Martin, Angela Fan, Éric de la Clergerie, Antoine Bordes, Benoît Sagot

Meta AI Paris & Inria Paris

Access to Information is Hard



EMPLOYMENT AGREEMENT

NOTE: This is not a Workplace Agreement.

THIS EMPLOYMENT AGREEMENT (the "Agreement") is made and entered into as of ________ *[insert date]* (the "Effective Date") by and between:

WHEREAS, the Employer desires to employ Employee as a ________ *[insert name of position]*, and Employee desires to accept employment with the Employer, on the terms and conditions set forth in this Agreement;

NOW, THEREFORE, in consideration of the premises and the mutual covenants and agreements contained herein, the parties hereto covenant and agree as follows:

1. **Employment**

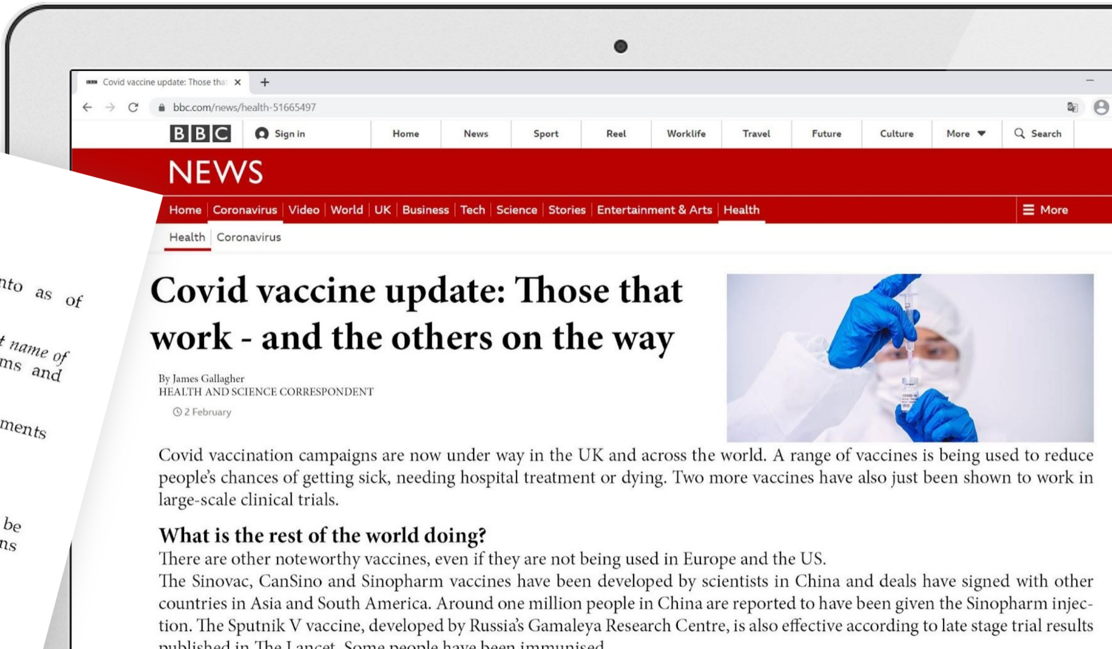
1.1 The Employer hereby agrees to employ the Employee, and the Employee agrees to be employed by the Employer, on a full-time basis at the remuneration and on the terms and conditions contained in this Agreement.

2. **Commencement; Term; Renewal**

2.1 The Employee shall commence employment on ________ *[insert date]*, at ____ AM/PM *[insert time]*.

2.2 *[If the agreement is for a fixed term, use the following two paragraphs.]* S
hereof, the term of the Employee's employment shall be for a period
he commencement date.

3 If the parties intend to renew th
reement, they shall, within th
ood faith to achieve
ot negotiat



Covid vaccine update: Those that
bbc.com/news/health-51665497
BBC
Sign in
Home
News
Sport
Reel
Worklife
Travel
Future
Culture
More
Search
NEWS
Home | Coronavirus | Video | World | UK | Business | Tech | Science | Stories | Entertainment & Arts | Health
More
Health | Coronavirus

Covid vaccine update: Those that work - and the others on the way

By James Gallagher
HEALTH AND SCIENCE CORRESPONDENT

2 February

Covid vaccination campaigns are now under way in the UK and across the world. A range of vaccines is being used to reduce people's chances of getting sick, needing hospital treatment or dying. Two more vaccines have also just been shown to work in large-scale clinical trials.

What is the rest of the world doing?

There are other noteworthy vaccines, even if they are not being used in Europe and the US.

The Sinovac, CanSino and Sinopharm vaccines have been developed by scientists in China and deals have signed with other countries in Asia and South America. Around one million people in China are reported to have been given the Sinopharm injection. The Sputnik V vaccine, developed by Russia's Gamaleya Research Centre, is also effective according to late stage trial results published in The Lancet. Some people have been immunised

… and many People struggle with Reading Difficulties

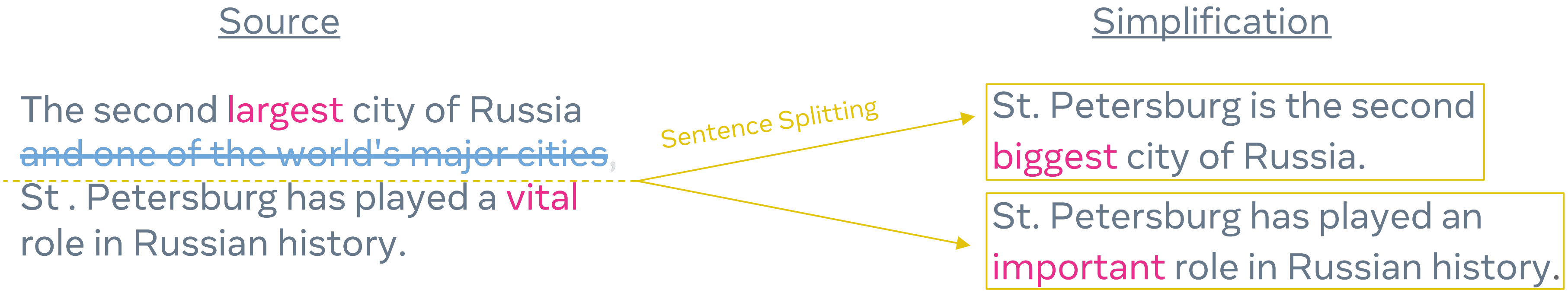
- Intellectual Disabilities
- Low literacy
- Non-native speakers

How can we make information easier to read and comprehend for each and everyone?

