Spanish is one of the top-5 spoken languages.

Not always easy to find Spanish NLP resources.

We provide a Spanish BERT model
plus training and evaluation data.

Spanish Pre-trained BERT Model and Evaluation Data

José Cañete

Gabriel Chaperón

Rodrigo Fuentes

Jou-Hui Ho

Hojing Kang

Jorge Pérez

dcc
CIENCIAS DE LA COMPUTACIÓN
UNIVERSIDAD DE CHILE

Instituto Milenio
Fundamentos
de los datos

die
DEPARTAMENTO DE INGENIERÍA ELÉCTRICA
UNIVERSIDAD DE CHILE

Outline

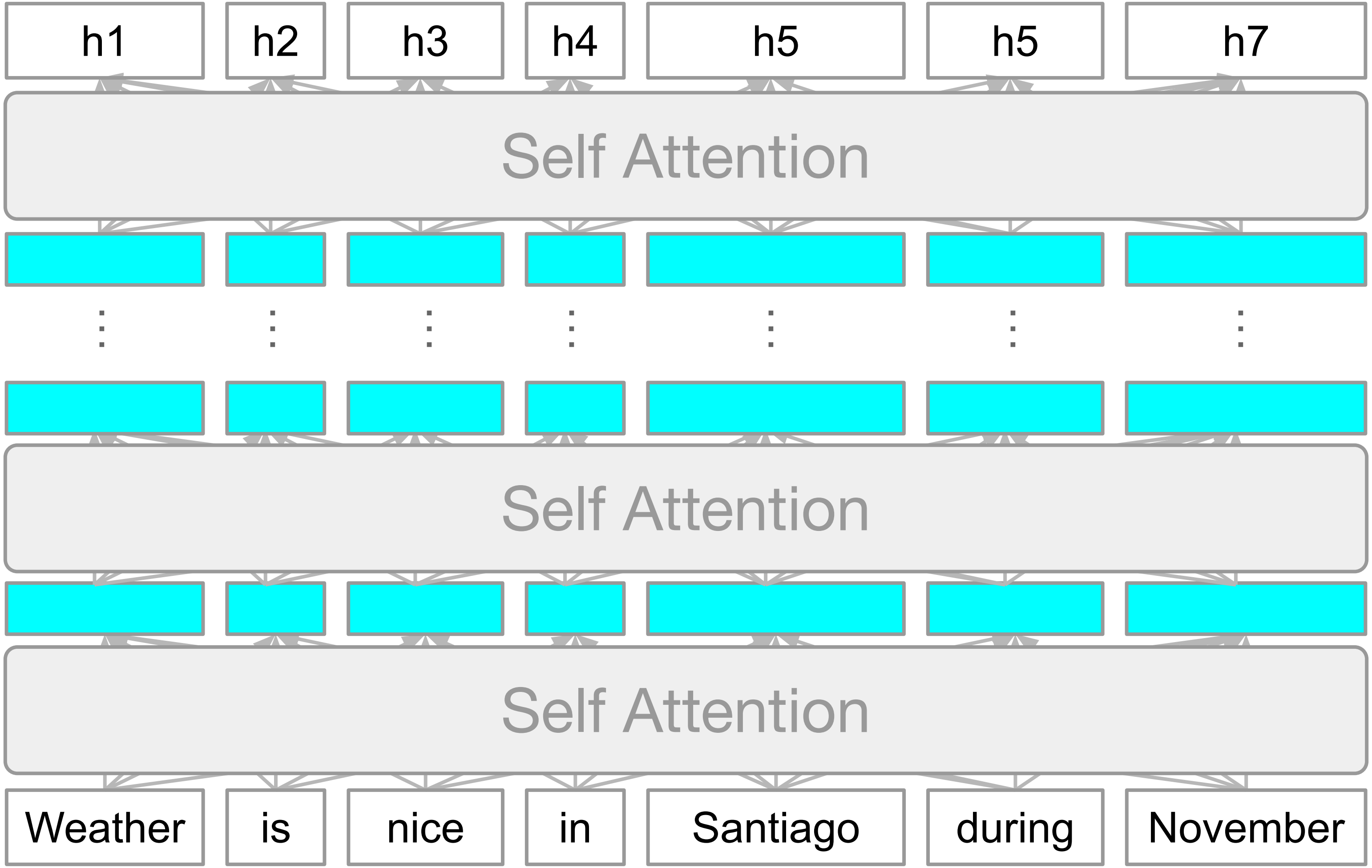
- Unsupervised pre-training
- Transformers and BERT
- Spanish BERT and Corpus
- Results: Spanish vs Multilingual BERT

