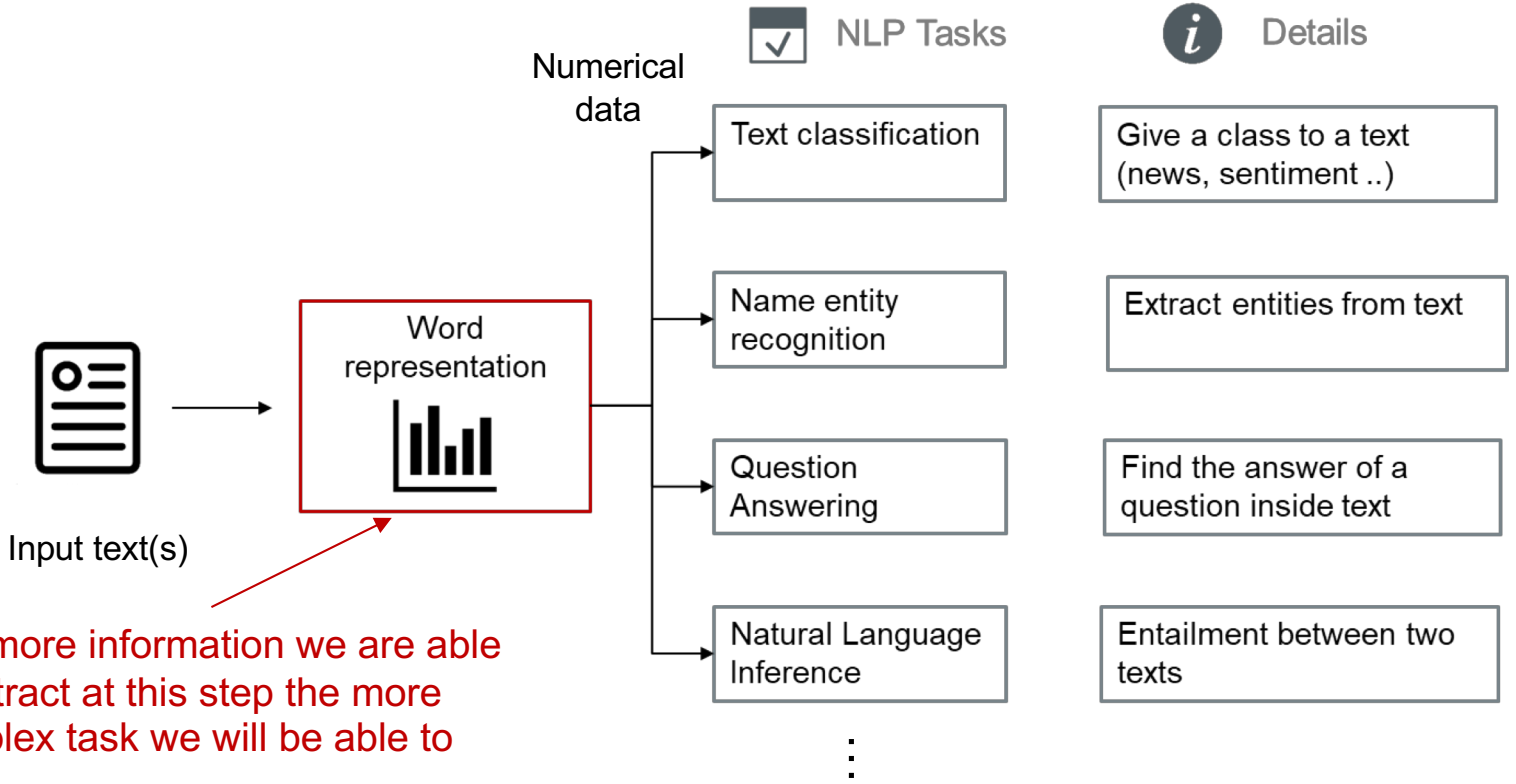
WORD REPRESENTATION FOR NATURAL LANGUAGE PROCESSING

CONTEXTUALIZED WORD REPRESENTATIONS

Tristan Karch
BNP Paribas AI Ted Talk
February 22, 2019

Reminder on NLP tasks

Word representation's goal: Convert text into numerical data so that it can be used by Machine Learning algorithms



Recap of previous presentation

Occurrence based

John likes to watch movies.

Mary likes movies too

“John”	1
“likes”	2
“to”	1
“watch”	1
“movies”	2
“Mary”	1
“too”	1

→ Represent documents by counting the occurrences of apparition of words

No **word level** information.

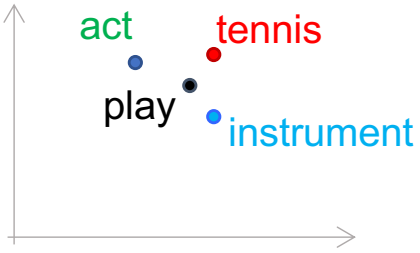
The counting of words only give a numerical representation of the whole document

Word embedding

play act

play tennis

play instrument



→ Represent the meaning of words by analyzing in which context they appear

Training: We only consider the context to **construct** the vector space

Predictions: We do not leverage the context in which word appears at prediction time.