Automatic Sentence Simplification

Goal: Simplify a sentence while preserving its meaning

A Typical Human Simplification



Source

The second largest city of Russia ~~and one of the world's major cities,~~ St . Petersburg has played a vital role in Russian history.

Sentence Splitting

Simplification

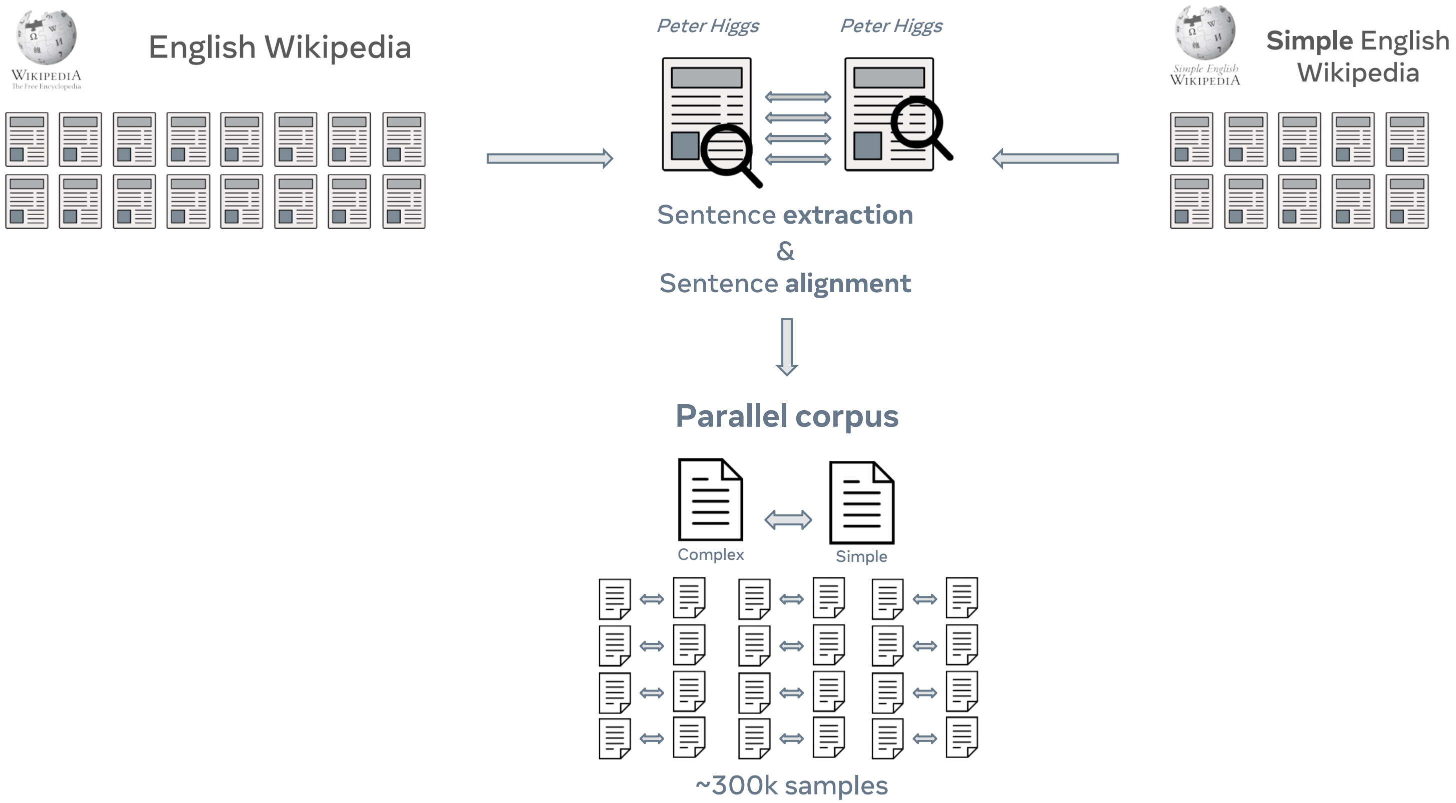
St. Petersburg is the second biggest city of Russia.

St. Petersburg has played an important role in Russian history.

- **Lexical Simplification** - Replace uncommon words
- **Syntactic Simplification** - Simplify complex syntactic structures
- **Compression** - Retain key information only

How to train sentence simplification models?

Traditional Simplification Datasets



English Wikipedia

Peter Higgs

Peter Higgs

Sentence **extraction**
&
Sentence **alignment**

Simple English Wikipedia

Parallel corpus

Complex

Simple

~300k samples

Problems with Simple English Wikipedia Alignment

- Contains alignment errors.
- Encyclopedic Domain only.
- Simple English Wikipedia **only in English**.

