UNIVERSITÉ DE GENÈVE

UGA Université Grenoble Alpes

Simple, Simpler and Beyond:

A Fine-Tuning BERT-Based Approach to Enhance Sentence Complexity Assessment for Text Simplification

Lucía Ormaechea[1,2], Nikos Tsourakis[1], Didier Schwab[2], Pierrette Bouillon[1] and Benjamin Lecouteux[2]

[1] Department of Translation Technology – University of Geneva – Switzerland

[2] GETALP Team – University of Grenoble-Alpes – France

Overview

1. Introduction
2. Corpora
3. Fine-grained simplicity assessment method
4. Conclusions and further work

1. Introduction

What is automatic text simplification (ATS)?

- **A**utomatic **T**ext **S**implification (**ATS**) is an area of NLP that aims at automatically converting texts into **simpler variants**, by **reducing their linguistic complexity**, albeit **preserving their original meaning** [Horn *et al.*, 2014; Stajner, 2021].

ORIGINAL	The second ***largest*** city of China ***~~and one of the world's major cities~~***, Beijing has played ***a vital*** role in Chinese history.
SIMPLIFIED	***Beijing*** is the second ***biggest*** city of China. Beijing has played ***an important*** role in Chinese history.

What is automatic text simplification (ATS)?

- **A**utomatic **T**ext **S**implification (**ATS**) is an area of NLP that aims at automatically converting texts into **simpler variants**, by **reducing their linguistic complexity**, albeit **preserving their original meaning** [Horn *et al.*, 2014; Stajner, 2021].

ORIGINAL	The second ***largest*** city of China ~~***and one of the world's major cities***~~, Beijing has played ***a vital*** role in Chinese history.
SIMPLIFIED	***Beijing*** is the second ***biggest*** city of China. Beijing has played ***an important*** role in Chinese history.

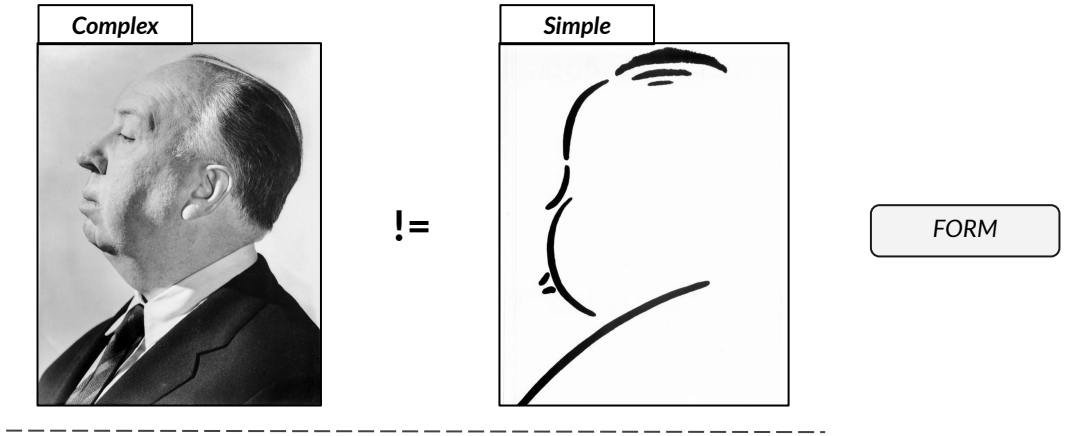
Complex

FORM

What is automatic text simplification (ATS)?

- **A**utomatic **T**ext **S**implification (**ATS**) is an area of NLP that aims at automatically converting texts into **simpler variants**, by **reducing their linguistic complexity**, albeit **preserving their original meaning** [Horn *et al.*, 2014; Stajner, 2021].

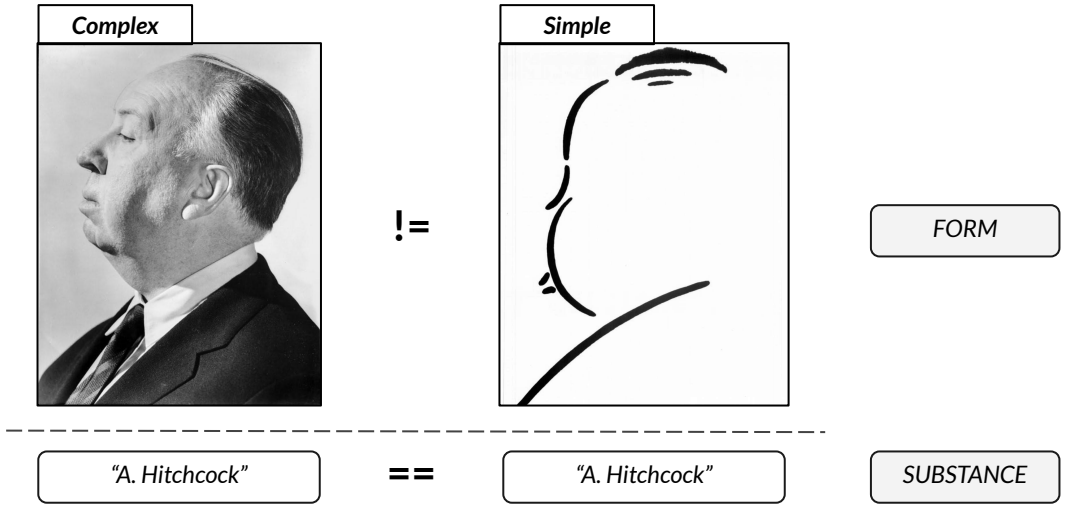
ORIGINAL	The second ***largest*** city of China ~~***and one of the world's major cities***~~, Beijing has played ***a vital*** role in Chinese history.
SIMPLIFIED	***Beijing*** is the second ***biggest*** city of China. Beijing has played ***an important*** role in Chinese history.



What is automatic text simplification (ATS)?

- **A**utomatic **T**ext **S**implification (**ATS**) is an area of NLP that aims at automatically converting texts into **simpler variants**, by **reducing their linguistic complexity**, albeit **preserving their original meaning** [Horn *et al.*, 2014; Stajner, 2021].

ORIGINAL	The second ***largest*** city of China ~~***and one of the world's major cities***~~, Beijing has played ***a vital*** role in Chinese history.
SIMPLIFIED	***Beijing*** is the second ***biggest*** city of China. Beijing has played ***an important*** role in Chinese history.



Sentence complexity assessment (SCA): an ancillary task to ATS

- **ATS**: interesting from a **text accessibility** and **comprehensibility aid perspective**. But the **scarcity of large-scale parallel monolingual data** prevents the advancement of the field, especially in **less resource-rich languages than English**.

Sentence complexity assessment (SCA): an ancillary task to ATS

- **ATS**: interesting from a **text accessibility** and **comprehensibility aid perspective**. But the **scarcity of large-scale parallel monolingual data** prevents the advancement of the field, especially in **less resource-rich languages than English**.
- Which points to the interest of **resorting to SCA** to obtain **complex-simple sentence pairs** from **comparable corpora** to **train ATS models**.

Two examples extracted from the French editions of Wikipedia and Vikidia.