In the vector space, the word representation is **frozen** and **unique**

→ One word can only have one meaning

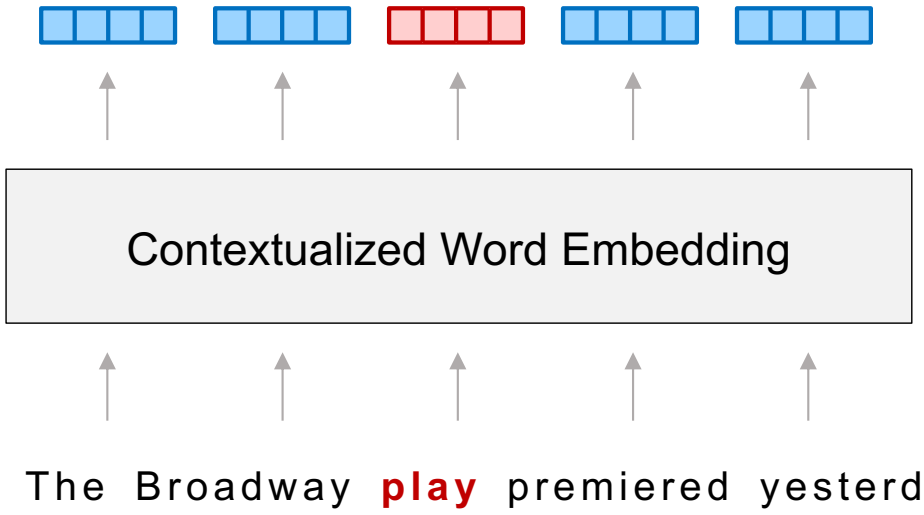
Contextualized word representation

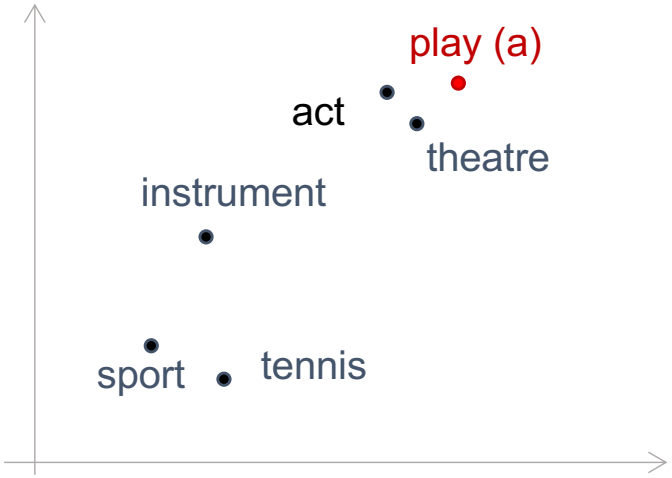
Understanding the different meaning of the same word

(a) The Broadway **play** premiered yesterday.

(b) I like to **play** tennis.

Overview of the model:





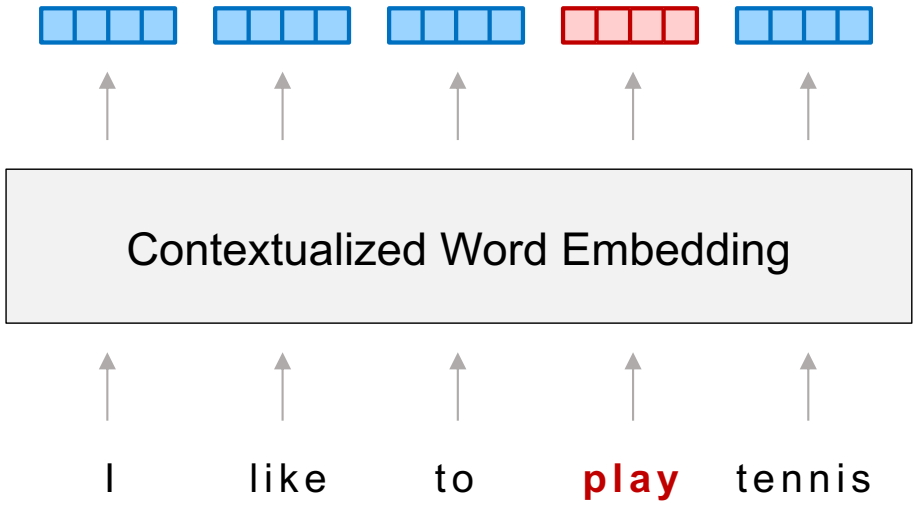
Contextualized word representation

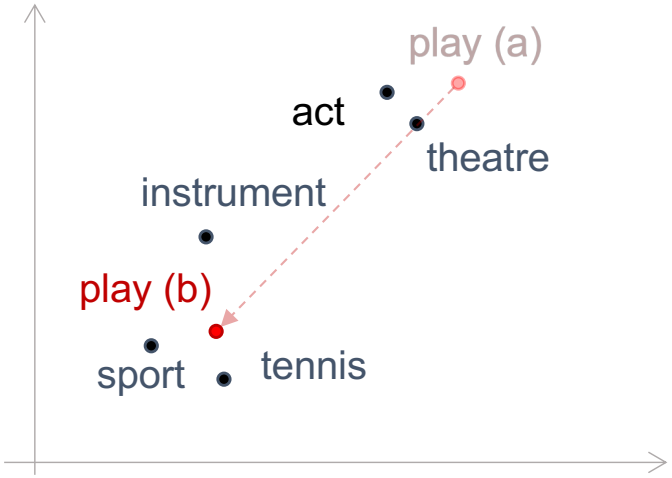
Understanding the different meaning of the same word

(a) The Broadway **play** premiered yesterday.

(b) I like to **play** tennis.

Overview of the model:





Contextualized word representation

Understanding the different meaning of the same word

(a) The Broadway **play** premiered yesterday.

(b) I like to **play** tennis.