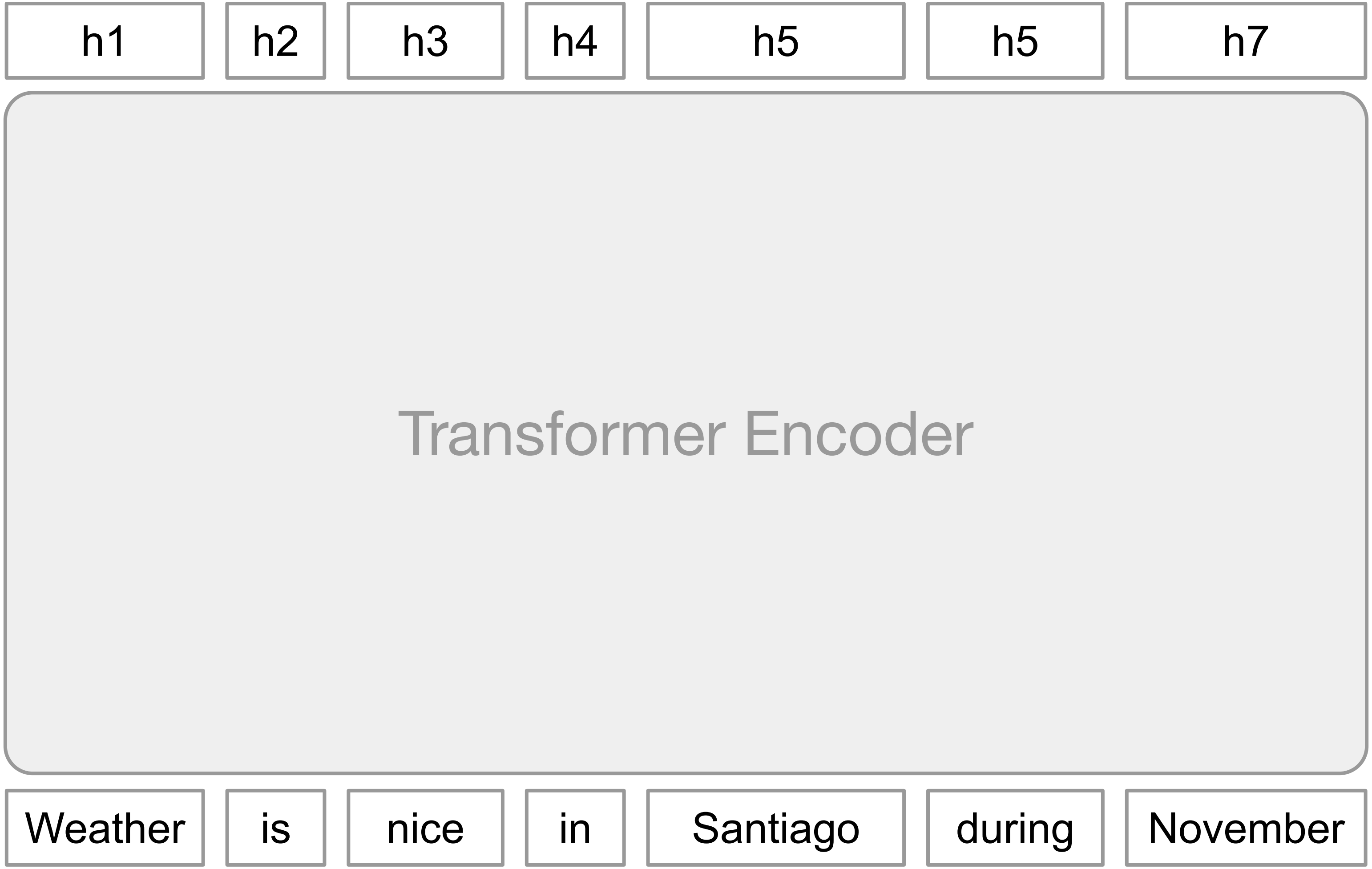
Unsupervised Pre-training

- Pick an architecture and tons of text
- Train the architecture with a general LM task (unsupervised)
- Use the weights as a starting point for training with a supervised task

