Exploring morphology-aware tokenization: A case study on Spanish language modeling

Alba Táboas García[1] Piotr Przybyła[1,2] Leo Wanner[3]

[1] Universitat Pompeu Fabra
[2] Institute of Computer Science, Polish Academy of Sciences
[3] Barcelona Supercomputing Center & ICREA

Overview

1. Introduction
2. Morphology-Aware Tokenization
3. Impact on Language Modeling and Downstream Applications
4. Conclusions and Future Work

Introduction

Tokenization Matters

Tokenization is a **crucial** part of neural language models:

Tokenization Matters

Tokenization is a **crucial** part of neural language models:

- Shapes the **vocabulary**:
  Affects model size, efficiency, and coverage

Tokenization Matters

Tokenization is a **crucial** part of neural language models:

- Shapes the **vocabulary**:
  Affects model size, efficiency, and coverage
- Influences how **morphology** and **meaning** are represented in embeddings

Tokenization Matters

Tokenization is a **crucial** part of neural language models:

- Shapes the **vocabulary**:
  Affects model size, efficiency, and coverage
- Influences how **morphology** and **meaning** are represented in embeddings
- Has been shown to **affect performance** on downstream tasks

Tokenization Matters

Tokenization is a **crucial** part of neural language models:

- Shapes the **vocabulary**:
  Affects model size, efficiency, and coverage
- Influences how **morphology** and **meaning** are represented in embeddings
- Has been shown to **affect performance** on downstream tasks

Early models:	Full words as tokens WORD2VEC, GLOVE
Current standard:	Purely statistical subword tokenizers BPE, WordPiece, SentencePiece, Unigram

BPE

Byte-Pair Encoding is **trained** to efficiently represent words in a given corpus.

1. Convert input text to a sequence of bytes.
2. Initiate vocabulary V, adding every possible byte as a token,
3. Repeat until $|V|$ reaches desired size:
   - 3.1 find the most frequente pair of subsequent tokens $\langle a, b \rangle$,
   - 3.2 add new token ab to the vocabulary,
   - 3.3 replace all occurences of $\langle a, b \rangle$ with ab.

$\rightarrow$ Common words will have individual tokens, less common are split into sub-words. Everything can be represented – as individual bytes if need be.

What Do Subwords Look Like?

superbizarre	**electroneutral**	**undeniability**
direction	**unidirectional**	**bidirectional**

What Do Subwords Look Like?

super - bizarre

electro - neutral

un - deni - abil - ity

direct - ion

uni - direct - ion - al

bi - direct - ion - al

What Do Subwords Look Like?

superb - izarre	electron - eu - tral	unden - iability
direction	un - idi - re - ction - al	bid - ire - ction - al

What Do Subwords Look Like? (Polish)

przyszedłem

przy - szedł - em

pr - z - ys - zed - łem

Morphological Misalignment

Problem: Purely statistical tokenization ignores morphology

Subwords often cut across meaningful morphemes.

Morphological Misalignment

Problem: Purely statistical tokenization ignores morphology

Subwords often cut across meaningful morphemes.

Consequences:

- Poor **generalization** to unseen forms
- Trouble with derivation & inflection $\Rightarrow$
  $\Rightarrow$ Trouble with **semantics** and **morpho-syntax**
- Reduced **interpretability**

Morphological Misalignment

Problem: Purely statistical tokenization ignores morphology

Subwords often cut across meaningful morphemes.

Consequences:

- Poor **generalization** to unseen forms
- Trouble with derivation & inflection $\Rightarrow$
  $\Rightarrow$ Trouble with **semantics** and **morpho-syntax**
- Reduced **interpretability**

Even more problematic for morphologically rich languages

Example: verb conjugation of indicative in Spanish

PARTICIPLES

Present: hablando

Past: hablado

☐ Include *vos*

☑ Include *vosotros*

Indicative of "hablar"

	Present	Preterite	Imperfect	Conditional	Future
yo	hablo	hablé	hablaba	hablaría	hablaré
tú	hablas	hablaste	hablabas	hablarías	hablarás
él/ella/Ud.	habla	habló	hablaba	hablaría	hablará
nosotros	hablamos	hablamos	hablábamos	hablaríamos	hablaremos
vosotros	habláis	hablasteis	hablabais	hablaríais	hablaréis
ellos/ellas/Uds.	hablan	hablaron	hablaban	hablarían	hablarán

Prior Work: Mixed Evidence

- **Positive results:**
  Morph-aware tokenization has been shown to improve downstream performance in English, Dutch, Turkish, Korean, Portuguese.

Prior Work: Mixed Evidence

- **Positive results:**
  Morph-aware tokenization has been shown to improve downstream performance in English, Dutch, Turkish, Korean, Portuguese.
- **Conflicting findings:**
  Some studies report only marginal gains or even advantages for purely statistical approaches.

Prior Work: Mixed Evidence

- **Positive results:**
  Morph-aware tokenization has been shown to improve downstream performance in English, Dutch, Turkish, Korean, Portuguese.
- **Conflicting findings:**
  Some studies report only marginal gains or even advantages for purely statistical approaches.

  However, some of them:
  - Focus only on narrow linguistic phenomena
  - Extract conclusions by comparing typologically different languages, instead of statistical vs. morphological tokenization in a fixed language.

Our Contribution

General Research Question: Can morphology-aware tokenization improve LM performance?

Our Contribution

General Research Question: Can morphology-aware tokenization improve LM performance?

- Focus on **Spanish** (fusional morphology, underexplored)

Our Contribution

General Research Question: Can morphology-aware tokenization improve LM performance?

- Focus on **Spanish** (fusional morphology, underexplored)
- Build a **morphology-aware tokenizer**:
  - Using a state-of-the-art morphological segmentation model
  - Integrating it into standard BPE training

Our Contribution

General Research Question: Can morphology-aware tokenization improve LM performance?

- Focus on **Spanish** (fusional morphology, underexplored)
- Build a **morphology-aware tokenizer**:
  - Using a state-of-the-art morphological segmentation model
  - Integrating it into standard BPE training
- Use said tokenizer and **pre-train a LM**
  - Trying different masking strategies

Our Contribution

General Research Question: Can morphology-aware tokenization improve LM performance?

- Focus on **Spanish** (fusional morphology, underexplored)
- Build a **morphology-aware tokenizer**:
  - Using a state-of-the-art morphological segmentation model
  - Integrating it into standard BPE training
- Use said tokenizer and **pre-train a LM**
  - Trying different masking strategies
- Perform comprehensive **evaluation**:
  - Intrinsic (perplexity, word prediction, linguistic probing)
  - Extrinsic (NLI, paraphrase detection, STS)

Morphology-Aware Tokenization

Two-Stage Approach

Stage 1: Morphological segmentation

- Train segmentation models for Spanish (MorphAGram)
- Select the best model (highest morphological quality)

Stage 2: Morph-aware BPE tokenization

- Use selected model to segment a dataset
- Train BPE tokenizer on the segmented dataset

Two-Stage Approach

Stage 1: Morphological segmentation

- Train segmentation models for Spanish (MorphAGram)
- Select the best model (highest morphological quality)