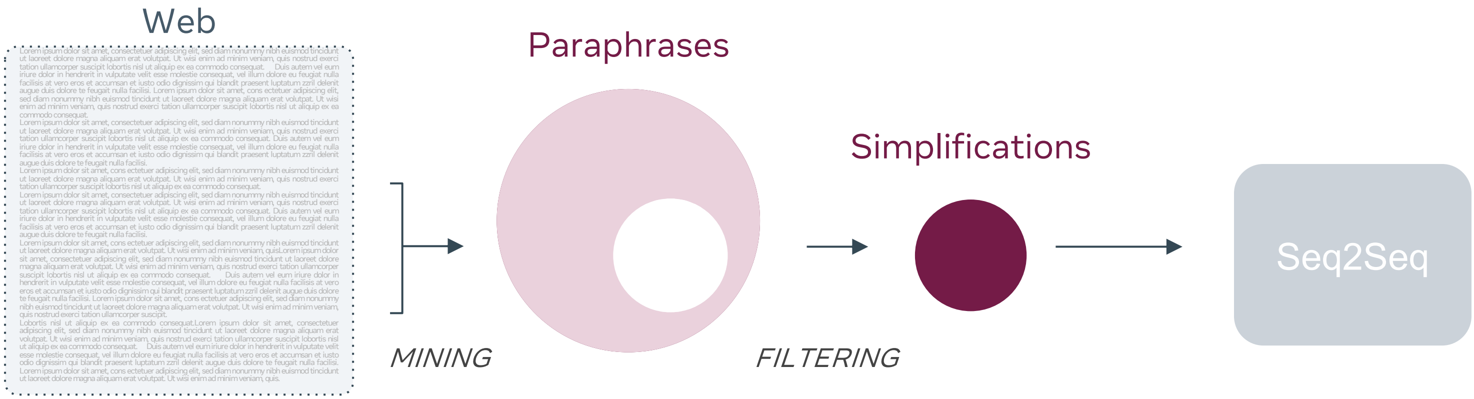
Can we find Parallel Sentences in any Language on the **Web**?

MUSS: Multilingual Unsupervised Sentence Simplification by Mining Paraphrases

Louis Martin, Angela Fan, Éric de la Clergerie, Antoine Bordes, Benoît Sagot

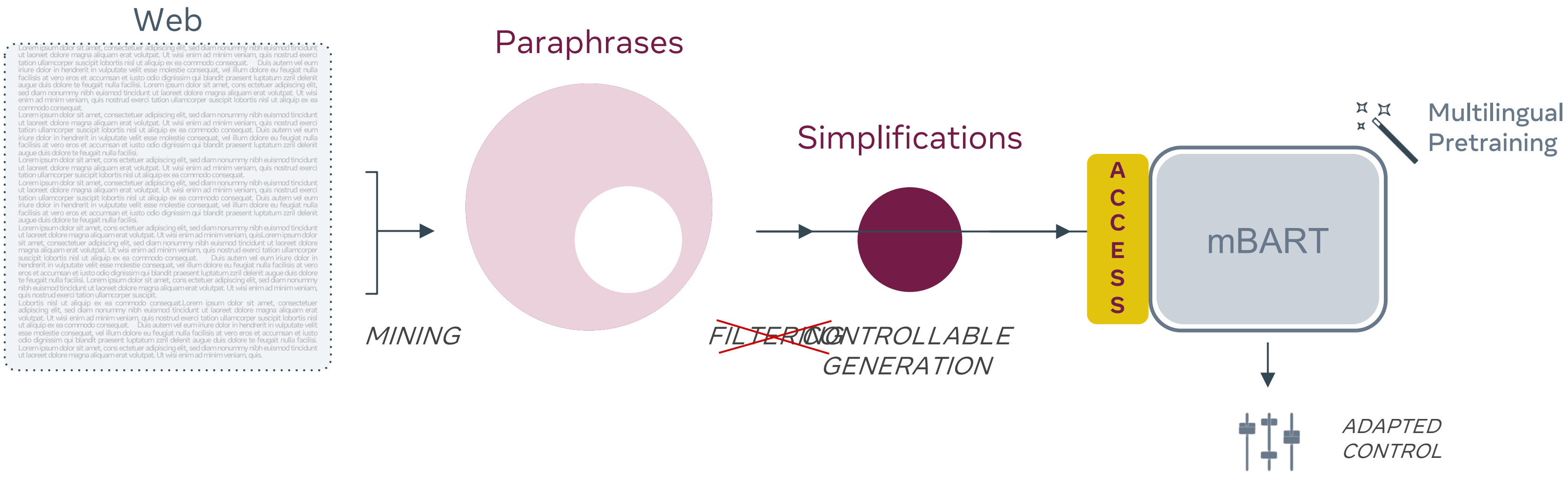
FACEBOOK AI

Mining Simplifications on the Web



What if we didn’t need this extra filtering step?