Sentence complexity assessment (SCA): an ancillary task to ATS

- **ATS**: interesting from a **text accessibility** and **comprehensibility aid perspective**. But the **scarcity of large-scale parallel monolingual data** prevents the advancement of the field, especially in **less resource-rich languages than English**.
- Which points to the interest of **resorting to SCA** to obtain **complex-simple sentence pairs** from **comparable corpora** to **train ATS models**.

Two examples extracted from the French editions of Wikipedia and Vikidia.

- **Limitations** of **current approaches** to **SCA**:
  - ✗ **Operated in an absolute manner.**
  - ✗ **Overly coarse.**
  - ✗ **Not suited for ATS.**

Sentence complexity assessment (SCA): an ancillary task to ATS

- **ATS**: interesting from a **text accessibility** and **comprehensibility aid perspective**. But the **scarcity of large-scale parallel monolingual data** prevents the advancement of the field, especially in **less resource-rich languages than English**.
- Which points to the interest of **resorting to SCA** to obtain **complex-simple sentence pairs** from **comparable corpora** to **train ATS models**.

Two examples extracted from the French editions of Wikipedia and Vikidia.

- **Limitations** of **current approaches** to **SCA**:
  - ✗ **Operated in an absolute manner.**
  - ✗ **Overly coarse.**
  - ✗ **Not suited for ATS.**

SIMPLER ! = SIMPLE

Simplification is **inherently** a **relative process** → A given text is transformed into a relatively **simpler** version, which does not necessarily equate to **simple**.

Bridging the gap: introducing a finer-grained method for SCA

Our proposed solution

- Contribute with a **BERT-based finer-grained method** to assess **SCA**.
- Help as a **preliminary step** in creating **labeled simplification data**.

Bridging the gap: introducing a finer-grained method for SCA

Our proposed solution

- Contribute with a **BERT-based finer-grained method** to assess **SCA**.
- Help as a **preliminary step** in creating **labeled simplification data**.
- Specifically targeted for the **French language**.

Bridging the gap: introducing a finer-grained method for SCA

Our proposed solution

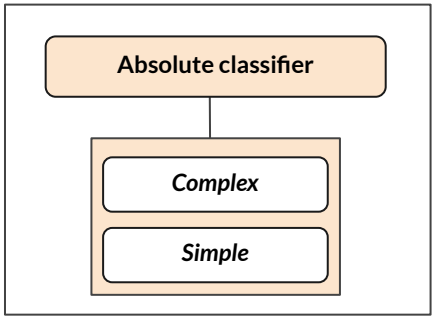
- Contribute with a **BERT-based finer-grained method** to assess **SCA**.
- Help as a **preliminary step** in creating **labeled simplification data**.
- Specifically targeted for the **French language**.
- Introduction of a new **triad** of **increasingly fine-grained models** so as to:

Bridging the gap: introducing a finer-grained method for SCA

Our proposed solution

- Contribute with a **BERT-based finer-grained method** to assess **SCA**.
- Help as a **preliminary step** in creating **labeled simplification data**.
- Specifically targeted for the **French language**.
- Introduction of a new **triad** of **increasingly fine-grained models** so as to:

1) Determine whether a sentence is complex or simple.



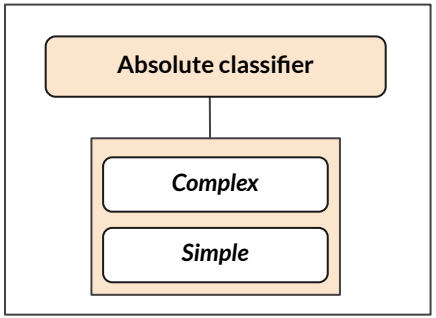
Increasingly fine-grained method for SCA

Bridging the gap: introducing a finer-grained method for SCA

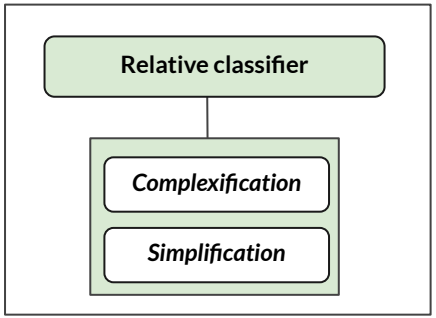
Our proposed solution

- Contribute with a **BERT-based finer-grained method** to assess **SCA**.
- Help as a **preliminary step** in creating **labeled simplification data**.
- Specifically targeted for the **French language**.
- Introduction of a new **triad** of **increasingly fine-grained models** so as to:

1) *Determine whether a sentence is complex or simple.*



2) *Assess if the second sentence in a pair is simpler than the first.*



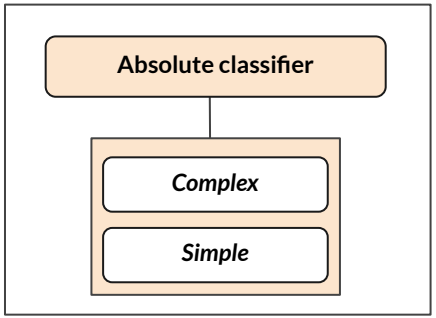
Increasingly fine-grained method for SCA

Bridging the gap: introducing a finer-grained method for SCA

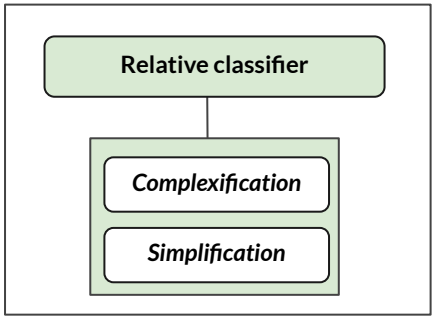
Our proposed solution

- Contribute with a **BERT-based finer-grained method** to assess **SCA**.
- Help as a **preliminary step** in creating **labeled simplification data**.
- Specifically targeted for the **French language**.
- Introduction of a new **triad** of **increasingly fine-grained models** so as to:

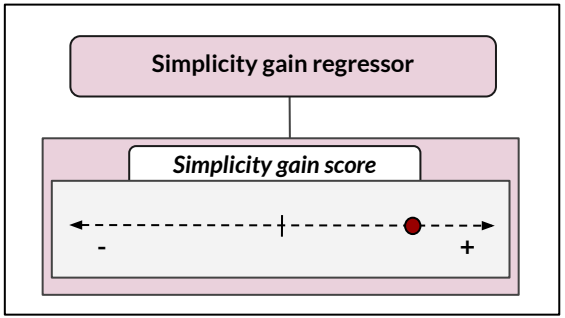
1) *Determine whether a sentence is complex or simple.*



2) *Assess if the second sentence in a pair is simpler than the first.*



3) *Measure the simplification gain achieved by the second sentence.*



Increasingly fine-grained method for SCA

2. Corpora

2.1) Choosing a dataset to train SCA models

To assess **sentence complexity** in a **data-driven manner**:

- Relied on **WikiLarge** parallel simplification dataset [Zhang & Lapata, 2017].
- **Originally in English**, monolingual.

2.1) Choosing a dataset to train SCA models

To assess **sentence complexity** in a **data-driven manner**:

- Relied on **WikiLarge** parallel simplification dataset [Zhang & Lapata, 2017].
- **Originally in English**, monolingual.
- Resort to **Google Translate** to obtain **French translations**.
- **Filtering** of too similar pairs → *lev* < 0.95

WikiLarge-Fr	
Train	105,420
Dev	13,177
Test	13,179

Overview of size (in sentence pairs).