Stage 2: Morph-aware BPE tokenization

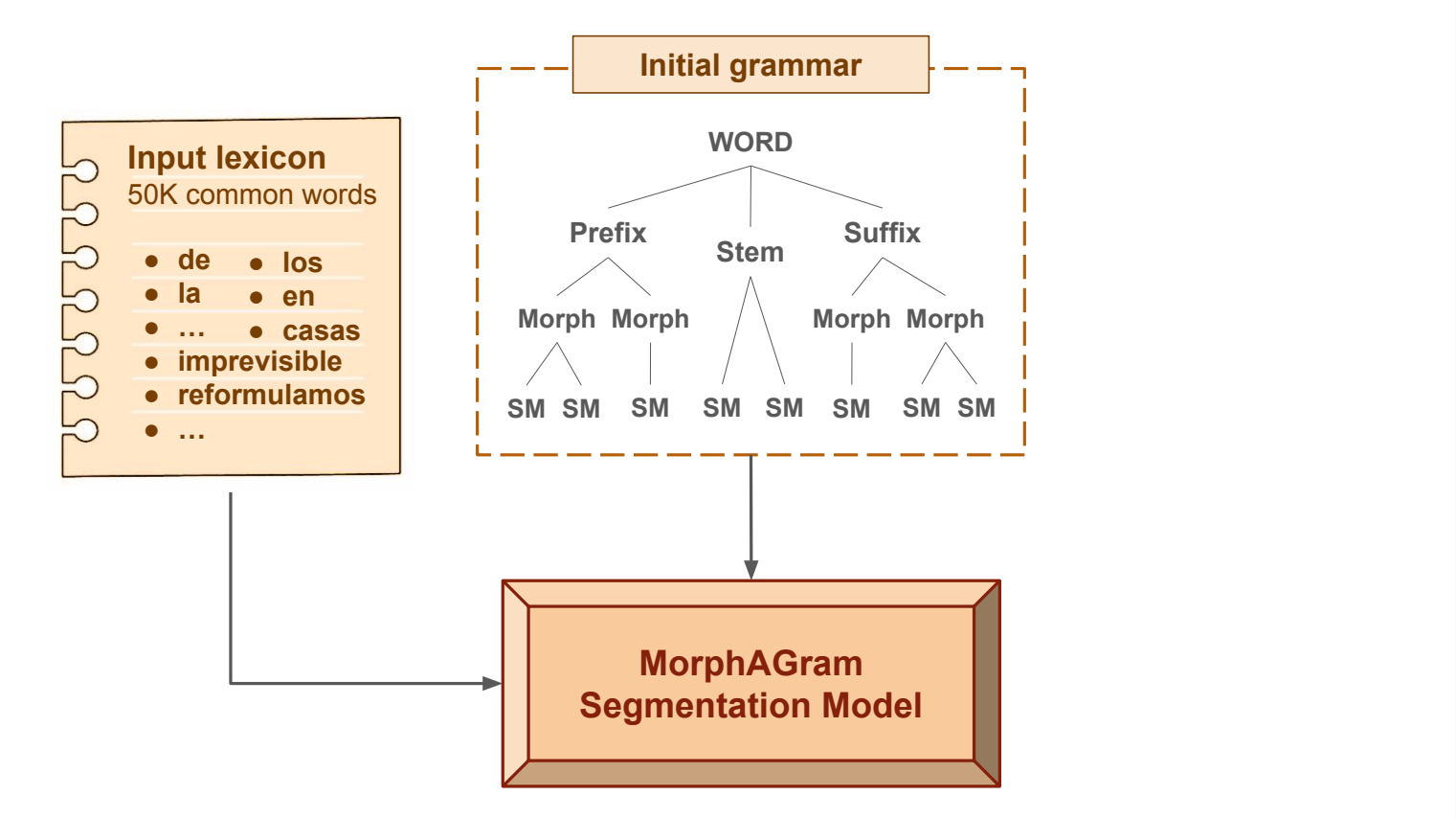
- Use selected model to segment a dataset
- Train BPE tokenizer on the segmented dataset

Benefit: Strike a **balance** between linguistic **informativeness** and statistical **efficiency**.

STAGE 1: MorphAGram for Spanish

Unsupervised setup:

- 50k-word lexicon
- Initial grammar



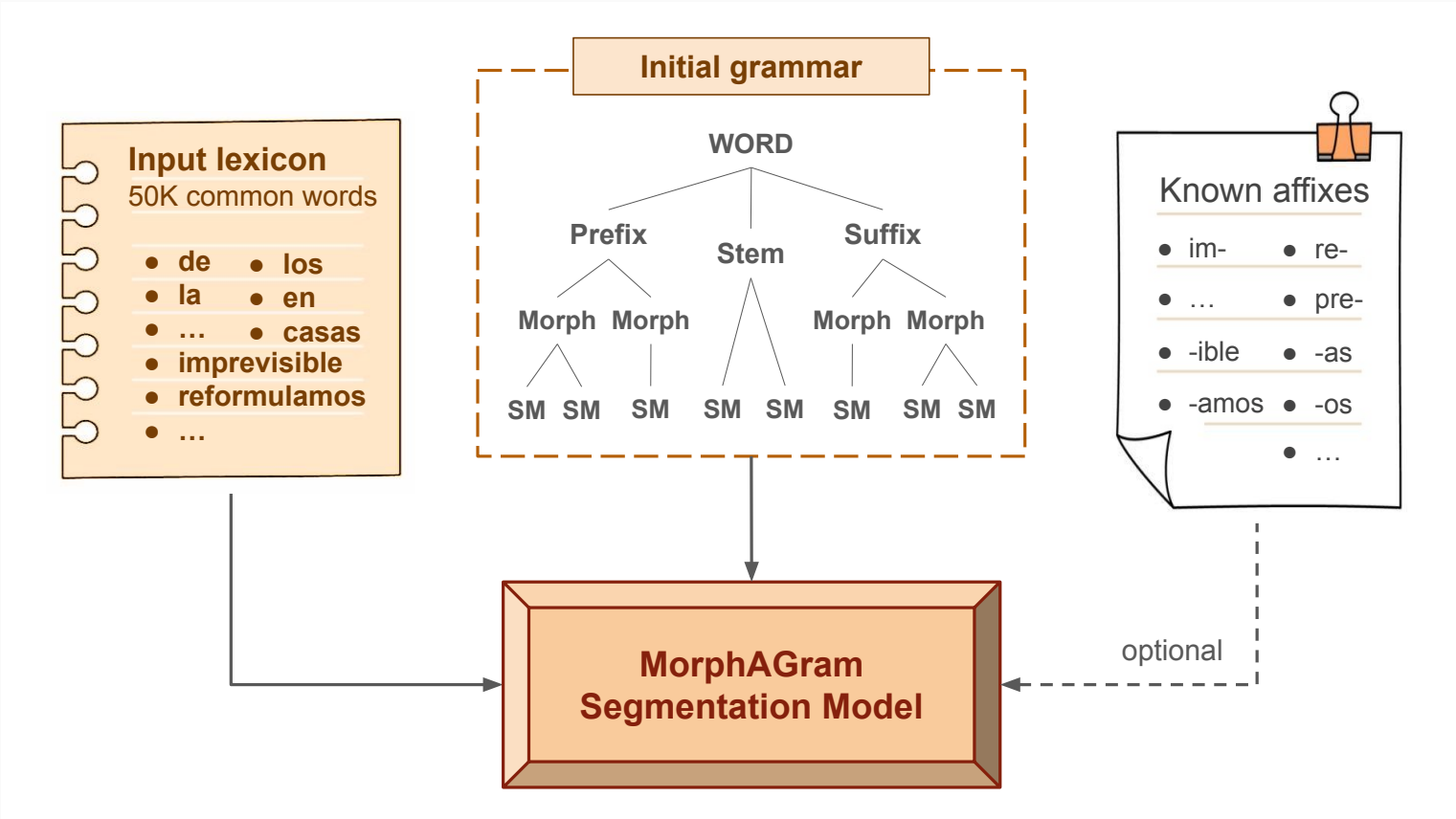
STAGE 1: MorphAGram for Spanish

Unsupervised setup:

- 50K-word lexicon
- Initial grammar

Semi-supervised setup:

- List of known inflectional and derivational affixes



STAGE 1: Segmentation Models and Examples

Compared Models

- Baseline: Morfessor 2.0 (via Polyglot library)
- MorphAGram: - Unsupervised - Semi-supervised

Segmentation Samples vs. Gold Reference

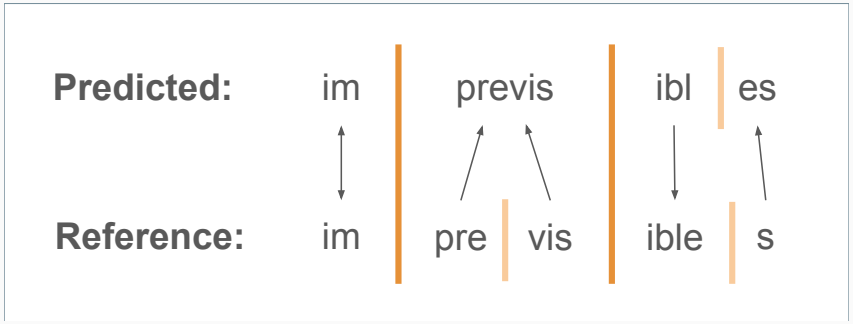
Stem morphs appear in bold

Morfessor 2.0	MorphAGram Unsupervised	MorphAGram Semi-supervised	Reference	Translation
impre-visible	impre-vis-**ible**	im-pre-**vis**-ible	im-pre-**vis**-ible	*unpredictable*
inter-nacional	inter-n-**acional**	**intern**-a-cion-al	inter-**nacion**-al	*international*
des-cuida-da-mente	des-cuid-**adamente**	des-**cuid**-ada-mente	des-**cuid**-ada-mente	*carelessly*
rápida-mente	rá-pid-**amente**	**ráp**-ida-mente	**rápid**-a-mente	*quickly*
transforma-ción	trans-form-**ación**	trans-**form**-ación	trans-**form**-ación	*transformation*
re-conocimiento	re-conoc-**imiento**	**reconoc**-imiento	re-**conoc**-imiento	*acknowledgment*
re-formula-mos	re-formul-**amos**	re-**formul**-amos	re-**formul**-amos	*(we) reformulate*
configura-s-te	con-figur-**aste**	**configuraste**	con-**figur**-aste	*(you) configurated*

STAGE 1: Segmentation Model Evaluation

Metrics

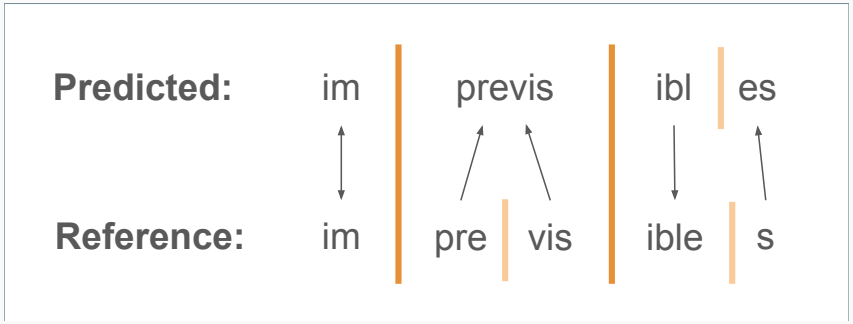
- Boundary Precision and Recall
- EMMA-2 morph-level matching



STAGE 1: Segmentation Model Evaluation

Metrics

- Boundary Precision and Recall
- EMMA-2 morph-level matching



Gold references

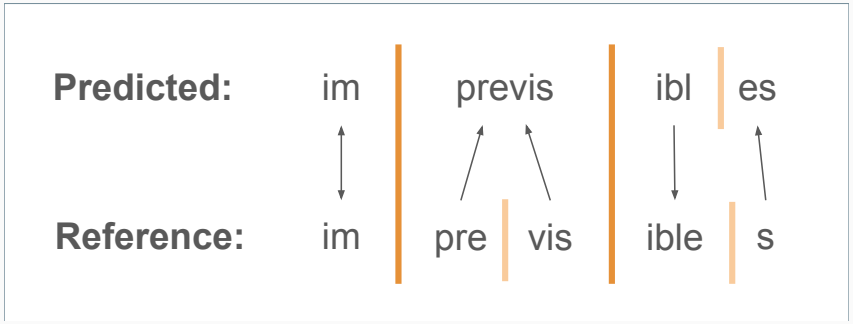
Annotated by hand

- 1,200 unseen words
- 5,400 words from texts

STAGE 1: Segmentation Model Evaluation

Metrics

- Boundary Precision and Recall
- EMMA-2 morph-level matching



Gold references

Annotated by hand

- 1,200 unseen words
- 5,400 words from texts

Segmentation Model	Words		Texts	
	BPR	EMMA2	BPR	EMMA2
Morfessor	0.29	0.72	0.64	0.74
MorphAGram				
Unsupervised	0.68	0.78	0.84	0.84
Semisupervised	**0.77**	**0.88**	**0.89**	**0.89**

Selected model: Semisupervised MorphAGram

STAGE 2: Dataset Pre-processing

Training dataset: 10% random subset of Spanish OSCAR (16GB)

> Nuestros clientes también tienen la posibilidad de jugar a La Primitiva online en alguno de nuestros grupos, y así incrementar sus posibilidades de obtener uno de los numerosos premios en juego.

STAGE 2: Dataset Pre-processing

Training dataset: 10% random subset of Spanish OSCAR (16GB)

> Nuestros clientes también tienen la posibilidad de jugar a La Primitiva online en alguno de nuestros grupos, y así incrementar sus posibilidades de obtener uno de los numerosos premios en juego.

Pre-process the dataset with our selected MorphAGram model:

Insert special symbol $< + >$ between morphs within each word

> Nuestr<+>os client<+>es también tien<+>en la pos<+>ibilidad de jug<+>ar a La Prim<+>itiva online en algun<+>o de nuestr<+>os grup<+>os, y así in<+>crement<+>ar sus pos<+>ibilidades de obten<+>er un<+>o de l<+>os numer<+>osos premios en jueg<+>o.

STAGE 2: Dataset Pre-processing

Training dataset: 10% random subset of Spanish OSCAR (16GB)

> Nuestros clientes también tienen la posibilidad de jugar a La Primitiva online en alguno de nuestros grupos, y así incrementar sus posibilidades de obtener uno de los numerosos premios en juego.

Pre-process the dataset with our selected MorphAGram model:

Insert special symbol $< + >$ between morphs within each word

> Nuestr<+>os client<+>es también tien<+>en la pos<+>ibilidad de jug<+>ar a La Prim<+>itiva online en algun<+>o de nuestr<+>os grup<+>os, y así in<+>crement<+>ar sus pos<+>ibilidades de obten<+>er un<+>o de l<+>os numer<+>osos premios en jueg<+>o.