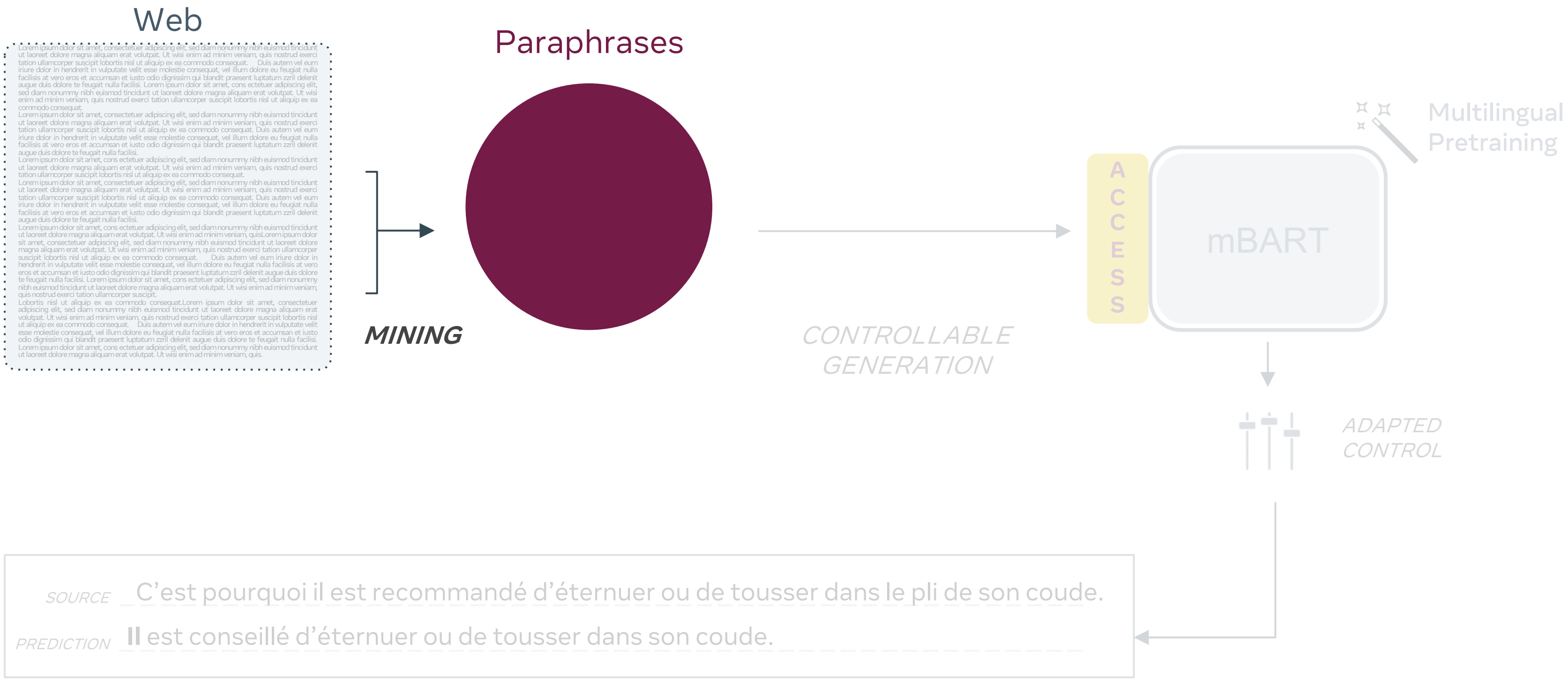
The MUSS Approach



SOURCE C’est pourquoi il est recommandé d’éternuer ou de tousser dans le pli de son coude.

PREDICTION **Il** est **donc conseillé** d'éternuer ou de tousser dans son coude / dans le **pli du** coude.

Mine Paraphrases



SOURCE C’est pourquoi il est recommandé d’éternuer ou de tousser dans le pli de son coude.

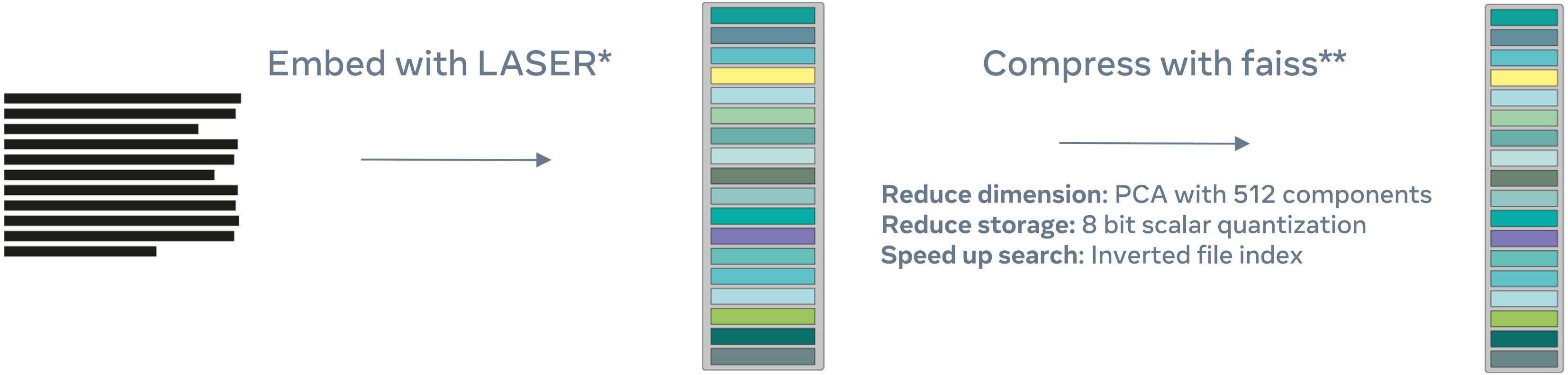
PREDICTION Il est conseillé d’éternuer ou de tousser dans son coude.

Mine Paraphrases

Idea: Mine Paraphrases using Sentence Embeddings

Paraphrases = **Nearest Neighbours** in Embedding Space

Index Creation



Embed with LASER*

Compress with faiss**

Reduce dimension: PCA with 512 components
Reduce storage: 8 bit scalar quantization
Speed up search: Inverted file index

1B extracted sequences
(180GB of text)

Latent representation
(dim = 1024)