2.1) Choosing a dataset to train SCA models

To assess **sentence complexity** in a **data-driven manner**:

- Relied on **WikiLarge** parallel simplification dataset [Zhang & Lapata, 2017].
- **Originally in English**, monolingual.
- Resort to **Google Translate** to obtain **French translations**.
- **Filtering** of too similar pairs → *lev* < 0.95

	WikiLarge-Fr
Train	105,420
Dev	13,177
Test	13,179

Overview of size (in sentence pairs).

2.2) Compiling Wikipedia-Vikidia data to implement SCA models (1/2)

- We relied on **register-diversified comparable corpora**:
  - **Wikipedia** → French edition.
  - **Vikidia** → Simplified version of the former.

2.1) Choosing a dataset to train SCA models

To assess **sentence complexity** in a **data-driven manner**:

- Relied on **WikiLarge** parallel simplification dataset [Zhang & Lapata, 2017].
- **Originally in English**, monolingual.
- Resort to **Google Translate** to obtain **French translations**.
- **Filtering** of too similar pairs → *lev* < 0.95

WikiLarge-Fr	
Train	105,420
Dev	13,177
Test	13,179

Overview of size (in sentence pairs).

2.2) Compiling Wikipedia-Vikidia data to implement SCA models (1/2)

- We relied on **register-diversified comparable corpora**:
  - **Wikipedia** → French edition.
  - **Vikidia** → Simplified version of the former.
- We scraped both **encyclopedias**, following the **pipeline** described in [Ormaechea & Tsourakis, 2023].

2.1) Choosing a dataset to train SCA models

To assess **sentence complexity** in a **data-driven manner**:

- Relied on **WikiLarge** parallel simplification dataset [Zhang & Lapata, 2017].
- **Originally in English**, monolingual.
- Resort to **Google Translate** to obtain **French translations**.
- **Filtering** of too similar pairs → *lev* < 0.95

WikiLarge-Fr	
Train	105,420
Dev	13,177
Test	13,179

Overview of size (in sentence pairs).

2.2) Compiling Wikipedia-Vikidia data to implement SCA models (1/2)

- We relied on **register-diversified comparable corpora**:
  - **Wikipedia** → French edition.
  - **Vikidia** → Simplified version of the former.
- We scraped both **encyclopedias**, following the **pipeline** described in [Ormaechea & Tsourakis, 2023].

Data acquisition	Wiki-texts	Viki-texts
# documents	34,806	
# sentences	165,806	134,348
# tokens	4,030,148	2,373,045

2.2) Compiling Wikipedia-Vikidia data to implement SCA models (2/2)

Meaning preservation pre-filtering step:

- **Why?**

2.2) Compiling Wikipedia-Vikidia data to implement SCA models (2/2)

Meaning preservation pre-filtering step:

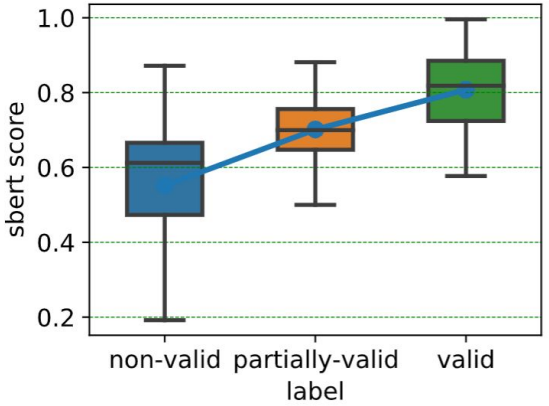
- **Why?**
- **SBERT** (Sentence-BERT) using **multilingual** sentence transformers*:
  - Generate fixed-length sentence **embeddings**.
  - **Compute cosine similarity** between ***Wiki***:***Viki*** sentences.

*https://huggingface.co/sentence-transformers/paraphrase-xlm-r-multilingual-v1

2.2) Compiling Wikipedia-Vikidia data to implement SCA models (2/2)

Meaning preservation pre-filtering step:

- **Why?**
- **SBERT** (Sentence-BERT) using **multilingual** sentence transformers*:
  - Generate fixed-length sentence **embeddings**.
  - **Compute cosine similarity** between ***Wiki***:***Viki*** sentences.
- **Manual annotation:**
  - Determine to which extent ***Wiki***:***Viki*** sentences conveyed the same meaning.
  - Definition of a **cutoff threshold** for the pairs that exhibit a **high semantic overlap**.

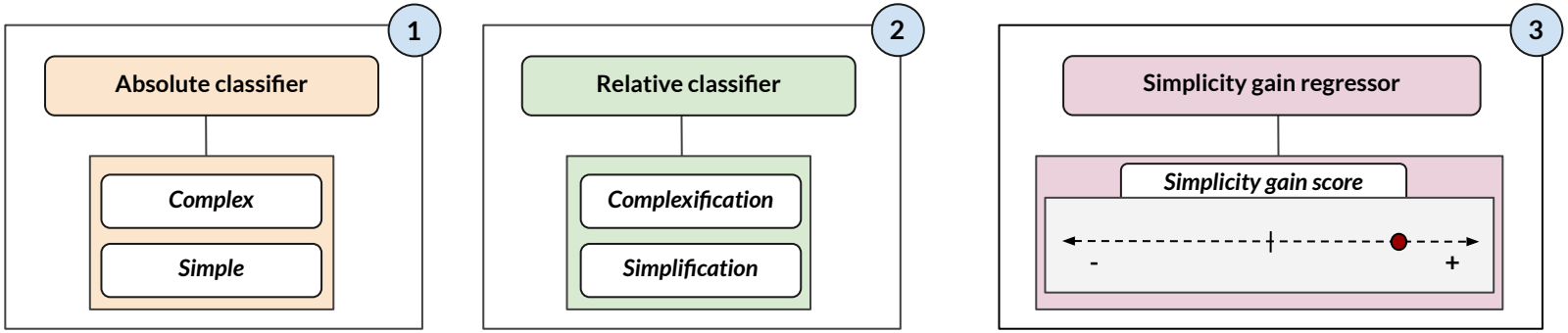


*https://huggingface.co/sentence-transformers/paraphrase-xlm-r-multilingual-v1

3. Fine-grained simplicity assessment method

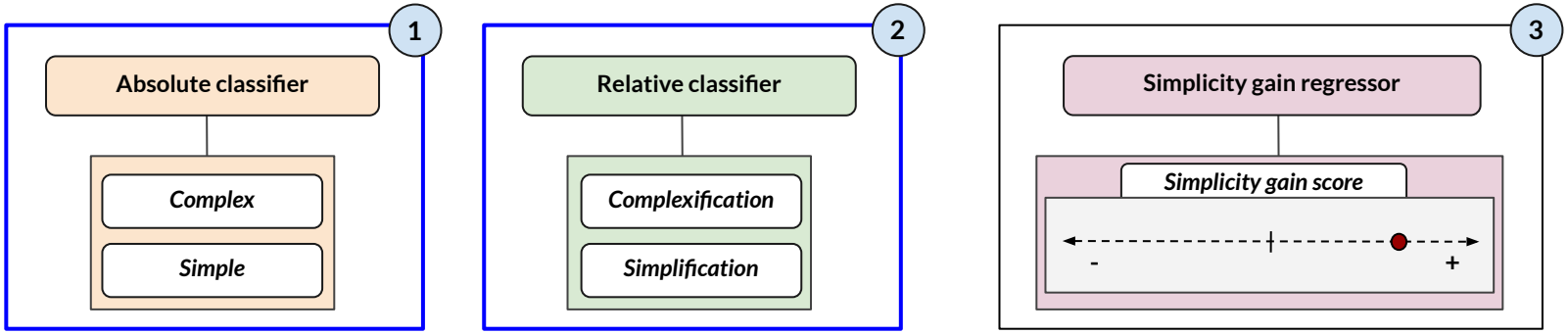
GOAL	Elicit relevant complex-simple pairs from Wikipedia-Vikidia compiled data.