Main idea:

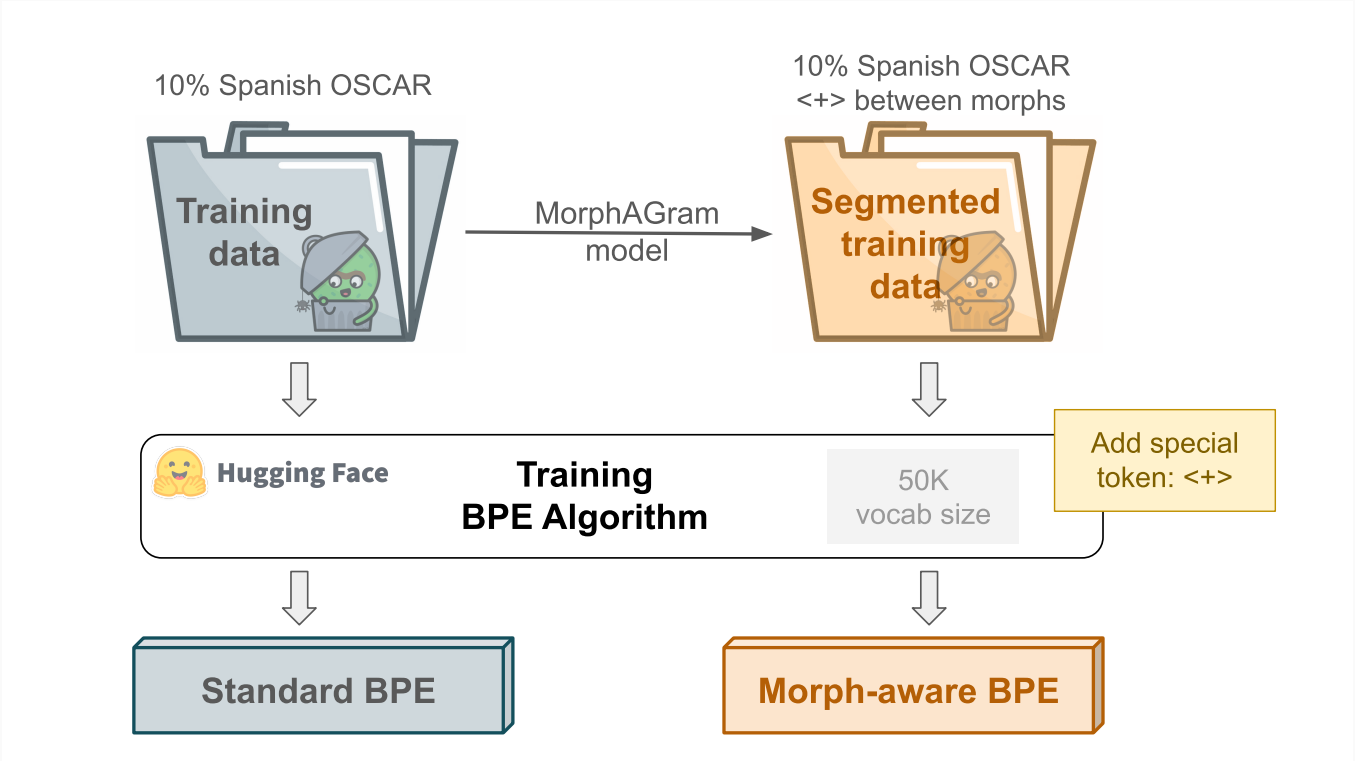
Add $< + >$ as a special token

Force BPE to consider **morph boundaries**

STAGE 2: Tokenizer Training

We train two **BPE** tokenizers (50k tokens each):

- **Standard**
  Raw dataset
- **Morphology-aware**
  Pre-segmented dataset & special token $< + >$



STAGE 2: Tokenization Examples

Tokenization Examples: Standard vs. Morph-aware BPE

Blanks mark the separation between tokens, '␣' stands for the beginning of a word
Differing tokenizations appear highlighted in blue (standard) or yellow (morph-aware)

Standard BPE
La ␣**heroica** ␣ciudad ␣**dormía** ␣la ␣**siesta** . ␣El ␣**viento** ␣Sur , ␣**caliente** ␣y ␣perez oso , ␣**empu jaba** ␣las ␣**nubes** ␣**blanque cinas** ␣que ␣se ␣**ras gaban** ␣al ␣**correr** ␣hacia ␣el ␣Norte .
Morphology-aware BPE
La ␣**hero ica** ␣ciudad ␣**dorm ía** ␣la ␣**siest a** . ␣El ␣**v iento** ␣Sur , ␣**cal iente** ␣y ␣perez oso , ␣**empuj aba** ␣las ␣**nub es** ␣**blanque cin as** ␣que ␣se ␣**rasg aban** ␣al ␣**corr er** ␣hacia ␣el ␣Norte .

'The heroic city was taking a nap. The hot, lazy South wind pushed the chalky clouds, which tore as they raced North.'

STAGE 2: Tokenizer Evaluation

Evaluation metrics:

- BPR, EMMA-2
  morph alignment with the same
  gold references as before
- Subword fertility
  average #tokens per word

STAGE 2: Tokenizer Evaluation

Evaluation metrics:

- BPR, EMMA-2
  morph alignment with the same gold references as before
- Subword fertility
  average #tokens per word

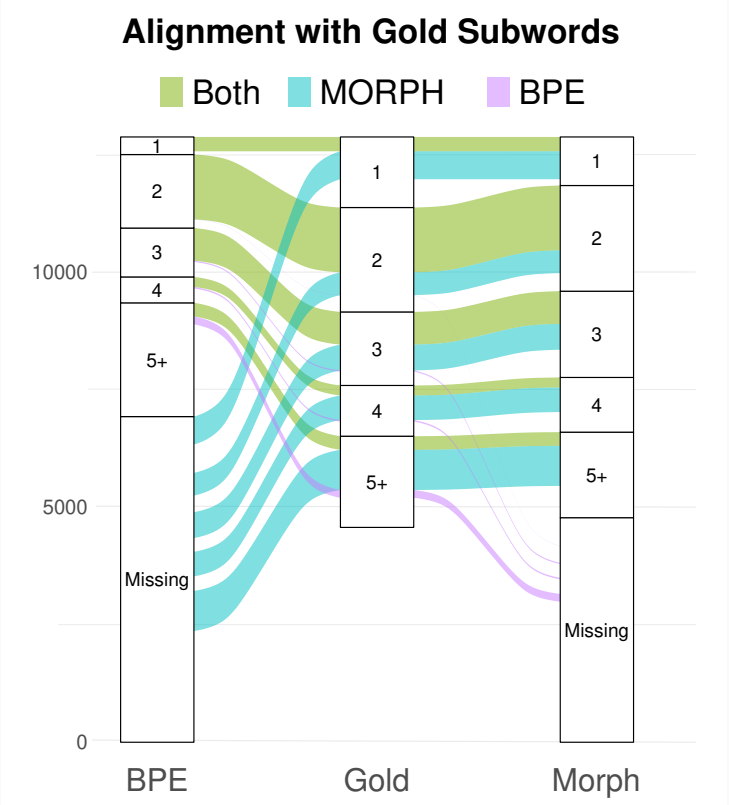
	Words		Texts		Subword
Tokenizer	BPR	EMMA2	BPR	EMMA2	**fertility**
Morph-aware	**0.67**	**0.84**	**0.83**	**0.84**	1.45
Standard	0.39	0.74	0.70	0.68	1.12

STAGE 2: Tokenizer Evaluation

Evaluation metrics:

- BPR, EMMA-2
  morph alignment with the same gold references as before
- Subword fertility
  average #tokens per word