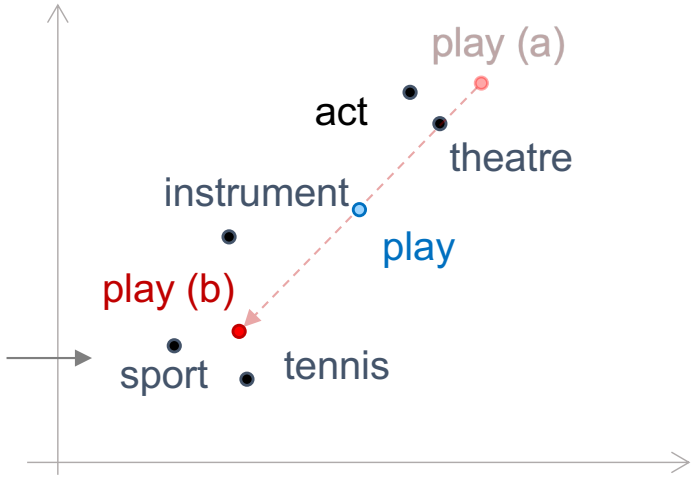
Comparison with non-contextualized word representation

play act

play tennis

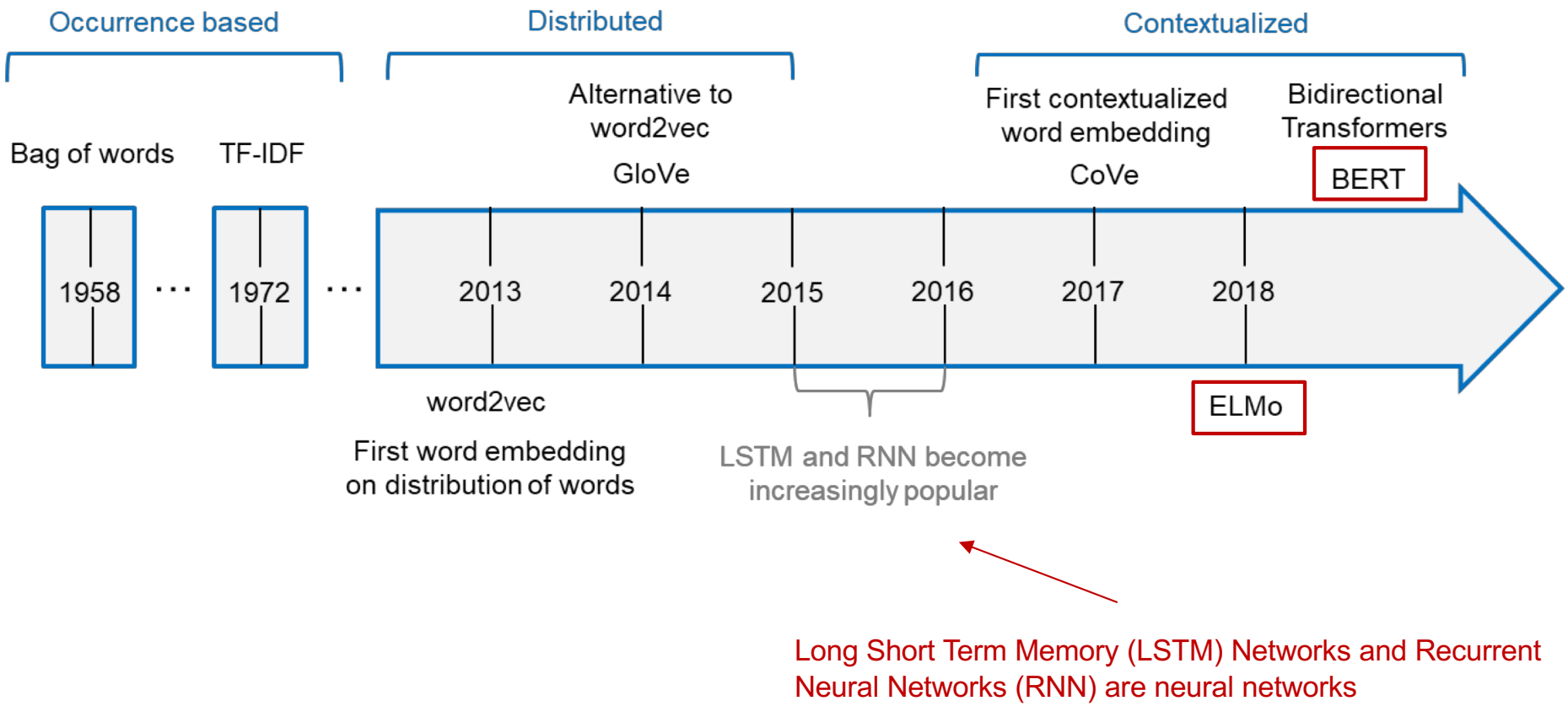
play instruction

Non-contextualized word embedding

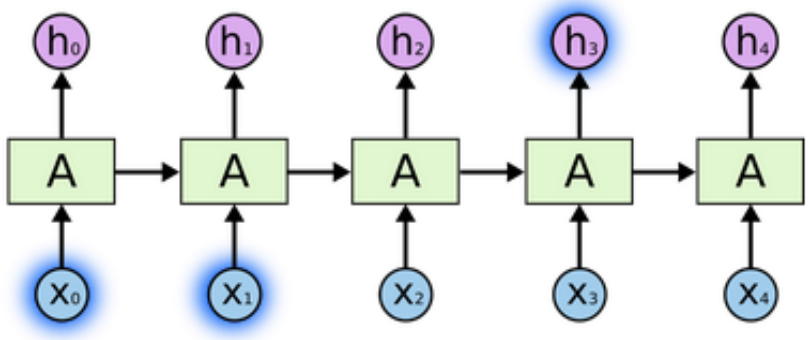


→ Word position analyzes **all contexts** and is not able to adapt to the different contexts

Different Types of Word Representations



Long Short Term Memory (LSTM) Network



Neural network that allows to analyze sequences (a sentence is a sequence of words)

Internal state (memory) that is conveyed across the different states of a sequence

- **Forgetting function**: What information should I forget from the previous steps?
- **Input function:** What information does the input bring to the internal state?
- **Output function:** What information carried by my internal state should I output?