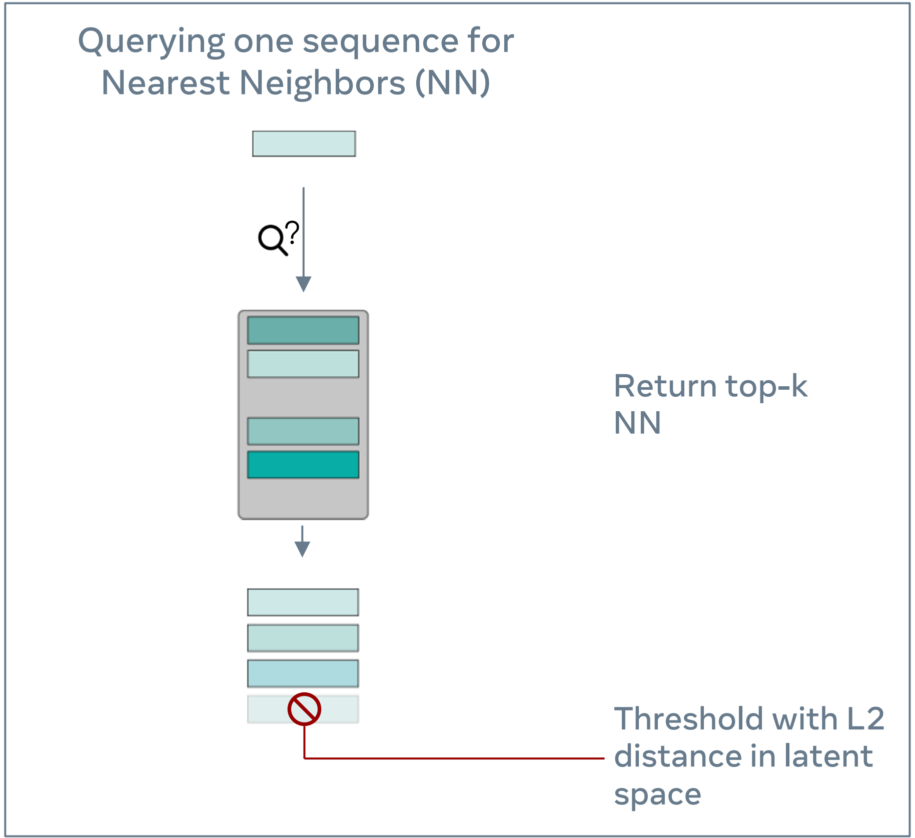
Optimized faiss index
(~500GB per language on disk)

*LASER: Multilingual sentence embeddings model
**faiss: Fast nearest neighbour search library

Paraphrase Mining

Optimized faiss index

(~500GB per language on disk)

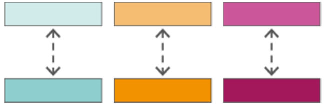


Querying one sequence for Nearest Neighbors (NN)