	Words		Texts		Subword
Tokenizer	BPR	EMMA2	BPR	EMMA2	**fertility**
Morph-aware	**0.67**	**0.84**	**0.83**	**0.84**	1.45
Standard	0.39	0.74	0.70	0.68	1.12



Impact on Language Modeling and Downstream Applications

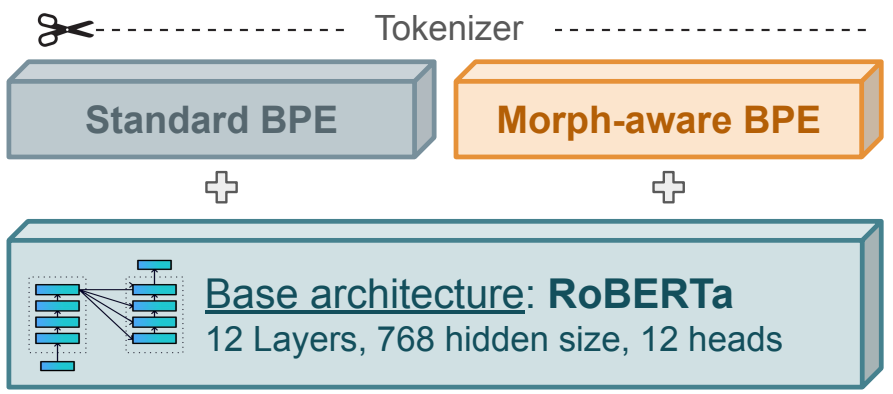
Language Model Setup and Pretraining Details

Base architecture:

- RoBERTa-base
  12L, 768H, 12H

Tokenizers

- Standard BPE
- Morphology-aware



Language Model Setup and Pretraining Details

Base architecture:

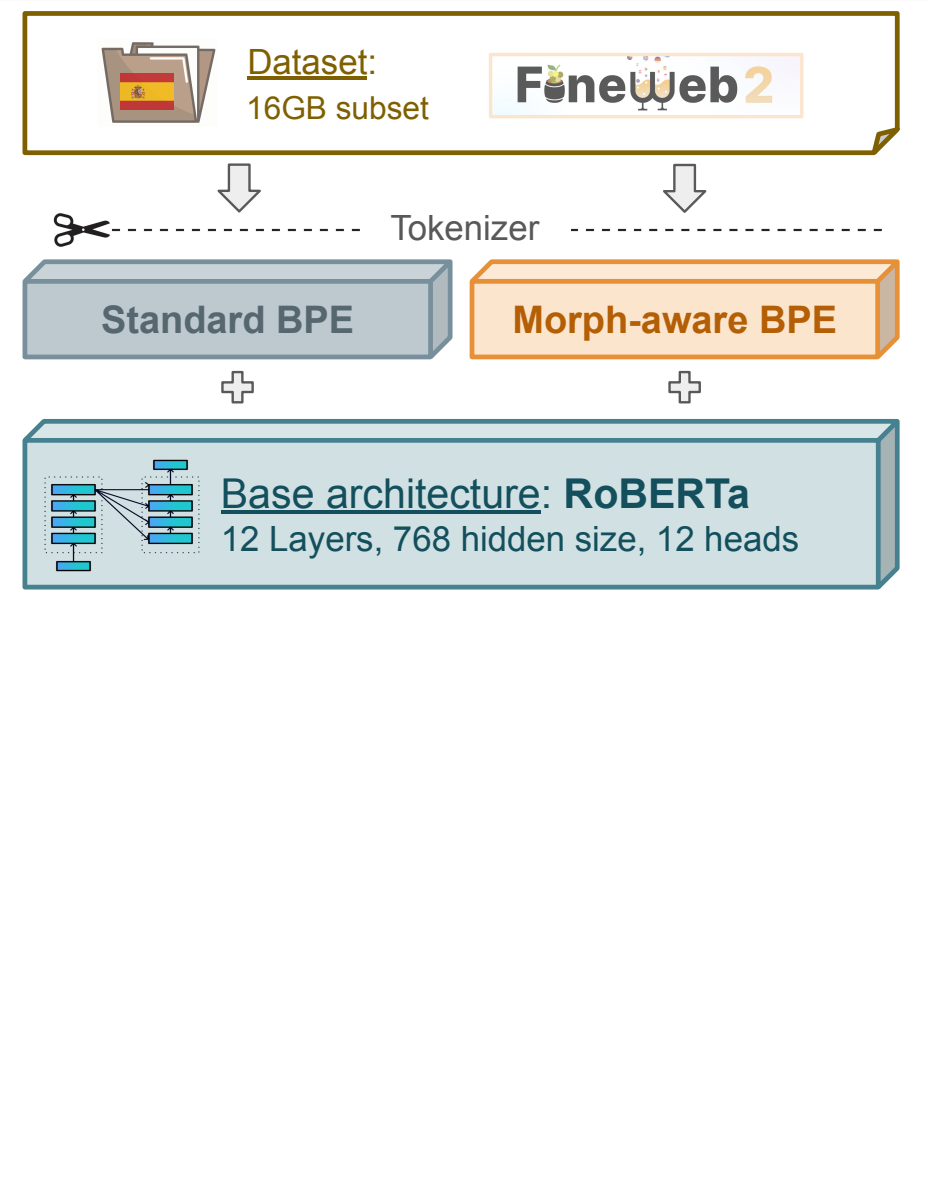
- RoBERTa-base
  12L, 768H, 12H

Tokenizers

- Standard BPE
- Morphology-aware

Pre-training details

- FineWeb-2
  (16 GB subset)