Transformer

[Vaswani et al. 2017]


h1
h2
h3
h4
h5
h5
h7
Self Attention
Self Attention
Self Attention
Weather
is
nice
in
Santiago
during
November



h1 h2 h3 h4 h5 h5 h7

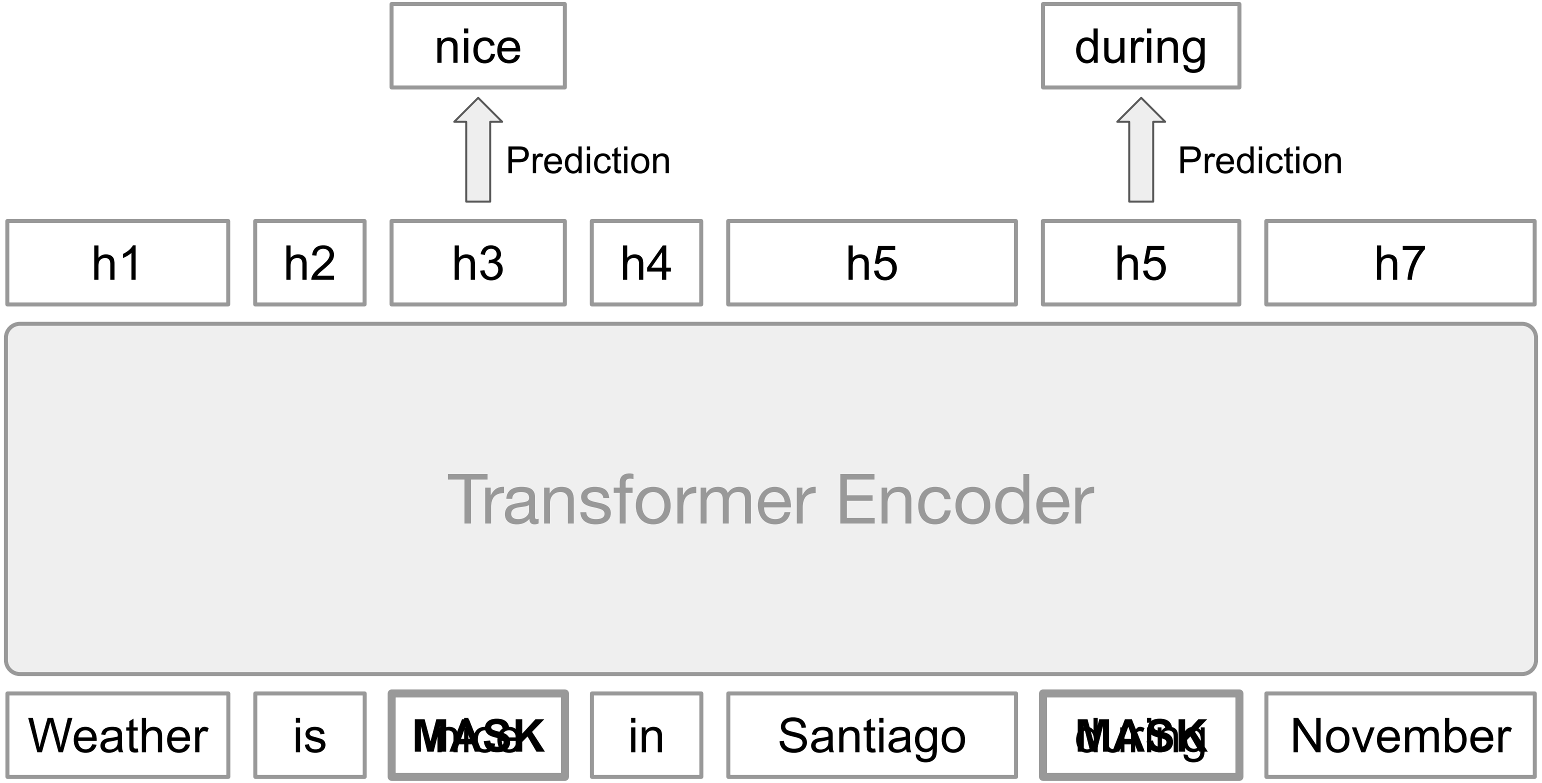