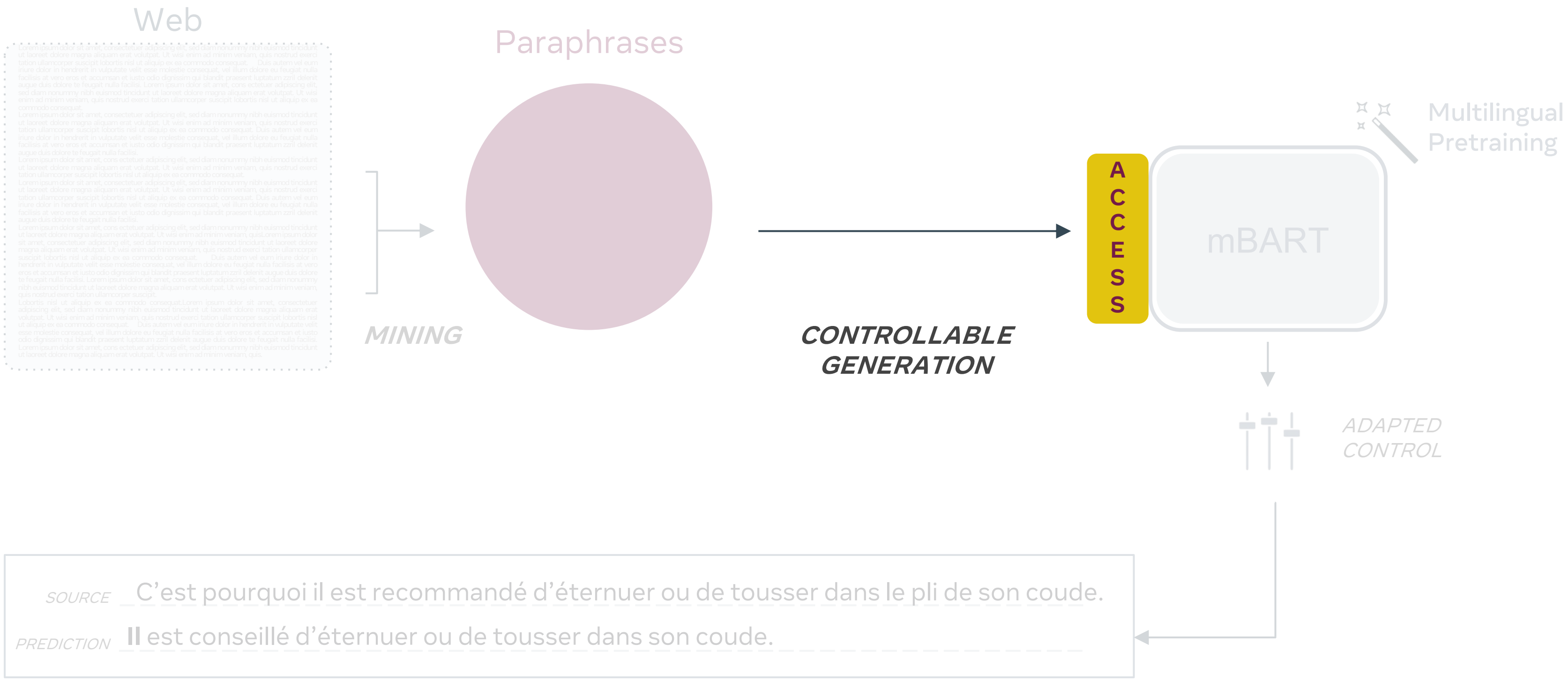
Return top-k NN

Threshold with L2 distance in latent space

Repeat in parallel for 1B queries
⇒ 1M extracted paraphrases



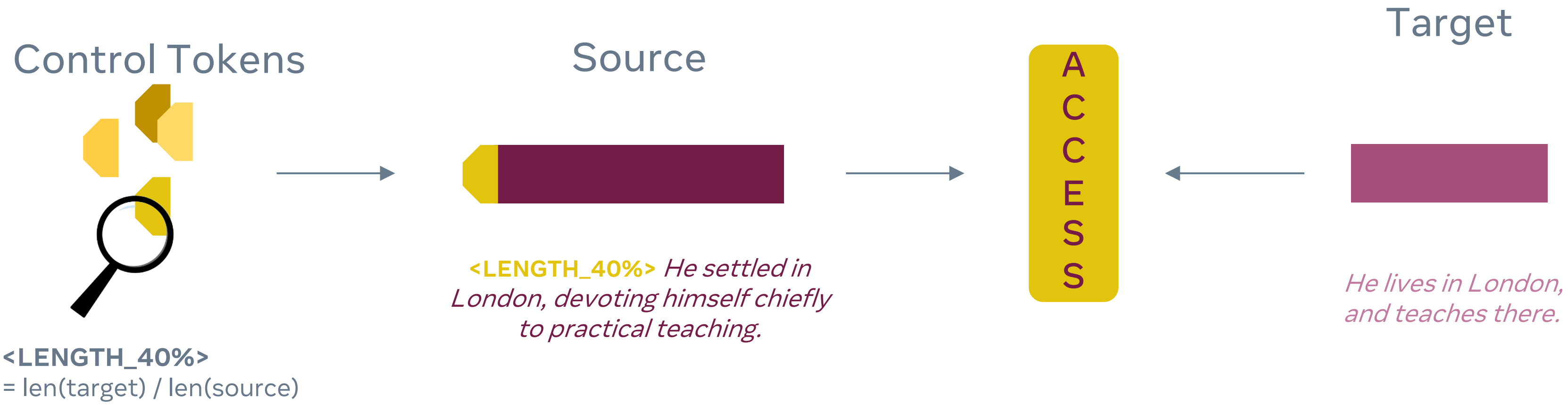
Use ACCESS for Controllable Generation



SOURCE C’est pourquoi il est recommandé d’éternuer ou de tousser dans le pli de son coude.

PREDICTION Il est conseillé d’éternuer ou de tousser dans son coude.

Conditioning on Control Tokens during Training



Control Tokens

Source

Target

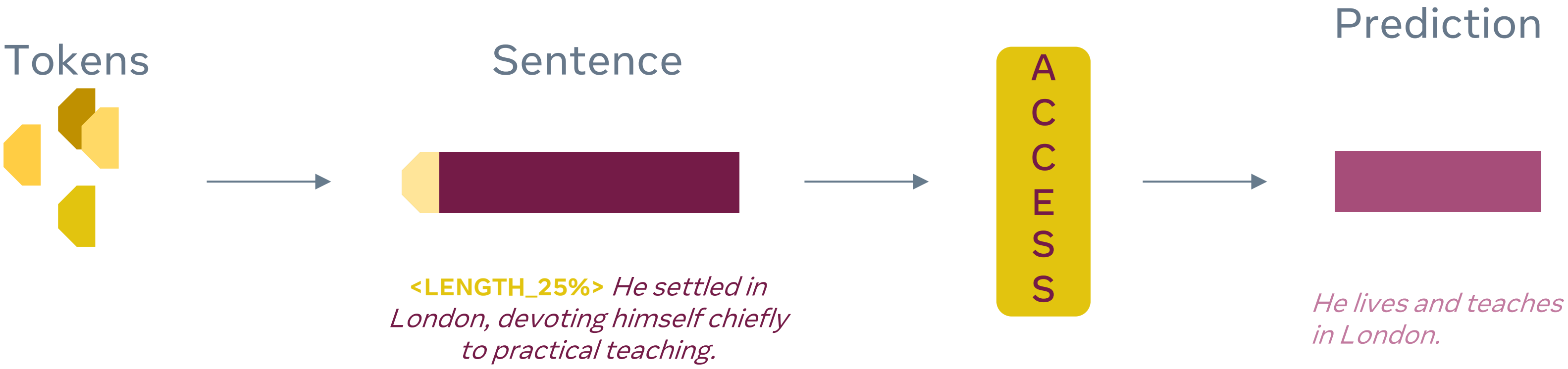
ACCESS

<LENGTH_40%> *He settled in London, devoting himself chiefly to practical teaching.*

He lives in London, and teaches there.

<LENGTH_40%>
= len(target) / len(source)

Choose Desired Length at Test Time



<LENGTH_25%> *He settled in London, devoting himself chiefly to practical teaching.*