Language Model Setup and Pretraining Details

Tokenizers

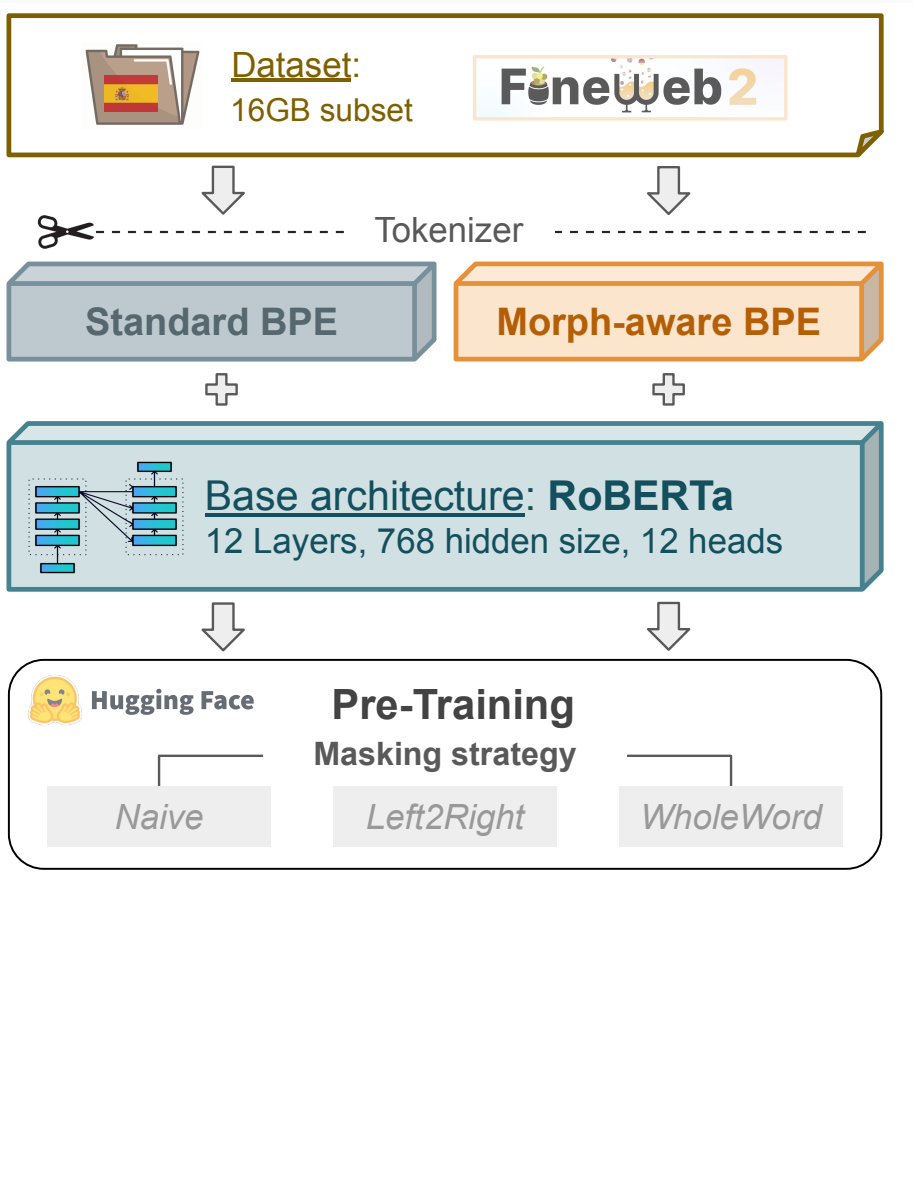
- Standard BPE
- Morphology-aware

Base architecture:

- RoBERTa-base
  12L, 768H, 12H

Pre-training details

- FineWeb-2
  (16 GB subset)
- 5 epochs
  2 × NVIDIA H100
- 3 masking strategies:
  Naive, L2R, WW



Language Model Setup and Pretraining Details

Base architecture:

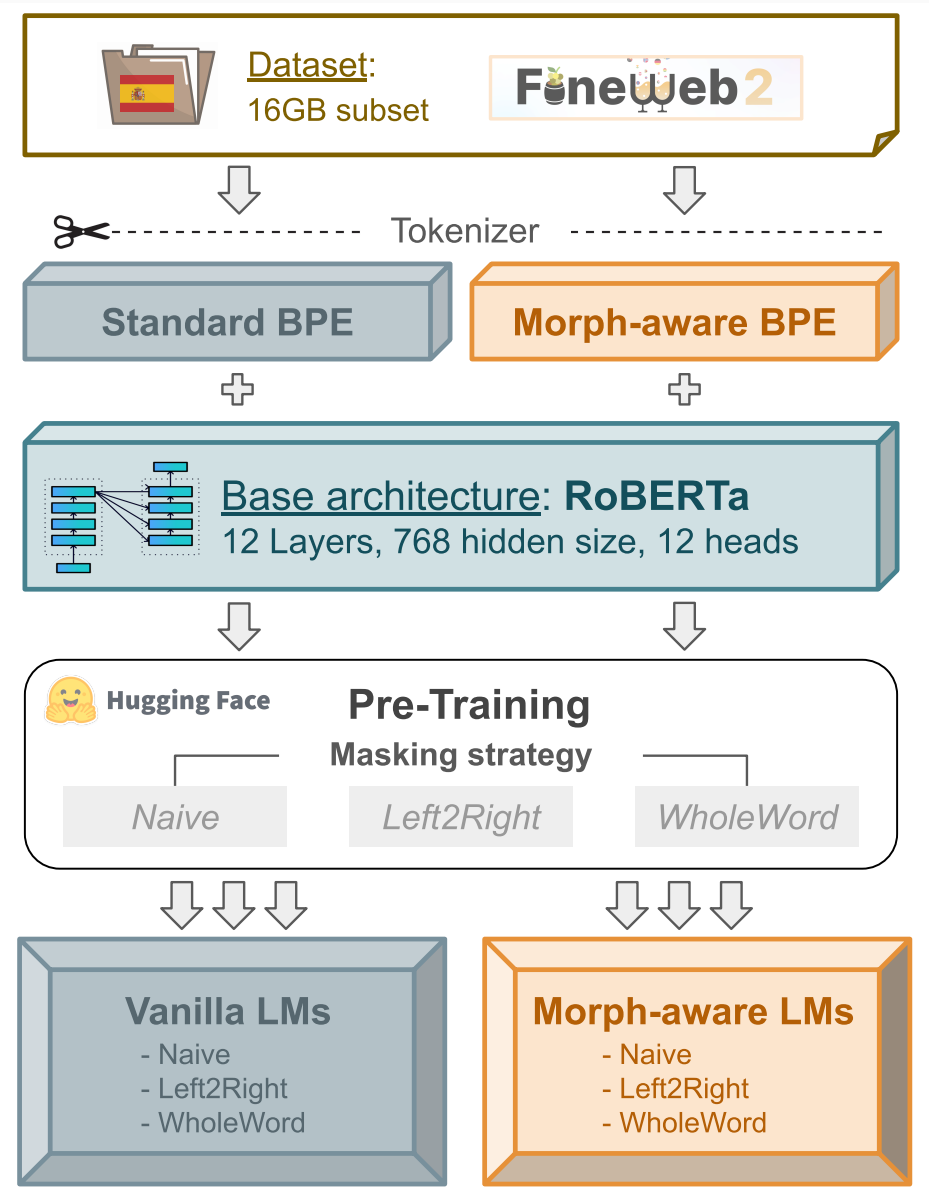
- RoBERTa-base
  12L, 768H, 12H

Tokenizers

- Standard BPE
- Morphology-aware

Pre-training details

- FineWeb-2
  (16 GB subset)
- 5 epochs
  2 × NVIDIA H100
- 3 masking strategies:
  Naive, L2R, WW



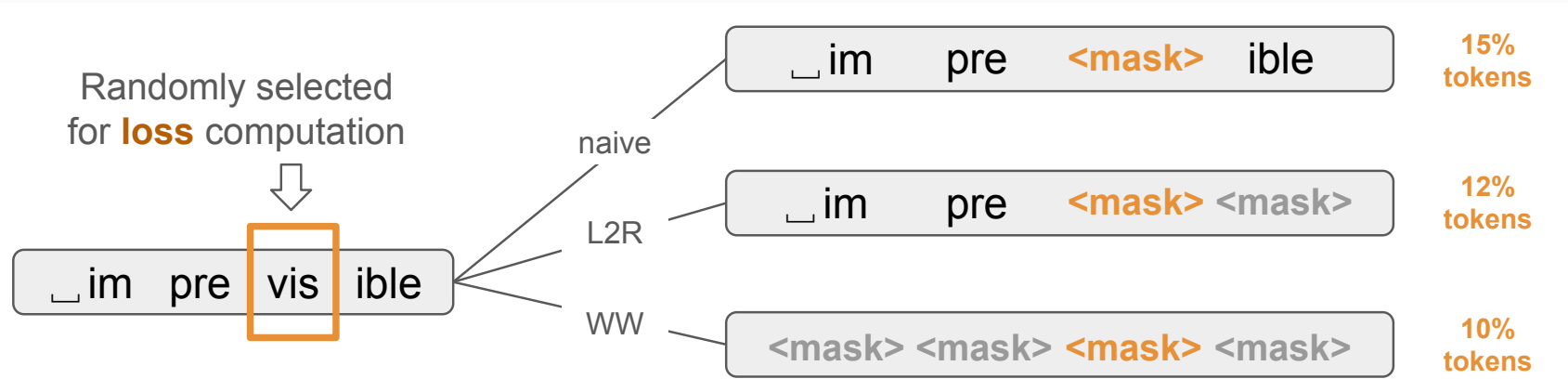
Masking Strategies

Inspired by recent literature on Masked Language Model scoring, we try three masking strategies during training:

– *Naive* – *Left-to-right* (L2R) – *Whole word* (WW)

Left-to-right and *whole word* masking are more appropriate for multi-token words, which are many when using a morph-aware tokenizer.