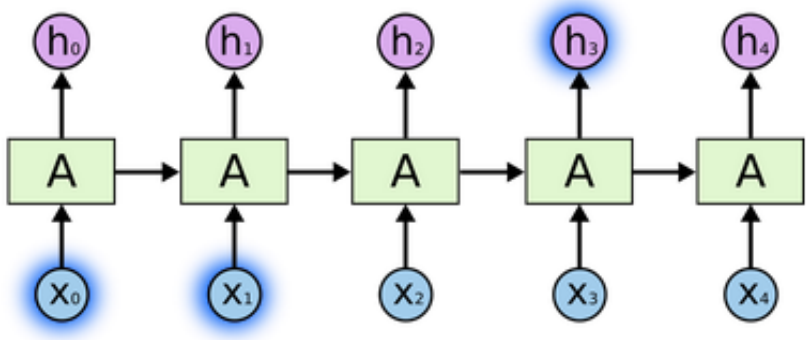
→ Memory that allows to save some information during the analysis of a sequence

Example: We want to build a model that predicts the next word of a sentence:

I grew up in France I speak fluent

x_0 x_1 x_2 x_3 x_4 x_5 x_6 x_7

Long Short Term Memory (LSTM) Network



Neural network that allows to analyze sequences (a sentence is a sequence of words)

Internal state (memory) that is conveyed across the different state of a sequence

- **Forgetting function**: What information should I forget from the previous steps?
- **Input function:** What information does the input bring to the internal state?
- **Output function:** What information carried by my internal state should I output?

→ Memory that allows to save some information during the analysis of a sequence

Example: We want to build a model that predicts the next word of a sentence:

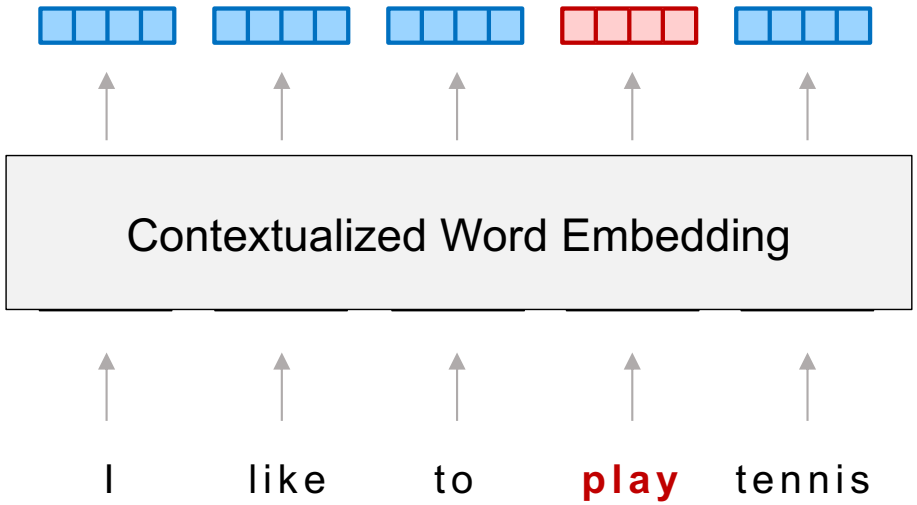
I grew up in France I speak fluent ⟶ french

x_0 x_1 x_2 x_3 x_4 x_5 x_6 x_7

ELMo - Deep contextualized word representations

Introducing ELMo

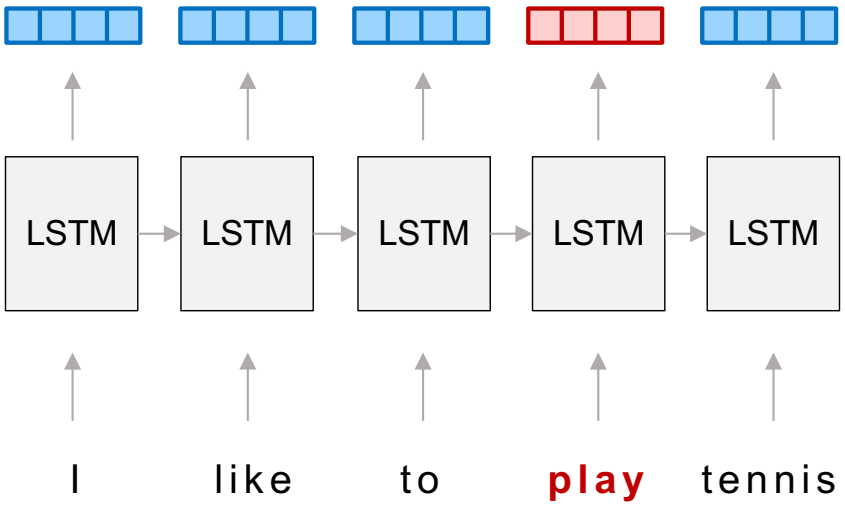
So LSTM seems like a good way to capture the influence of the context on the meaning of the words inside a sentence