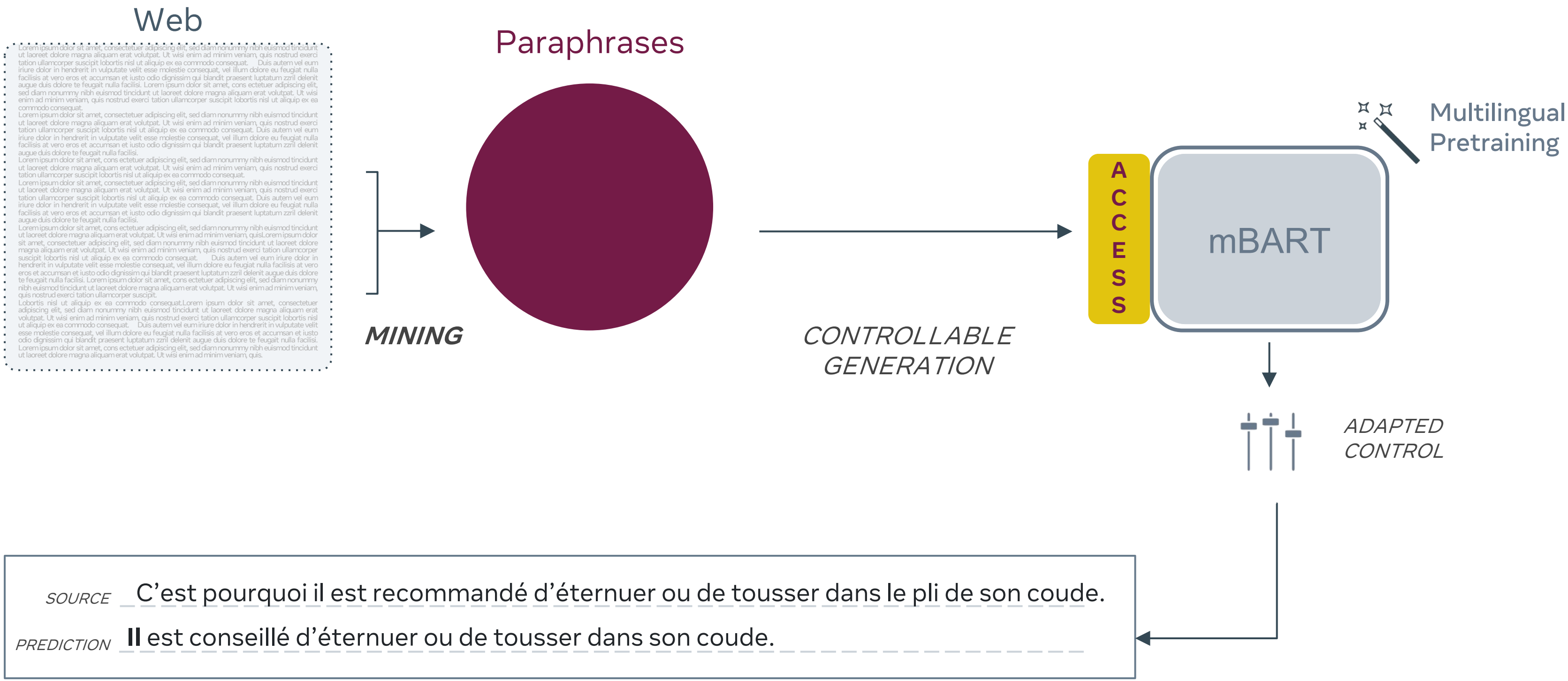
He lives and teaches in London.

In practice: **find ratio** with best results on valid

Condition on Many Attributes

- Length
- Lexical Complexity
- Syntactic Complexity
- Amount of Paraphrasing

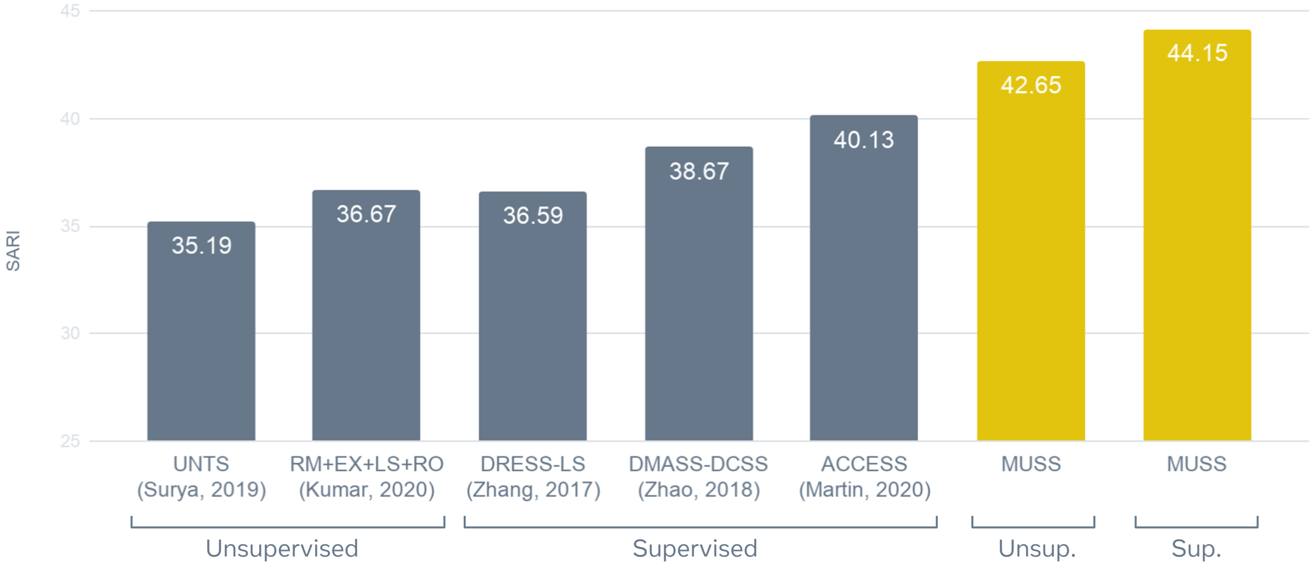
The MUSS Approach



SOURCE C’est pourquoi il est recommandé d’éternuer ou de tousser dans le pli de son coude.