Masking strategies



Perplexity Measurements

We compute both **perplexity** and **bits-per-byte**
for fair comparison among models with different tokenizers

- Small dataset
  $\neq$ pre-training data
- 3 masking strategies:
  naive, L2R, WW

Perplexity Measurements

We compute both **perplexity** and **bits-per-byte**
for fair comparison among models with different tokenizers

- Small dataset
  $\neq$ pre-training data
- 3 masking strategies:
  naive, L2R, WW

		Perplexity		
Tokenizer	**Masking**	Naive	L2R	WW
Standard BPE	Naive	**10.11**	**14.46**	**20.51**
Standard BPE	L2R	16.88	15.82	21.54
Standard BPE	WW	27.07	24.76	20.61
Morph-aware	L2R	**7.61**	**7.76**	20.71
Morph-aware	WW	23.41	25.64	**20.03**

		Bits-per-byte		
Tokenizer	**Masking**	Naive	L2R	WW
Standard BPE	Naive	**0.657**	**0.759**	**0.859**
Standard BPE	L2R	0.803	0.784	0.872
Standard BPE	WW	0.938	0.912	0.860
Morph-aware	L2R	**0.781**	**0.789**	1.168
Morph-aware	WW	1.214	1.249	**1.155**

Perplexity Measurements

We compute both **perplexity** and **bits-per-byte**
for fair comparison among models with different tokenizers

- Small dataset
  $\neq$ pre-training data
- 3 masking strategies:
  naive, L2R, WW

		Perplexity		
Tokenizer	**Masking**	Naive	L2R	WW
Standard BPE	Naive	**10.11**	**14.46**	**20.51**
Standard BPE	L2R	16.88	15.82	21.54
Standard BPE	WW	27.07	24.76	20.61
Morph-aware	L2R	**7.61**	**7.76**	20.71
Morph-aware	WW	23.41	25.64	**20.03**

		Bits-per-byte		
Tokenizer	**Masking**	Naive	L2R	WW
Standard BPE	Naive	**0.657**	**0.759**	**0.859**
Standard BPE	L2R	0.803	0.784	0.872
Standard BPE	WW	0.938	0.912	0.860
Morph-aware	L2R	**0.781**	**0.789**	1.168
Morph-aware	WW	1.214	1.249	**1.155**

Selected models:

Vanilla LM - Standard BPE tokenizer with **naive** masking

Morph-aware LM - Morph-aware tokenizer with **L2R** masking

LAMBADA word prediction

- Task: predict **final** and **random** words in a text

!! Models are bidirectional & target words can be multi-token

- Greedy prediction L2R/R2L
- Accuracy (exact match)

Un segundo después, Sibyl levantó lentamente la vista. Sus rizos rubios caían sin fuerza sobre sus hombros y una sombra embrujada oscurecía sus brillantes ojos azules. Tenía miedo.

random

La chica no mostró sorpresa ante la llegada de Nika, como si la hubiera estado esperando desde el principio.

final

"Ven y siéntate conmigo", dijo Sibyl.

Word Prediction and Linguistic Probing

LAMBADA word prediction

- Task: predict **final** and **random** words in a text

!! Models are bidirectional & target words can be multi-token

- Greedy prediction L2R/R2L
- Accuracy (exact match)

Un segundo después, Sibyl levantó lentamente la vista. Sus rizos rubios caían sin fuerza sobre sus hombros y una sombra embrujada oscurecía sus brillantes ojos azules. Tenía miedo.

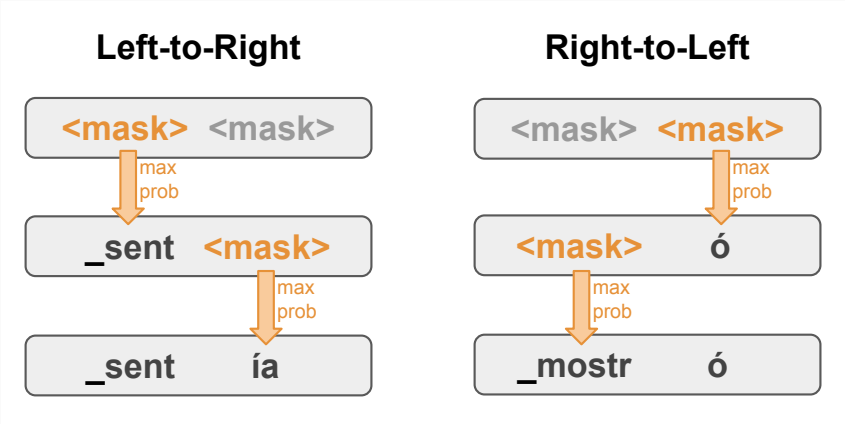
random

La chica no mostró sorpresa ante la llegada de Nika, como si la hubiera estado esperando desde el principio.

final

"Ven y siéntate conmigo", dijo Sibyl.

Greedy prediction:



We compute the tokens' joint probability for each word and select the highest among the two

$$p_{L2R}(\_sentia) = p(\_sent|BC) \cdot p(ia|BC + \_sent)$$

$$p_{R2L}(\_mostro) = p(o|BC) \cdot p(\_mostr|BC + o)$$

where BC is the bidirectional context (with the corresponding masks)

Word Prediction and Linguistic Probing

LAMBADA word prediction

- Task: predict **final** and **random** words in a text

!! Models are bidirectional & target words can be multi-token

- Greedy prediction L2R/R2L
- Accuracy (exact match)

Un segundo después, Sibyl levantó lentamente la vista. Sus rizos rubios caían sin fuerza sobre sus hombros y una sombra embrujada oscurecía sus brillantes ojos azules. Tenía miedo.

random

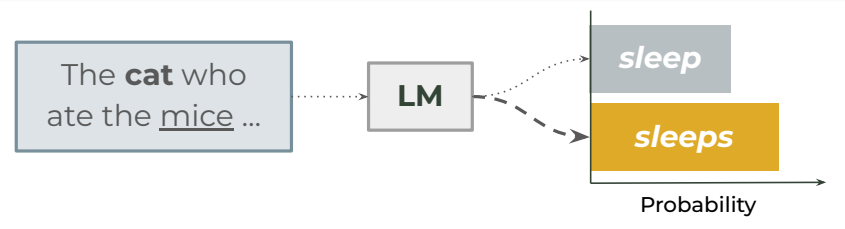
La chica no mostró sorpresa ante la llegada de Nika, como si la hubiera estado esperando desde el principio.

final

"Ven y siéntate conmigo", dijo Sibyl.