MEANS	• Increasingly fine-grained simplicity assessment approach. • Fine-tuning with WikiLarge-FR, by the use of FlauBERT (*small*, *base*, *large*) [Le *et al.*, 2020].



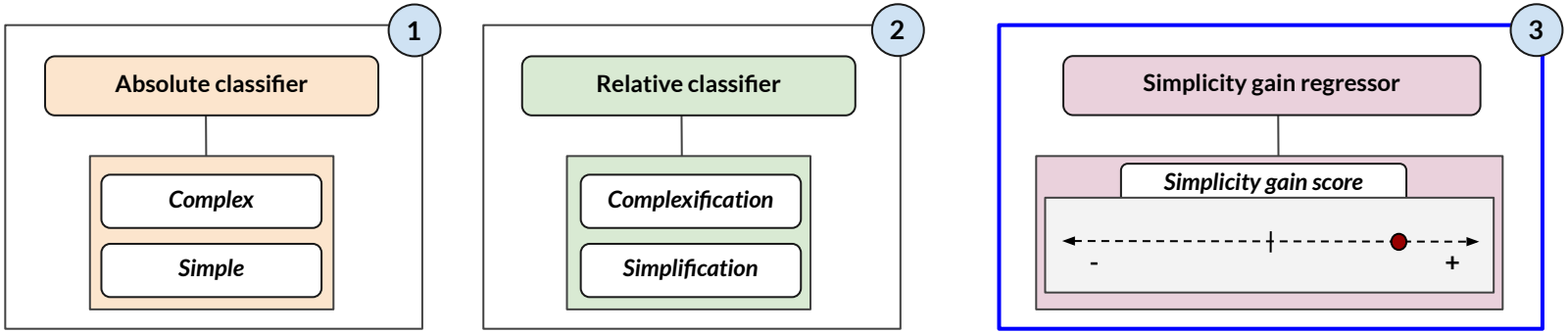
GOAL	Elicit relevant complex-simple pairs from Wikipedia-Vikidia compiled data.

MEANS	• Increasingly fine-grained simplicity assessment approach. • Fine-tuning with WikiLarge-FR, by the use of FlauBERT (*small*, *base*, *large*) [Le *et al.*, 2020].



GOAL	Elicit relevant complex-simple pairs from Wikipedia-Vikidia compiled data.

MEANS	• Increasingly fine-grained simplicity assessment approach. • Fine-tuning with WikiLarge-FR, by the use of FlauBERT (*small*, *base*, *large*) [Le *et al.*, 2020].



GOAL	Elicit relevant complex-simple pairs from Wikipedia-Vikidia compiled data.

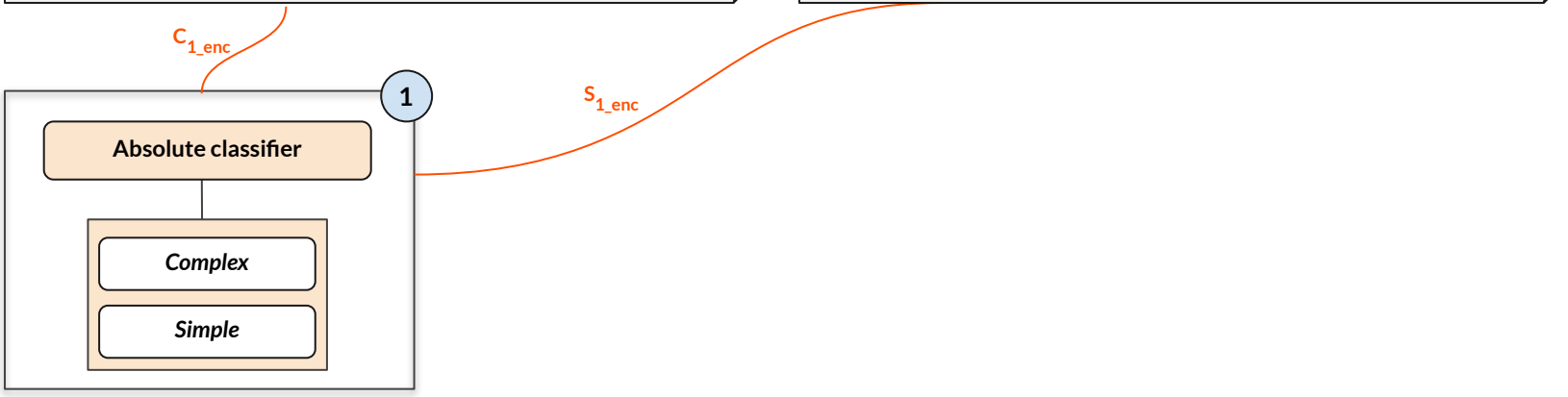
MEANS	• Increasingly fine-grained simplicity assessment approach. • Fine-tuning with WikiLarge-FR, by the use of FlauBERT (*small*, *base*, *large*) [Le *et al.*, 2020].

Complex sentence (c_1)

Juno est la patronne du mariage, et beaucoup de gens croient que le moment le plus favorable pour se marier est juin, le mois qui porte le nom de la déesse.

Simple sentence (s_1)

Beaucoup de gens considèrent le mois de juin, qui porte le nom de la déesse qui est la patronne du mariage, comme le moment le plus favorable pour se marier.



GOAL	Elicit relevant complex-simple pairs from Wikipedia-Vikidia compiled data.

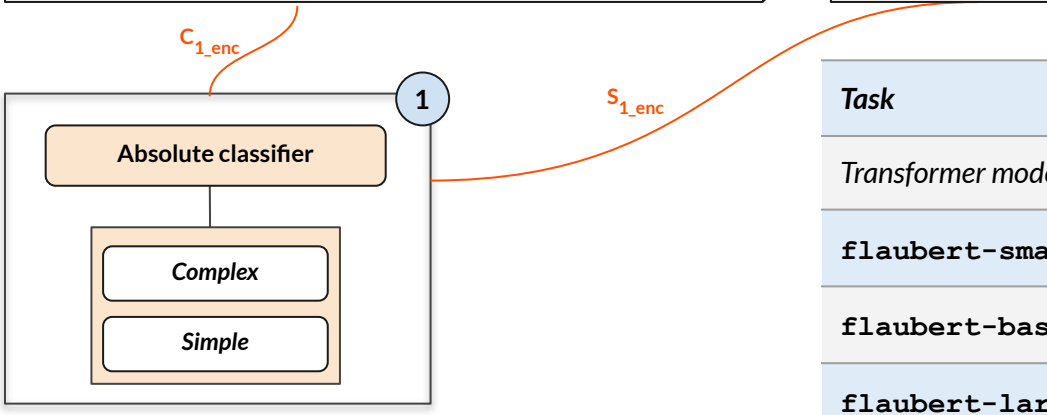
MEANS	• Increasingly fine-grained simplicity assessment approach. • Fine-tuning with WikiLarge-FR, by the use of FlauBERT (*small*, *base*, *large*) [Le *et al.*, 2020].