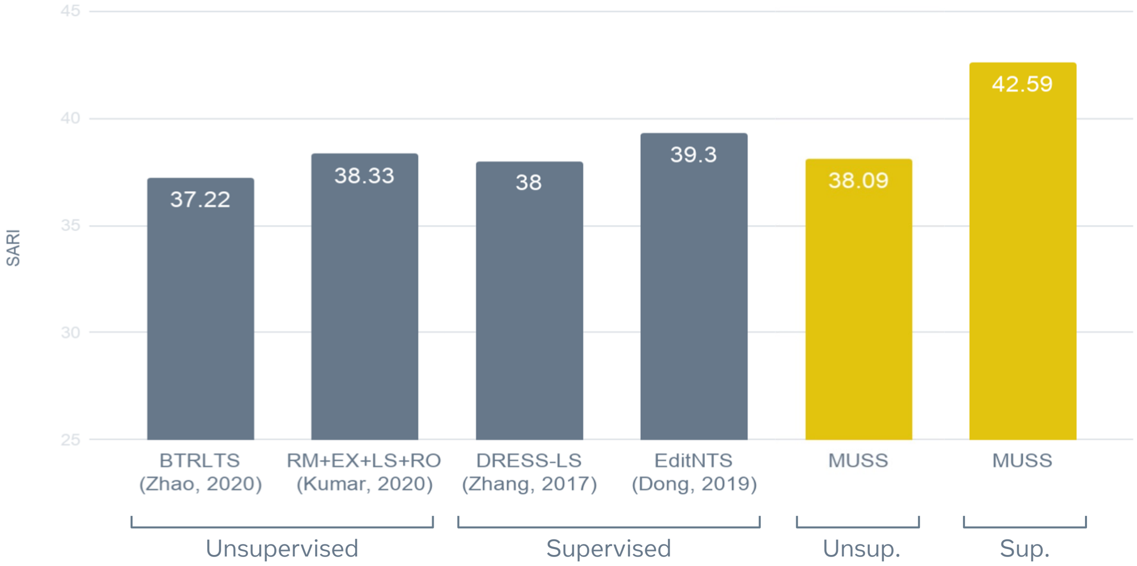
PREDICTION **Il** est conseillé d’éternuer ou de tousser dans son coude.

English Results - ASSET



- MUSS improves over previous methods
- Incorporating labelled data improves further.

English Results - Newsela



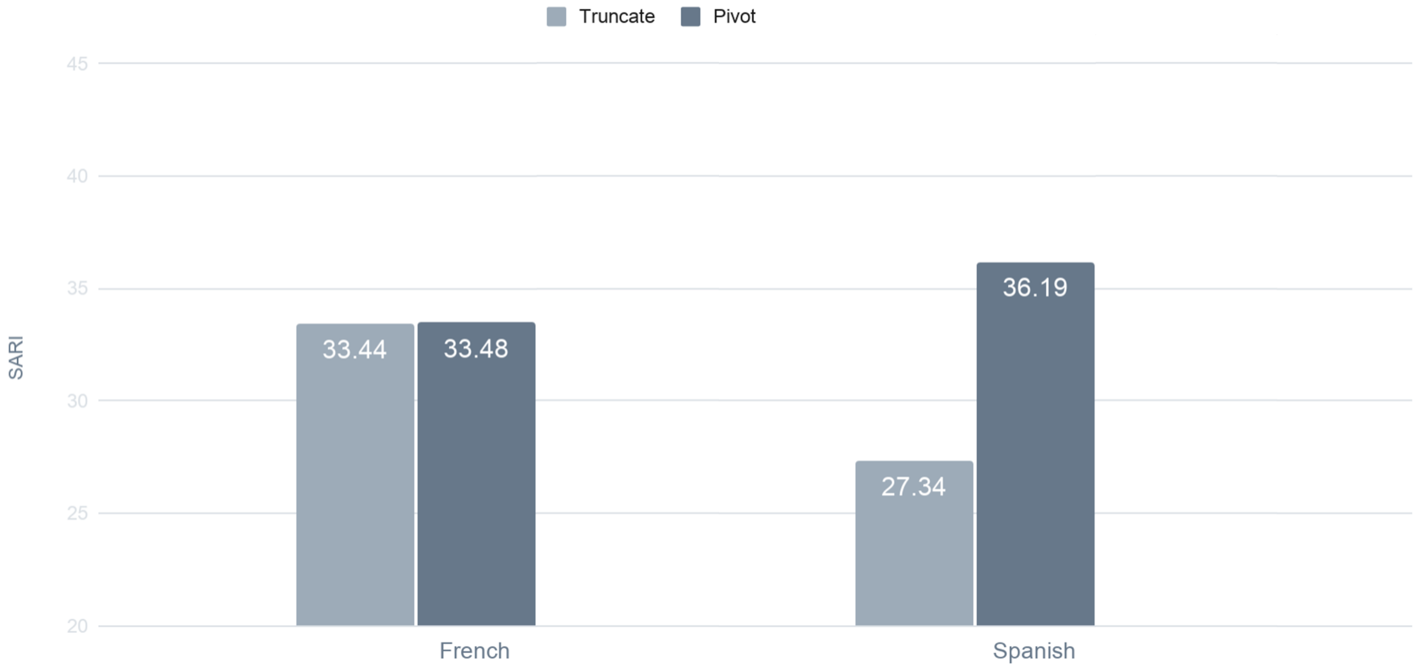
- In-domain data important for professional News Corpus

Multilingual Results

Baselines

- **Truncate**
  - Drop last 20% tokens.
- **Pivot**
  - **Fr**⇒**En** Translation - **En**⇒**En** Simplification - **En**⇒**Fr** Translation

Multilingual Results



- Good results compared to strong baselines…
- But benchmarks are still imperfect

MUSS Simplifications Example

Correct Simplification

It is situated at the coast of the Baltic Sea, where it encloses the city of Stralsund.
It is located on the Baltic Sea. The city of Stralsund is located in it.

Simplification Error

In 1998, Culver ran for Iowa Secretary of State and was victorious.
In 1998, Culver ran for Governor of Iowa and won.

Different meaning

Conclusion

MUSS: Paraphrase Mining + Controllable Generation

- Fully unsupervised sentence simplification
- Can be applied in any language

Perspectives

Towards document simplification

- How to mine full documents?
- Will controls be as effective?

Apply MUSS to other text rewriting tasks

- Paraphrasing
- Style transfer
- Summarization

Thank You