ELMo - Deep contextualized word representations

Introducing ELMo

So LSTM seems like a good way to capture the influence of the context on the meaning of the words inside a sentence

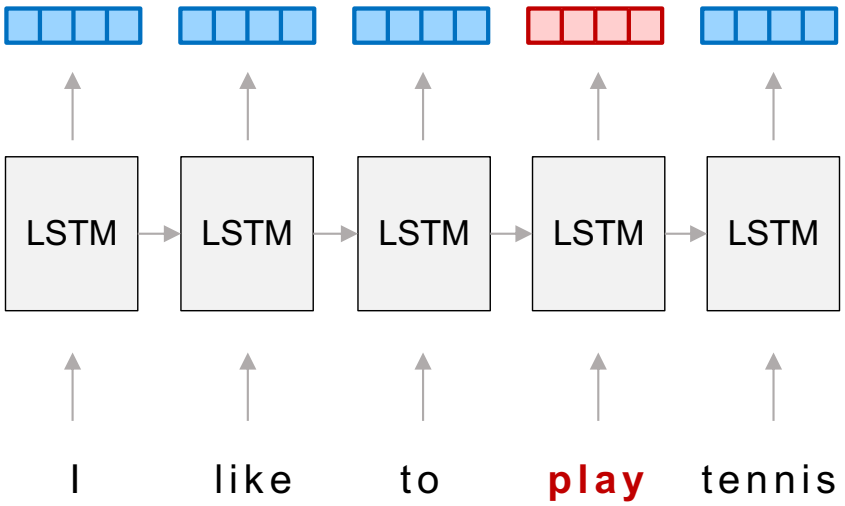


The LSTM sequentially takes the words of the sentence in order to output its **contextualized** word representation:

ELMo - Deep contextualized word representations

Introducing ELMo

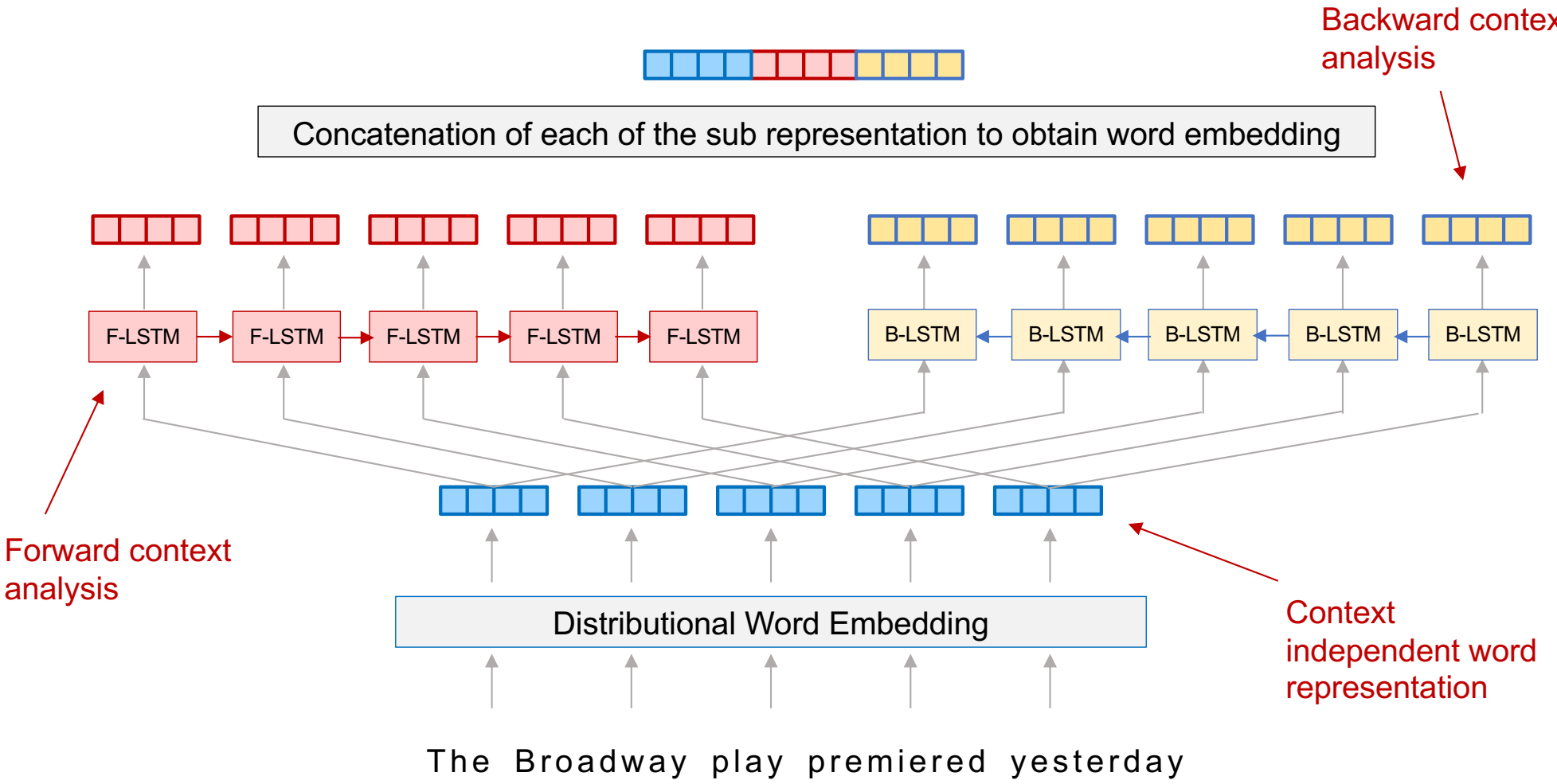
So LSTM seems like a good way to capture the influence of the context on the meaning of the words inside a sentence