Complex sentence (c_1)

Juno est la patronne du mariage, et beaucoup de gens croient que le moment le plus favorable pour se marier est juin, le mois qui porte le nom de la déesse.

Simple sentence (s_1)

Beaucoup de gens considèrent le mois de juin, qui porte le nom de la déesse qui est la patronne du mariage, comme le moment le plus favorable pour se marier.

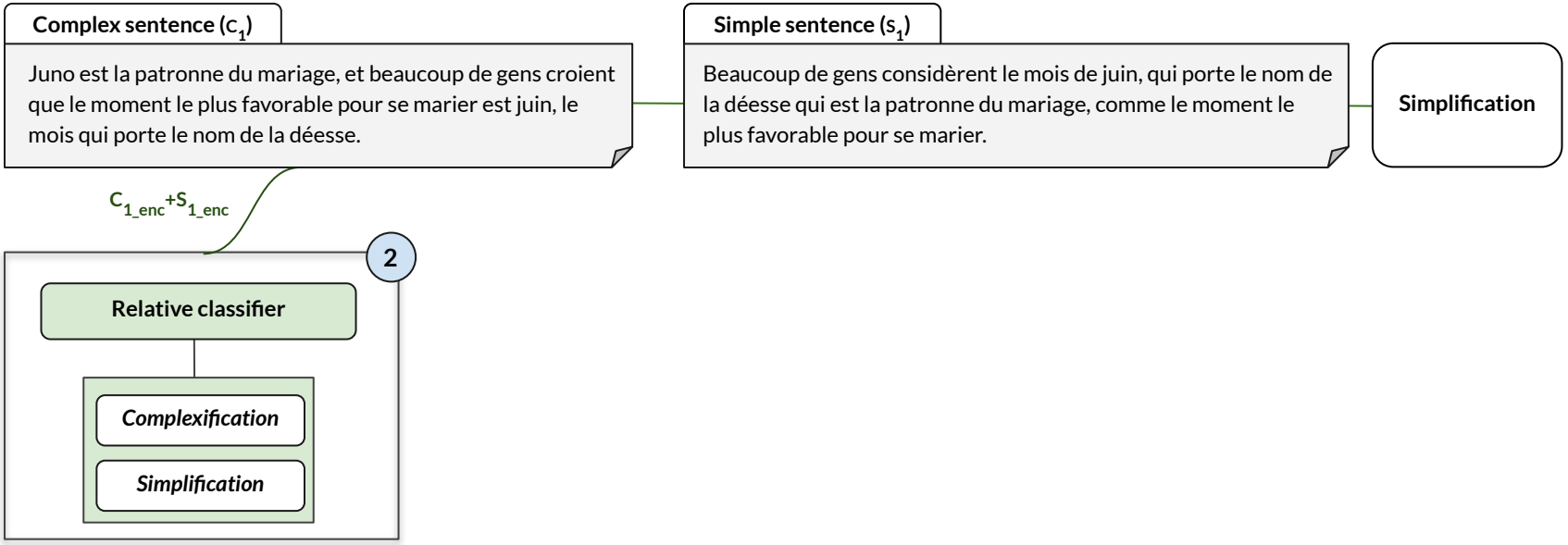


Task	Absolute classification	
Transformer model	*untuned*	*tuned*
`flaubert-small`	49.54	**70.11**
`flaubert-base`	50.97	**69.82**
`flaubert-large`	52.29	**69.19**

Accuracy results in % obtained for the absolute complexity classifier on the test set.

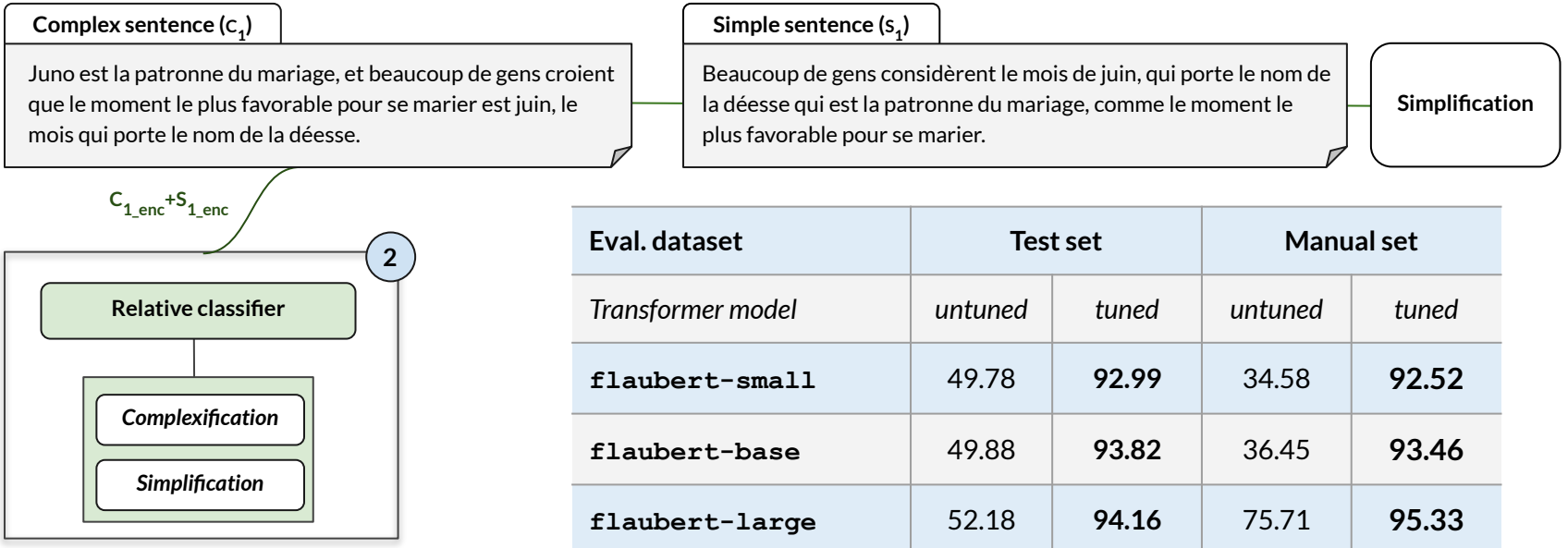
GOAL	Elicit relevant complex-simple pairs from Wikipedia-Vikidia compiled data.

MEANS	• Increasingly fine-grained simplicity assessment approach. • Fine-tuning with WikiLarge-FR, by the use of FlauBERT (*small*, *base*, *large*) [Le *et al.*, 2020].



GOAL	Elicit relevant complex-simple pairs from Wikipedia-Vikidia compiled data.