Agreement tests

- Task: assign higher probability to **correctly inflected** target
- Several types of agreement



Word Prediction and Linguistic Probing

LAMBADA word prediction

- Task: predict **final** and **random** words in a text

!! Models are bidirectional & target words can be multi-token

- Greedy prediction L2R/R2L
- Accuracy (exact match)

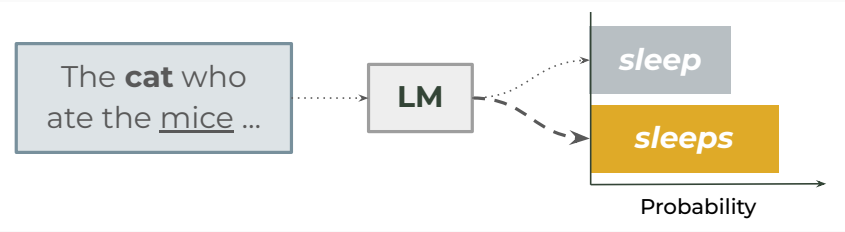
Un segundo después, Sibyl levantó lentamente la vista. Sus rizos rubios caían sin fuerza sobre sus hombros y una sombra embrujada oscurecía sus brillantes ojos azules. Tenía miedo.

La chica no mostró (random) sorpresa ante la llegada de Nika, como si la hubiera estado esperando desde el principio.

"Ven y siéntate conmigo", dijo Sibyl. (final)

Agreement tests

- Task: assign higher probability to **correctly inflected** target
- Several types of agreement



	Model	
Task	Vanilla	Morph-aware
LAMBADA word prediction		
Final word	0.338	**0.400**
Random word	0.393	**0.434**
Morpho-syntactic tests		
Agreement	0.864	**0.926**

Downstream Tasks: Finetuning Details and Results

We fine-tune both models (vanilla vs. morph-aware) on:

- tasks:
  - **Natural language inference**
    XNLI & InferES
  - **Paraphrase identification**
    PAWS-X
  - **Semantic text similarity**
    STS (SemEval)

Downstream Tasks: Finetuning Details and Results

We fine-tune both models (vanilla vs. morph-aware) on:

- tasks:
  - **Natural language inference**
    XNLI & InferES
  - **Paraphrase identification**
    PAWS-X
  - **Semantic text similarity**
    STS (SemEval)

  Hyperparameter search:
  - Batch sizes: 8, 16
  - Learning rates:
    1×10^{-5}, 3×10^{-5} , 5×10^{-5}
  - Weight decay: 0.1, 0.01

Downstream Tasks: Finetuning Details and Results

We fine-tune both models (vanilla vs. morph-aware) on:

- tasks:
  - **Natural language inference**
    XNLI & InferES
  - **Paraphrase identification**
    PAWS-X
  - **Semantic text similarity**
    STS (SemEval)

  Hyperparameter search:
  - Batch sizes: 8, 16
  - Learning rates:
    1×10^{-5}, 3×10^{-5} , 5×10^{-5}
  - Weight decay: 0.1, 0.01

Results

	Model	
Task	Vanilla	Morph-aware
Natural language inference		
XNLI (Accuracy)	0.733	**0.742**
InferES (Accuracy)	0.656	**0.666**
Paraphrase identification		
PAWS-X (F1)	0.841	**0.845**
Semantic text similarity		
STS (Combined)	0.776	**0.801**

Downstream Tasks: Finetuning Details and Results

We fine-tune both models (vanilla vs. morph-aware) on:

- tasks:
  - **Natural language inference**
    XNLI & InferES
  - **Paraphrase identification**
    PAWS-X
  - **Semantic text similarity**
    STS (SemEval)

Hyperparameter search:

- Batch sizes: 8, 16
- Learning rates:
  1×10^{-5}, 3×10^{-5} , 5×10^{-5}
- Weight decay: 0.1, 0.01

Results

	Model	
Task	Vanilla	Morph-aware
Natural language inference		
XNLI (Accuracy)	0.733	**0.742**
InferES (Accuracy)	0.656	**0.666**
Paraphrase identification		
PAWS-X (F1)	0.841	**0.845**
Semantic text similarity		
STS (Combined)	0.776	**0.801**

Morph-aware model achieves **consistent gains** across tasks

Conclusions and Future Work

Main Findings

- **Morphological segmentation**
  benefits from linguistic knowledge:
    - Our semi-supervised MorphAGram model outperforms the unspervised variant

Main Findings

- **Morphological segmentation**
  benefits from linguistic knowledge:
  - Our semi-supervised MorphAGram model outperforms the unspervised variant
- **Morph-aware tokenizer**:
  - Stronger alignment with gold segmentations
  - Slightly higher subword fertility (finer granularity)
  - More linguistically meaningful tokens

Main Findings

- **Morphological segmentation**
  benefits from linguistic knowledge:
  - Our semi-supervised MorphAGram model outperforms the unspervised variant
- **Morph-aware tokenizer**:
  - Stronger alignment with gold segmentations
  - Slightly higher subword fertility (finer granularity)
  - More linguistically meaningful tokens
- **Language Model** improvements:
  - Large gains in morphologically sensitive tasks
  - Consistent (though smaller) gains in general tasks

Concluding Remarks

Key takeaway: **Morphology-aware tokenization improves Spanish LM quality**

Concluding Remarks

Key takeaway: **Morphology-aware tokenization improves Spanish LM quality**

- Our approach is **simple but effective**: No need to modify tokenization algorithms or model architectures

Concluding Remarks

Key takeaway: **Morphology-aware tokenization improves Spanish LM quality**

- Our approach is **simple but effective**: No need to modify tokenization algorithms or model architectures
- Our approach is **generalizable** to other languages:
  - Best with an available morphological segmenter
  - But MorphAGram's framework provides a good language-agnostic alternative

Concluding Remarks

Key takeaway: **Morphology-aware tokenization improves Spanish LM quality**

- Our approach is **simple but effective**: No need to modify tokenization algorithms or model architectures
- Our approach is **generalizable** to other languages:
  - Best with an available morphological segmenter
  - But MorphAGram's framework provides a good language-agnostic alternative
- Released resources:
  - Segmentation model, reference data, lexicon, affix list
  - Tokenizers and trained models

`https://github.com/Albalbalba/morphtokenizer`

Future Work

- **Token-level tasks:**
  Explore benefits for POS tagging, parsing, SRL, and other morpho-sensitive tasks

Future Work

- **Token-level tasks:**
  Explore benefits for POS tagging, parsing, SRL, and other morpho-sensitive tasks
- **Broader language coverage:**
  Test on highly inflected and agglutinative languages

Future Work

- **Token-level tasks:**
  Explore benefits for POS tagging, parsing, SRL, and other morpho-sensitive tasks
- **Broader language coverage:**
  Test on highly inflected and agglutinative languages
- **Scaling up:**
  Evaluate impact on larger architectures and training data

Future Work

- **Token-level tasks:**
  Explore benefits for POS tagging, parsing, SRL, and other morpho-sensitive tasks
- **Broader language coverage:**
  Test on highly inflected and agglutinative languages
- **Scaling up:**
  Evaluate impact on larger architectures and training data
- **Multilingual models:**
  Adapt morphology-aware tokenization for multilingual settings

Future Work

- **Token-level tasks:**
  Explore benefits for POS tagging, parsing, SRL, and other morpho-sensitive tasks
- **Broader language coverage:**
  Test on highly inflected and agglutinative languages
- **Scaling up:**
  Evaluate impact on larger architectures and training data
- **Multilingual models:**
  Adapt morphology-aware tokenization for multilingual settings

Thinking outside the box:

Word-level tokenization is not compulsory!

Assess promising alternatives (CANINE, LCM) in morphology-sensitive tasks

Thanks for your attention!

Any questions?