Transformer Encoder

Weather is nice in Santiago during November

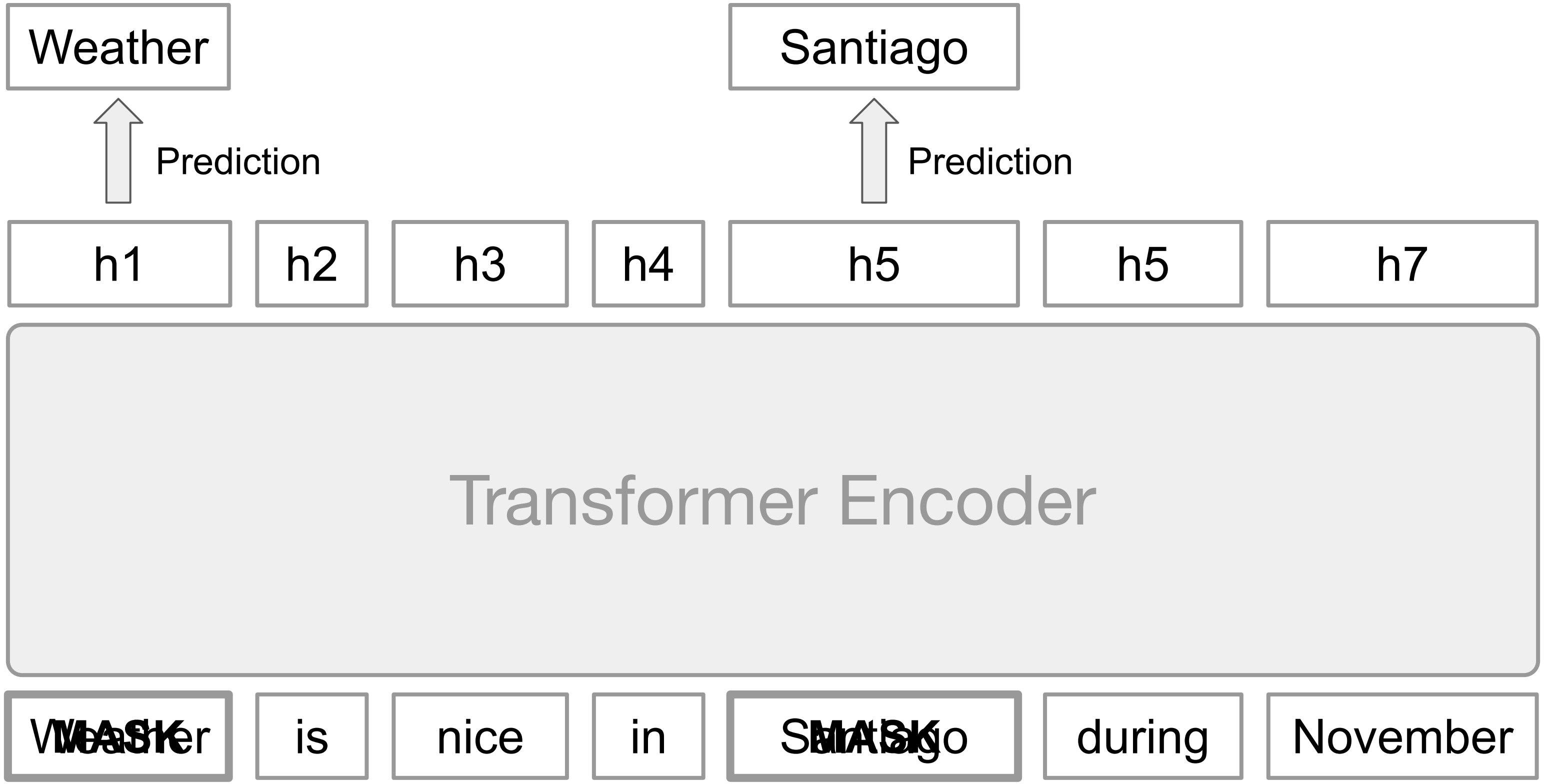
BERT: **B**idirectional **E**ncoder **R**epresentations from **T**ransformers