MEANS	• Increasingly fine-grained simplicity assessment approach. • Fine-tuning with WikiLarge-FR, by the use of FlauBERT (*small*, *base*, *large*) [Le *et al.*, 2020].

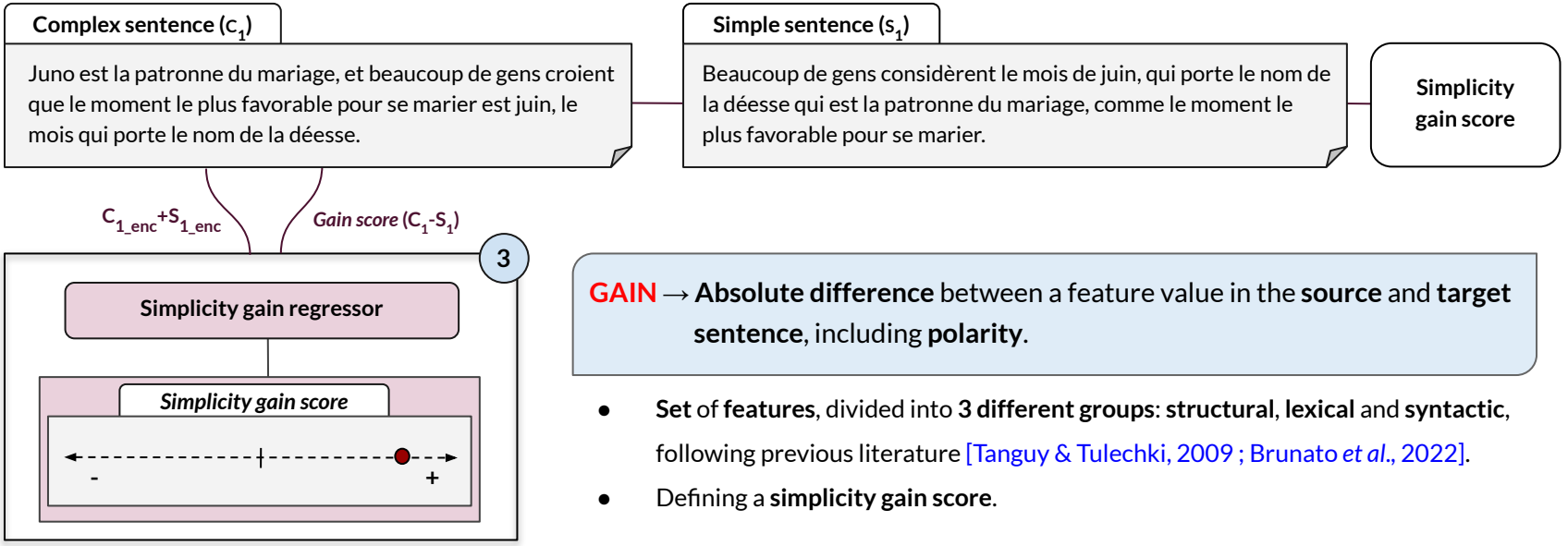


Eval. dataset	Test set		Manual set	
Transformer model	*untuned*	*tuned*	*untuned*	*tuned*
`flaubert-small`	49.78	**92.99**	34.58	**92.52**
`flaubert-base`	49.88	**93.82**	36.45	**93.46**
`flaubert-large`	52.18	**94.16**	75.71	**95.33**

Accuracy results in % obtained for the relative complexity classifier on the test and manual sets.

GOAL	Elicit relevant complex-simple pairs from Wikipedia-Vikidia compiled data.

MEANS	• Increasingly fine-grained simplicity assessment approach. • Fine-tuning with WikiLarge-FR, by the use of FlauBERT (*small*, *base*, *large*) [Le *et al.*, 2020].

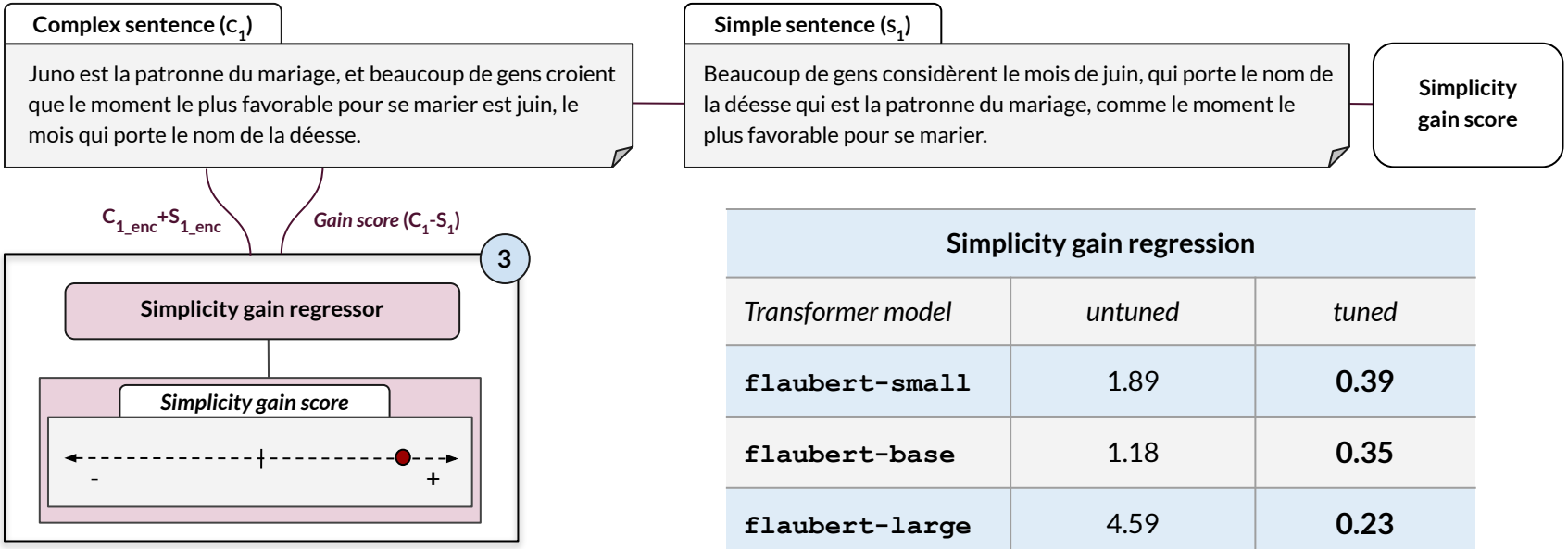


GAIN → **Absolute difference** between a feature value in the **source** and **target sentence**, including **polarity**.

- **Set** of **features**, divided into **3 different groups**: **structural**, **lexical** and **syntactic**, following previous literature [Tanguy & Tulechki, 2009 ; Brunato *et al.*, 2022].
- Defining a **simplicity gain score**.

GOAL	Elicit relevant complex-simple pairs from Wikipedia-Vikidia compiled data.