Issue: We can only use past information with LSTM
→ The contextualized word representation is only influenced by **past context**

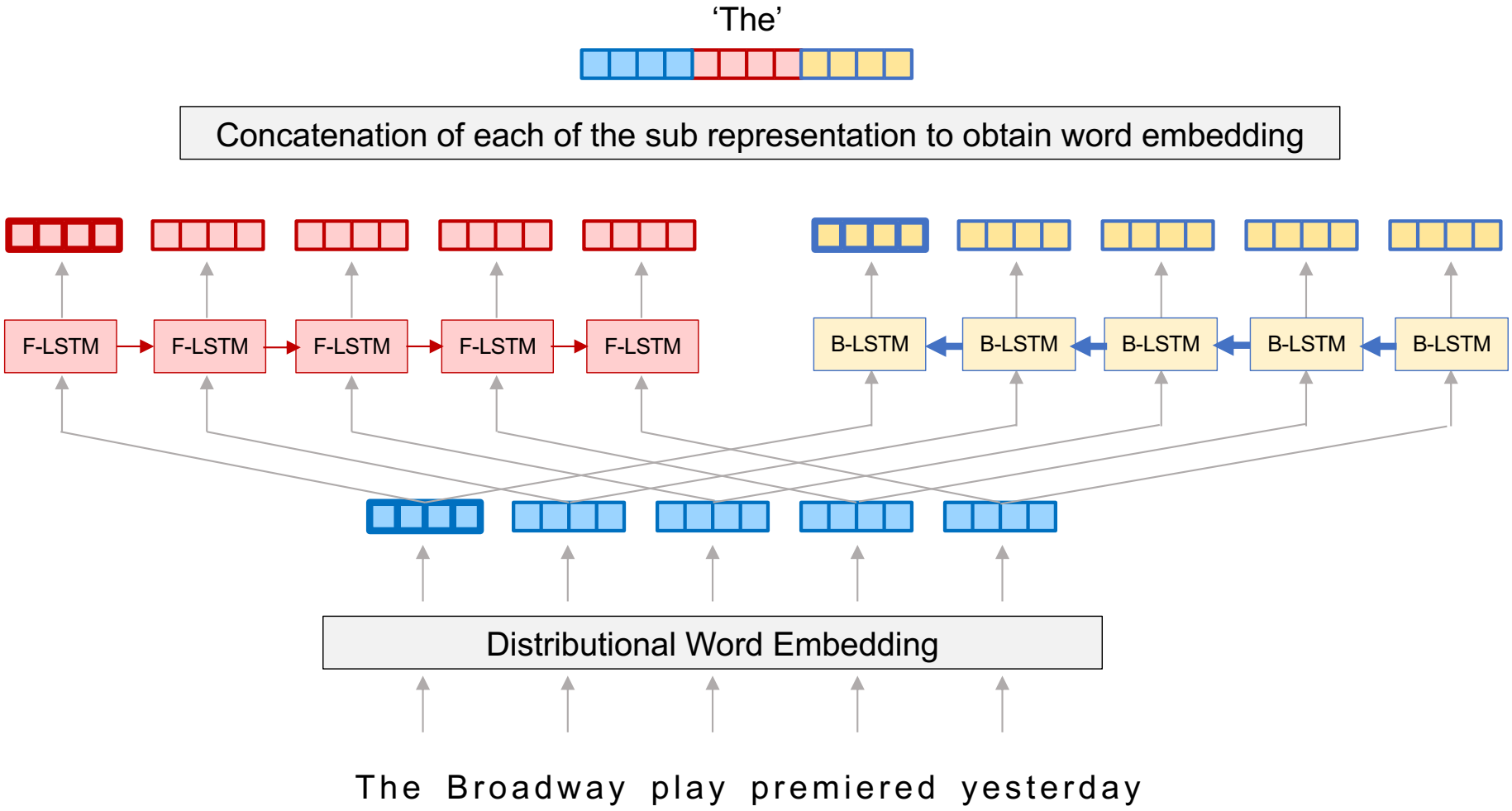
ELMo - Deep contextualized word representations