[Delvin et al. 2019]

Masked Language Model (MLM)



Masked Language Model (MLM)



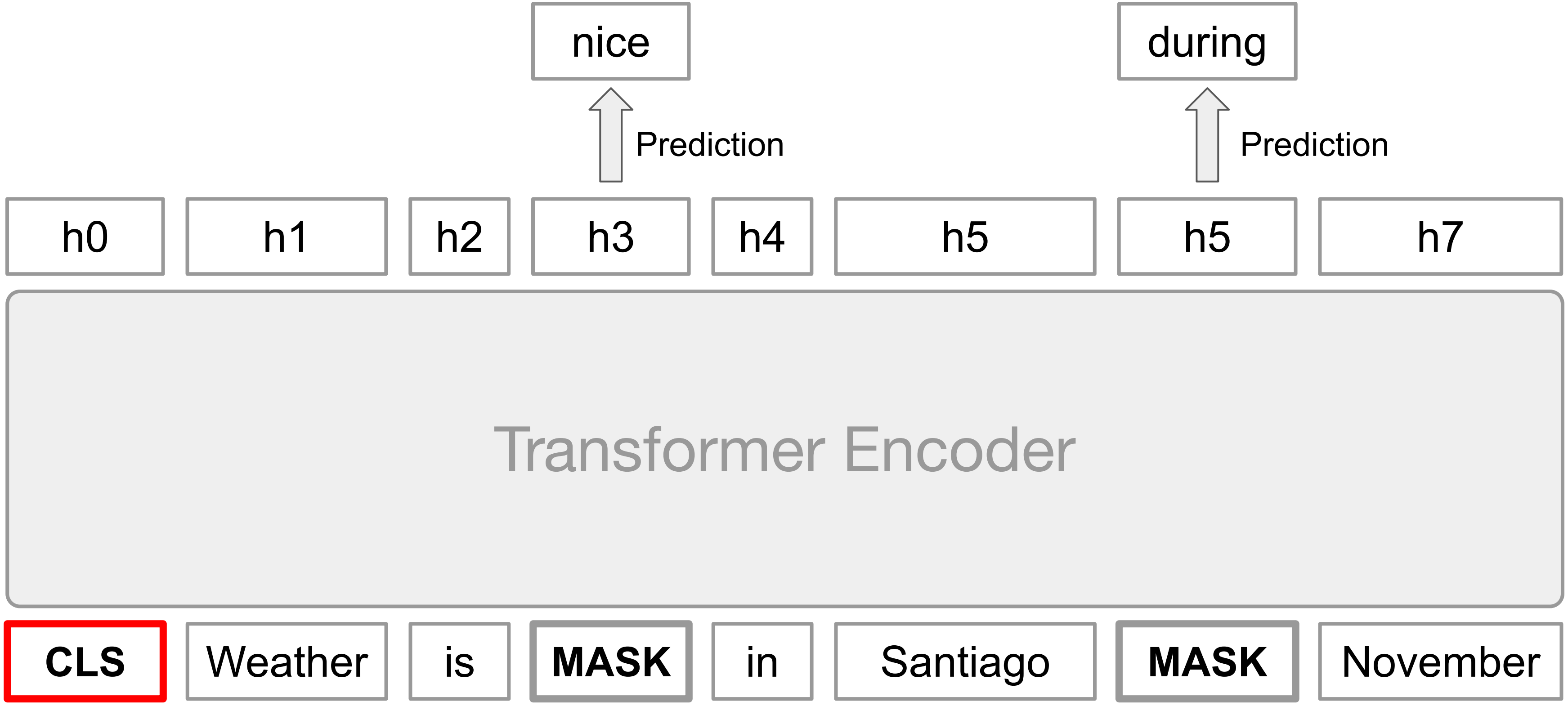
Unsupervised Pre-training

- **Pick an architecture** and tons of text

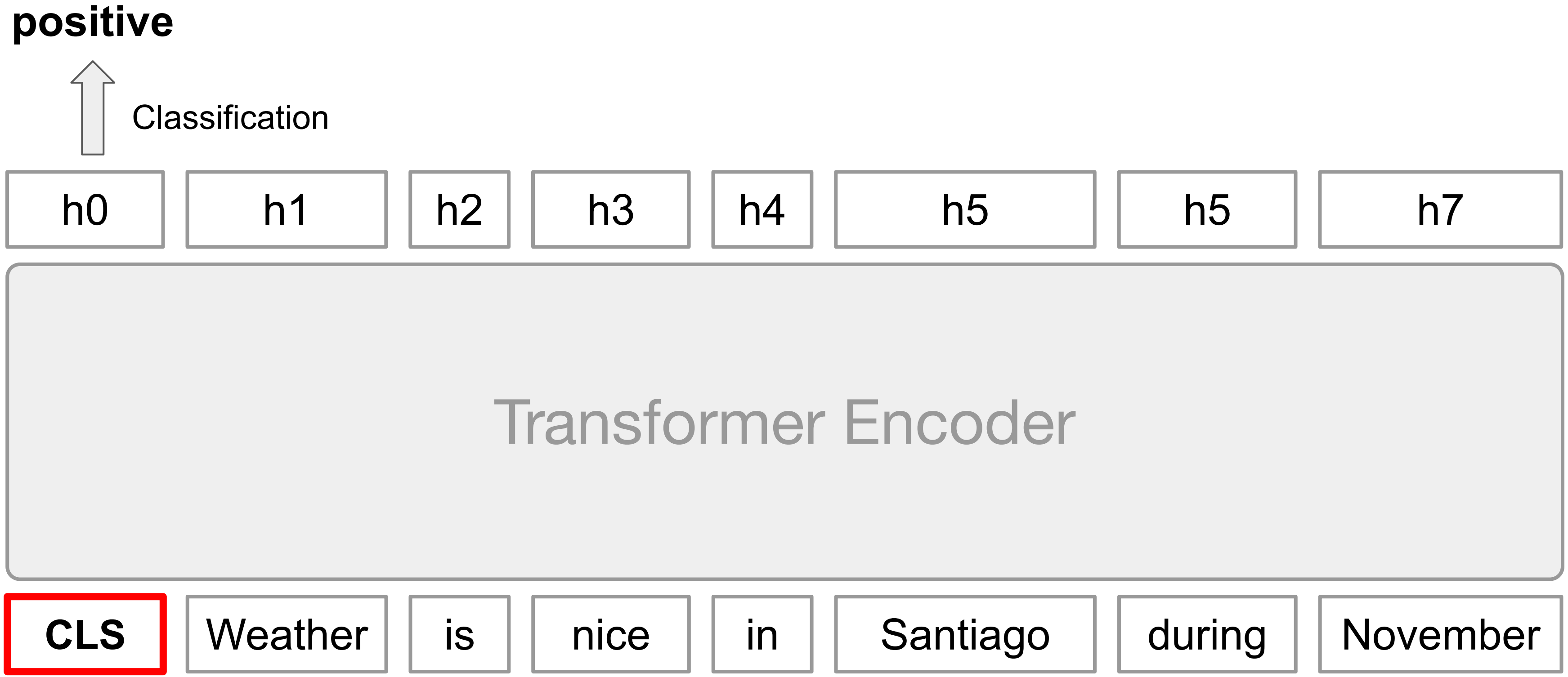
- Train the architecture with a
  general LM task (unsupervised)

- Use the weights as a starting point
  for training with a supervised task

Classification (**CLS**) token during pre-training

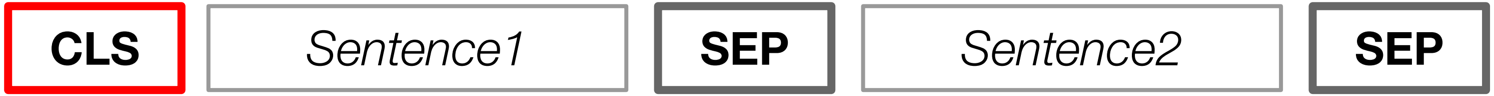


Classification token during fine tuning



Other BERT technicalities

- *Next Sentence Prediction*:



CLS *Sentence1* **SEP** *Sentence2* **SEP**

¿Is “Sentence2” the sentence that follows “Sentence1”?