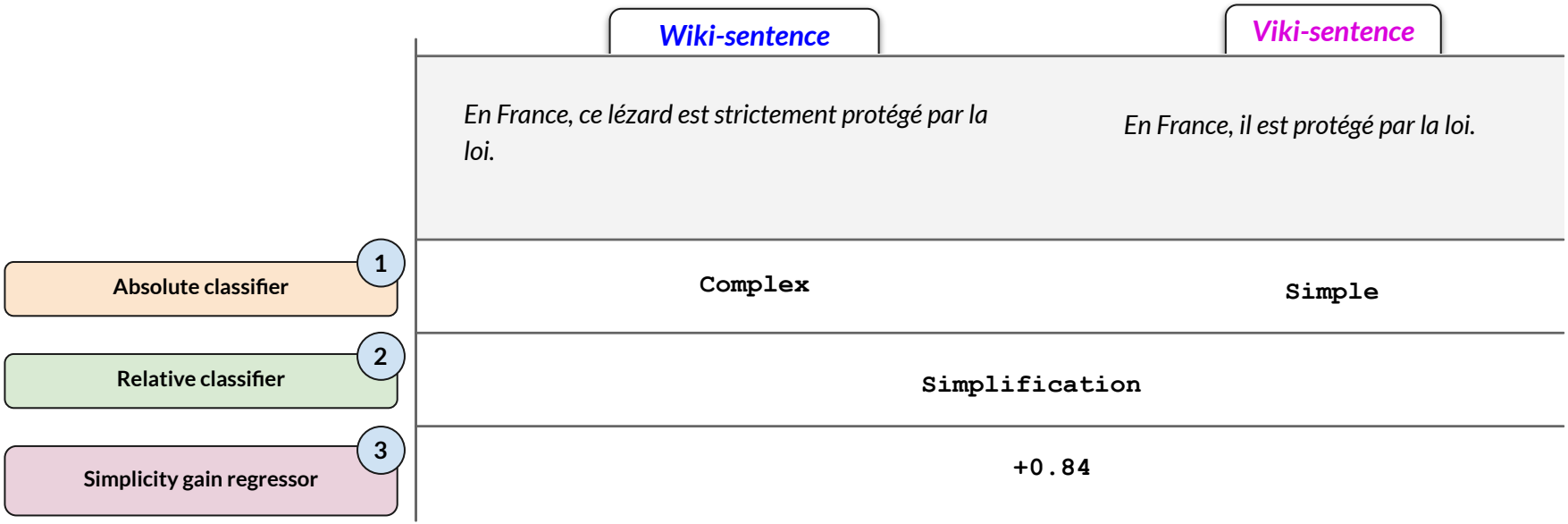
MEANS	• Increasingly fine-grained simplicity assessment approach. • Fine-tuning with WikiLarge-FR, by the use of FlauBERT (*small*, *base*, *large*) [Le *et al.*, 2020].



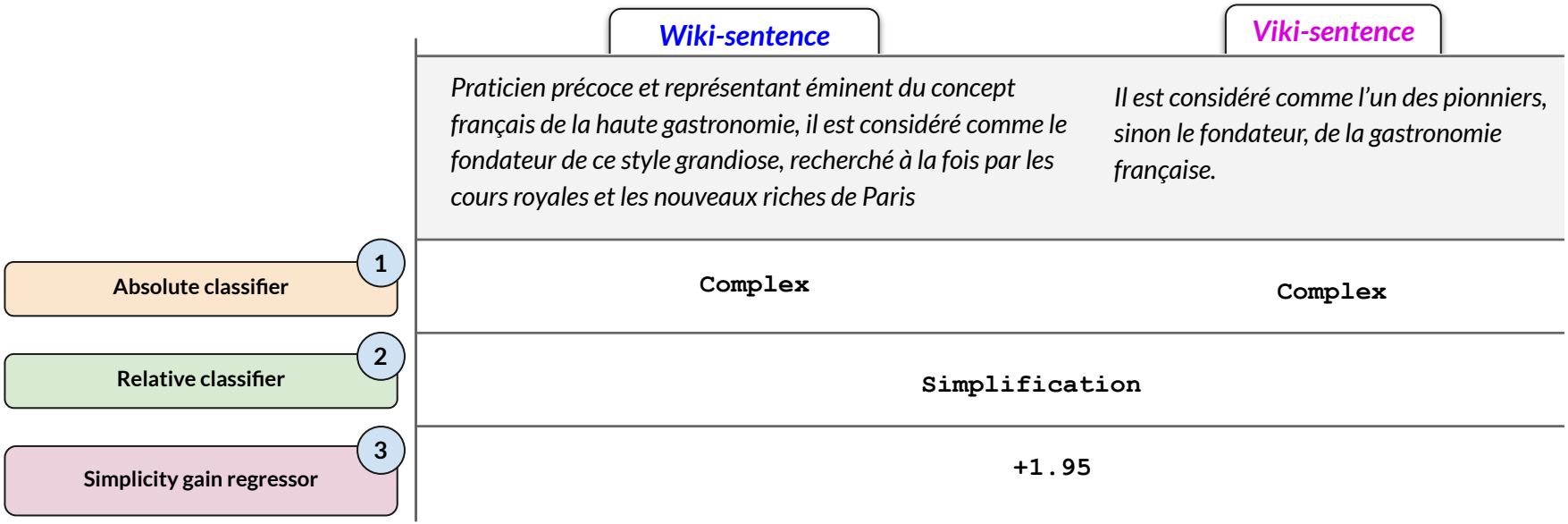
Simplicity gain regression		
Transformer model	*untuned*	*tuned*
`flaubert-small`	1.89	**0.39**
`flaubert-base`	1.18	**0.35**
`flaubert-large`	4.59	**0.23**

MSE scores obtained by the regressor on the test set.

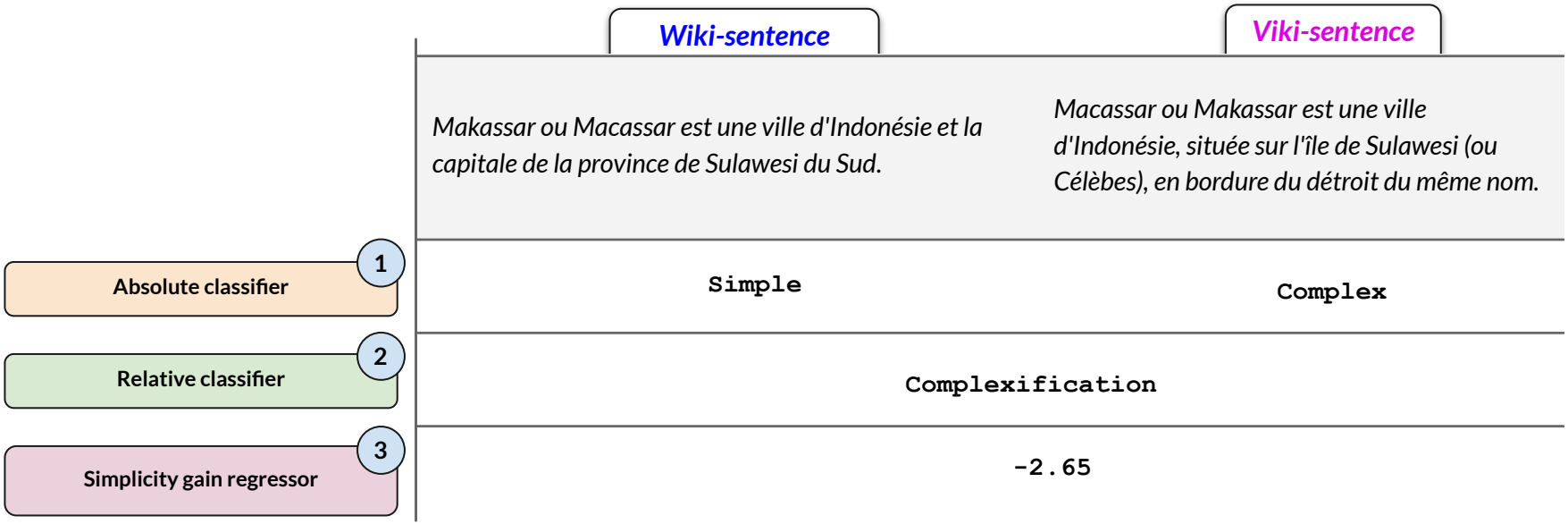
Implementing trained SCA models on Wikipedia-Vikidia data