The full model explained



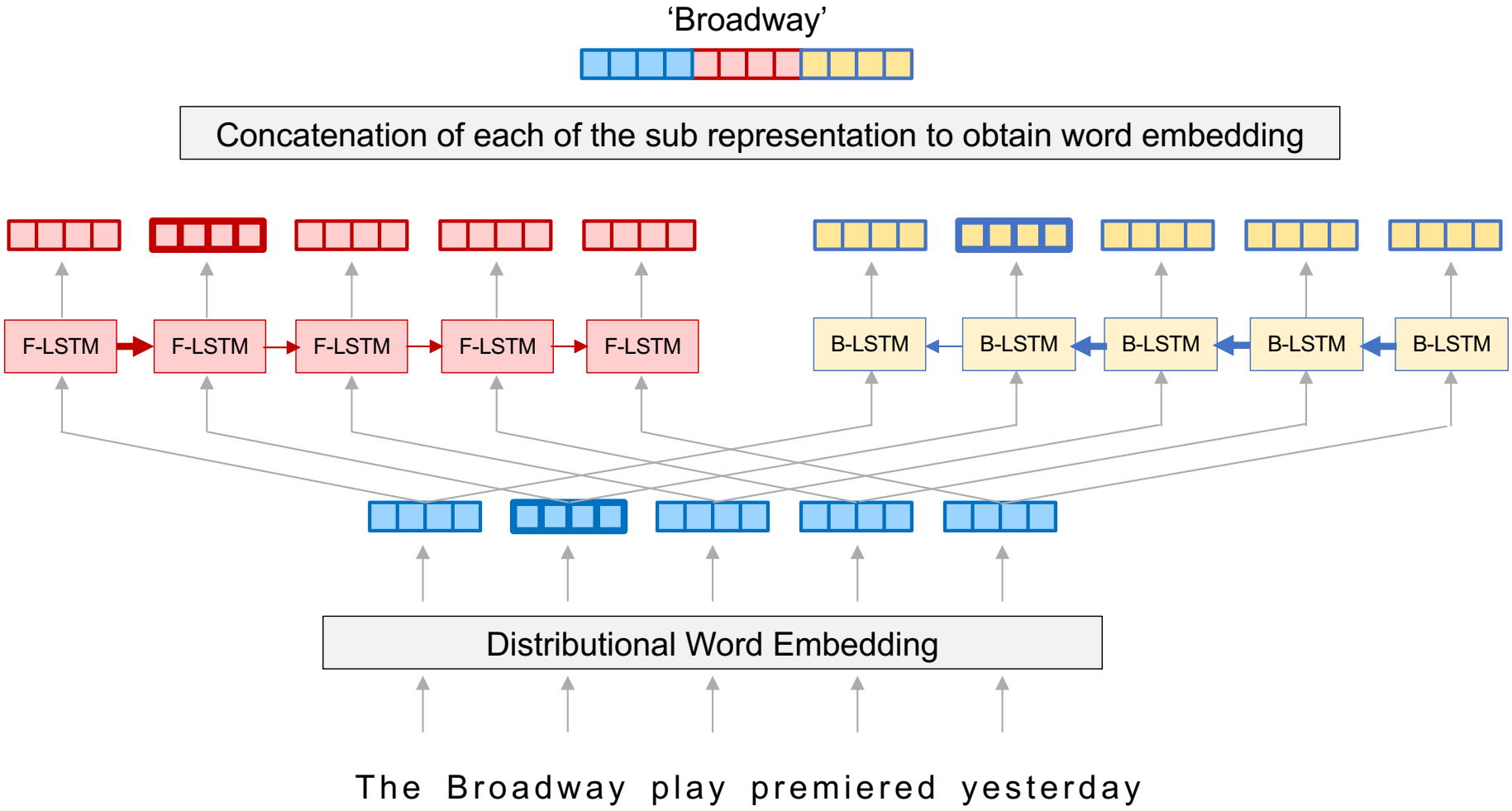
ELMo - Deep contextualized word representations

The full model explained



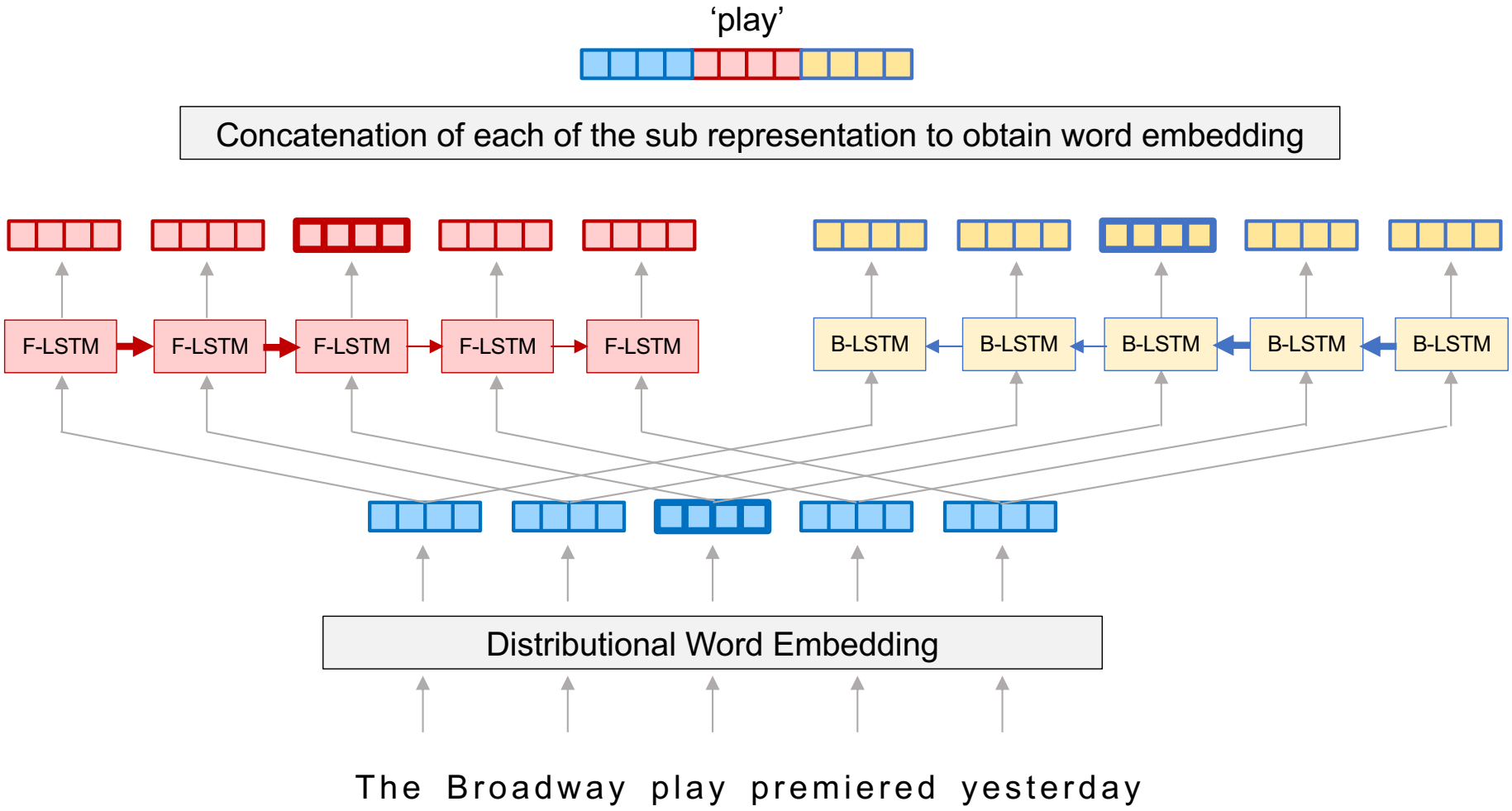
ELMo - Deep contextualized word representations