Other BERT technicalities

- *Positional* and *Segment Encodings*
- ~15% corruptions during pre-training
- Base: 12L - 768H - 12A (110M parameters)
- Large: 24L - 1024H - 16A (340M parameters)

Spanish BERT

Spanish Corpus for Pre-training

3 billion words, including:

- Spanish Wikipedia
- Spanish portion of EUBookshop
- Spanish OpenSubtitles, TED Talks, News, etc.

https://github.com/josecannete/spanish-corpora

Spanish BERT

- *Whole-Word Masking*
- *Dynamic Masking*
- cased/uncased

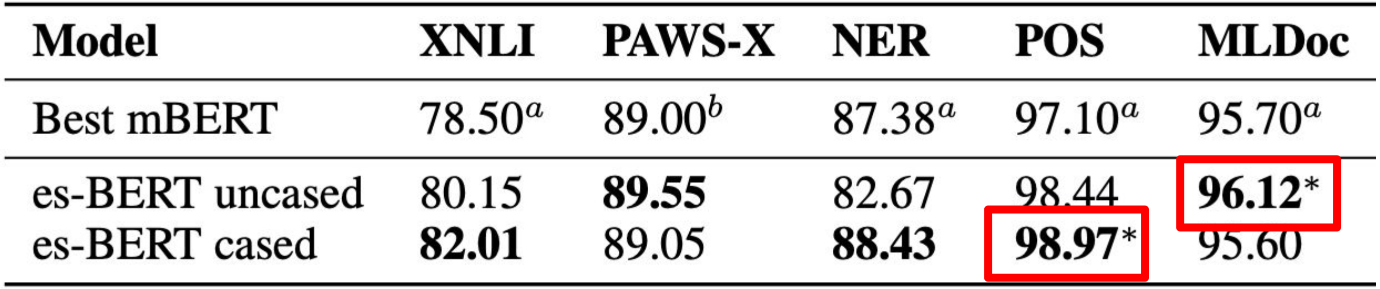
https://github.com/dccuchile/beto

Downstream Tasks

- Natural Language Inference
- Paraphrasing
- Named Entity Recognition
- Part of Speech
- Document Classification
- Question Answering

https://github.com/dccuchile/glues

Spanish BERT vs Multilingual BERT



Model	**XNLI**	**PAWS-X**	**NER**	**POS**	**MLDoc**
Best mBERT	78.50^a	89.00^b	87.38^a	97.10^a	95.70^a
es-BERT uncased	80.15	**89.55**	82.67	98.44	**96.12***
es-BERT cased	**82.01**	89.05	**88.43**	**98.97***	95.60

Model	**MLQA, MLQA**	**TAR, XQuAD**	**TAR, MLQA**
Best mBERT	53.90 / 37.40^c	**77.60** / **61.80**^{d}	68.10 / **48.30**^{d}
es-BERT uncased	67.85 / **46.03**	77.52 / 55.46	68.04 / 45.00
es-BERT cased	**68.01** / 45.88	77.56 / 57.06	**69.15** / 45.63

Future Work

- Add new tasks
- Train smaller models

Spanish Pre-trained BERT Model and Evaluation Data

https://github.com/dccuchile/beto

José Cañete Gabriel Chaperón Rodrigo Fuentes Jou-Hui Ho Hojing Kang Jorge Pérez

dcc
CIENCIAS DE LA COMPUTACIÓN
UNIVERSIDAD DE CHILE

Instituto Milenio
Fundamentos
de los datos

die
DEPARTAMENTO DE INGENIERÍA ELÉCTRICA
UNIVERSIDAD DE CHILE