Implementing trained SCA models on Wikipedia-Vikidia data

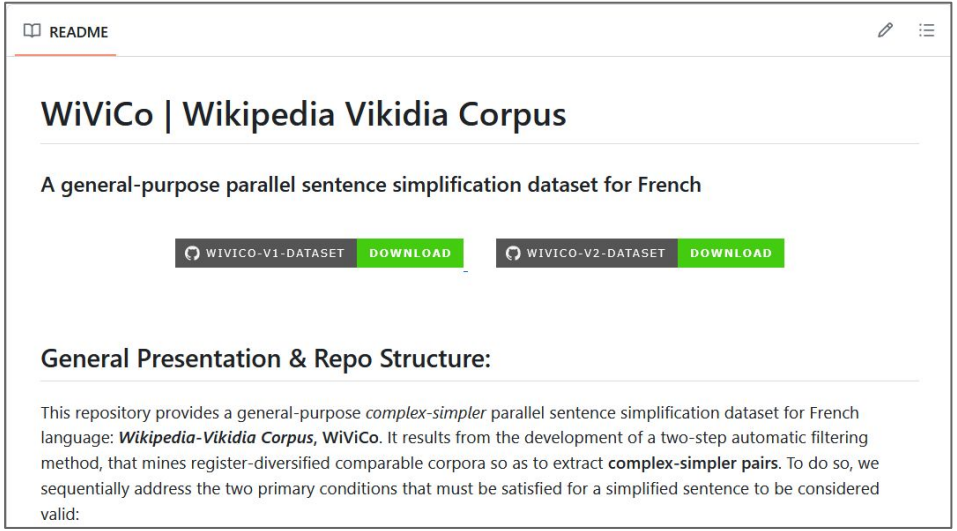


Implementing trained SCA models on Wikipedia-Vikidia data



Wikipedia-Vikidia Corpus (WiViCo)

After the implementation of the triad of SCA models…



https://github.com/lormaechea/wivico

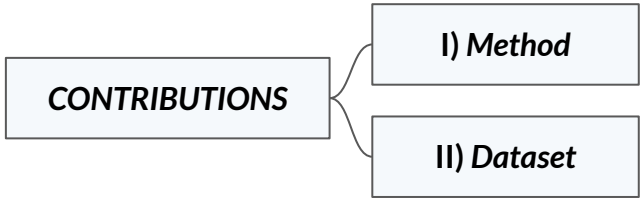
- General-purpose **parallel sentence simplification** dataset for **French language**.
- Current version (v.2): **~45*k* parallel complex-simple sentences.**
- Including **complex-simple standard** examples and:
  - `Complex-Complex → Simplification`
  - `Simple-Simple → Simplification`

4. Conclusions and further work

CONTRIBUTIONS

I) *Method*

- Introduction of a finer-grained approach for SCA.
- Suitable for ATS → Inherently a relative operation.



- Introduction of a finer-grained approach for SCA.
- Suitable for ATS → Inherently a relative operation.

- General-purpose parallel sentence simplification dataset → WiViCo

CONTRIBUTIONS

I) *Method*

- Introduction of a finer-grained approach for SCA.
- Suitable for ATS → Inherently a relative operation.

II) *Dataset*

- General-purpose parallel sentence simplification dataset → WiViCo

---

LIMITATIONS

- Use of Google Translate to obtain WikiLarge-Fr.
- Need to manually assess the correctness of the produced translations.

CONTRIBUTIONS

I) *Method*

- Introduction of a finer-grained approach for SCA.
- Suitable for ATS → Inherently a relative operation.

II) *Dataset*

- General-purpose parallel sentence simplification dataset → WiViCo

---

LIMITATIONS

- Use of Google Translate to obtain WikiLarge-Fr.
- Need to manually assess the correctness of the produced translations.

FURTHER WORK

I) ATS *Ranking*

- Embed SCA models into a larger pipeline for ATS.
- Rank candidate simplified sentences: most simplified with best meaning preservation.

CONTRIBUTIONS

I) *Method*

- Introduction of a finer-grained approach for SCA.
- Suitable for ATS → Inherently a relative operation.

II) *Dataset*

- General-purpose parallel sentence simplification dataset → WiViCo

---

LIMITATIONS

- Use of Google Translate to obtain WikiLarge-Fr.
- Need to manually assess the correctness of the produced translations.

FURTHER WORK

I) ATS *Ranking*

- Embed SCA models into a larger pipeline for ATS.
- Rank candidate simplified sentences: most simplified with best meaning preservation.

II) *Interpretability*

- Better interpretability of the simplicity gain score.
- Examine the correlation: linguistic features and human judgments.

UNIVERSITÉ DE GENÈVE

UGA Université Grenoble Alpes

Thank you for your attention!

Lucía Ormaechea

Ph.D. Candidate

Lucia.OrmaecheaGrijalba@unige.ch

https://luciaormaechea.com/

References (1/2)

C. Horn, C. Manduca and D. Kauchak (2014)

Learning a Lexical Simplifier using Wikipedia

Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, 458-463.

S. Stajner (2021)

Automatic Text Simplification for Social Good: Progress and Challenges

Findings of the Association for Computational Linguistics, 2637-2652.

X. Zhang and M. Lapata (2017)

Sentence Simplification with Deep Reinforcement Learning

Proceedings of the Conference on Empirical Methods in Natural Language Processing, 584-594.

D. Brunato, F. Dell'Orletta and G. Venturi (2022)

Linguistically-Based Comparison of Different Approaches to Building Corpora for Text Simplification: A Case Study on Italian

Frontiers in Psychology, 13.

References (2/2)

H. Le, L. Vial, J. Frej, V. Segonne, M. Coavoux, B. Lecouteux, A. Allauzen, B. Crabbé, L. Besacier and D. Schwab (2020)

FlauBERT: Unsupervised Language Model Pre-training for French

Proceedings of the 12th Language Resources and Evaluation Conference, 2479-2490.

L. Ormaechea and N. Tsourakis (2023)

Extracting Sentence Simplification Pairs from French Comparable Corpora Using a Two-Step Filtering Method

Proceedings of the 8th Swiss Text Analytics Conference (SwissText).

L. Tanguy and N. Tulechki (2009)

Sentence Complexity in French: a Corpus-Based Approach

Intelligent Information Systems (IIS), 131-145.