The full model explained



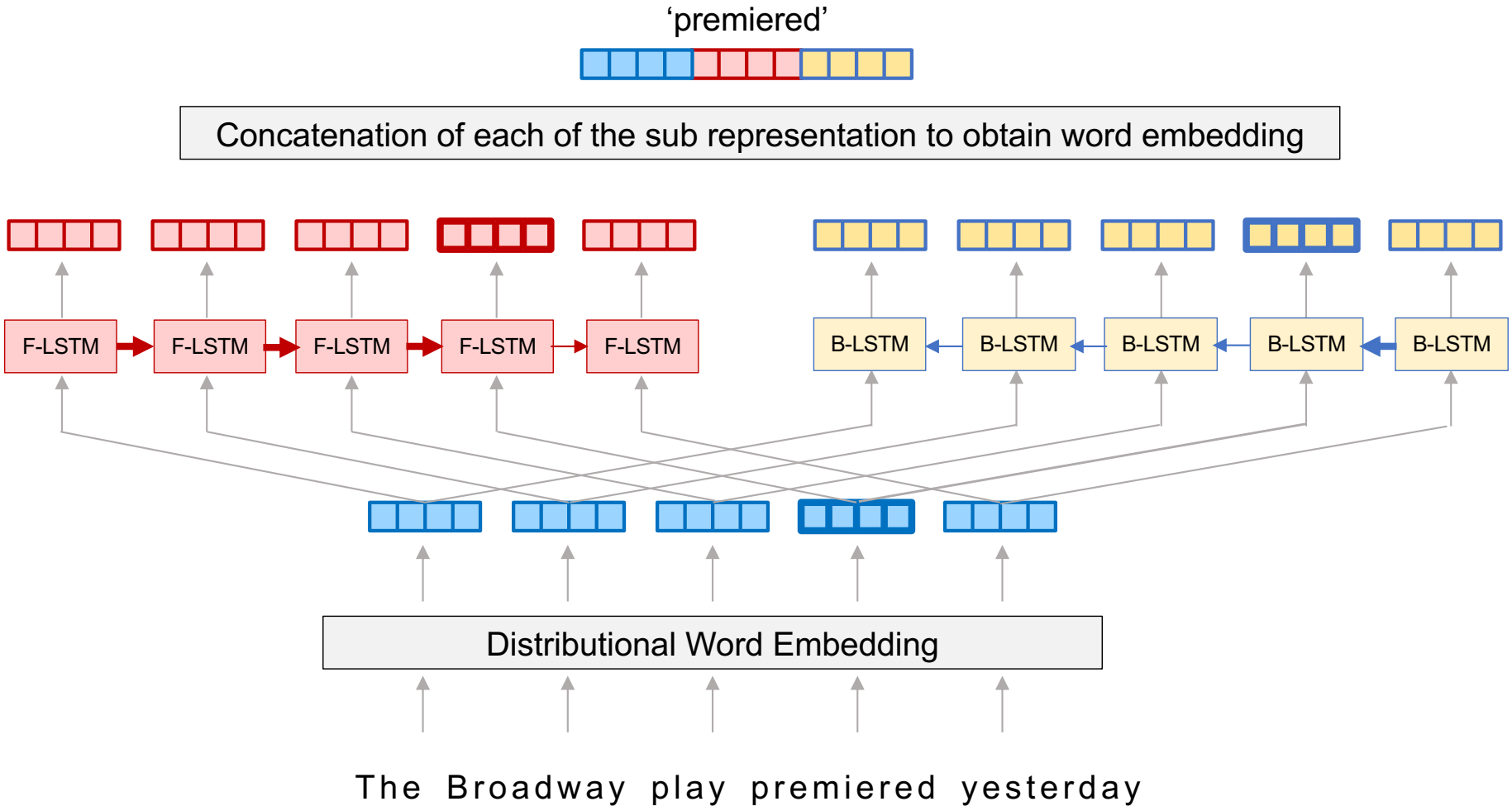
ELMo - Deep contextualized word representations

The full model explained



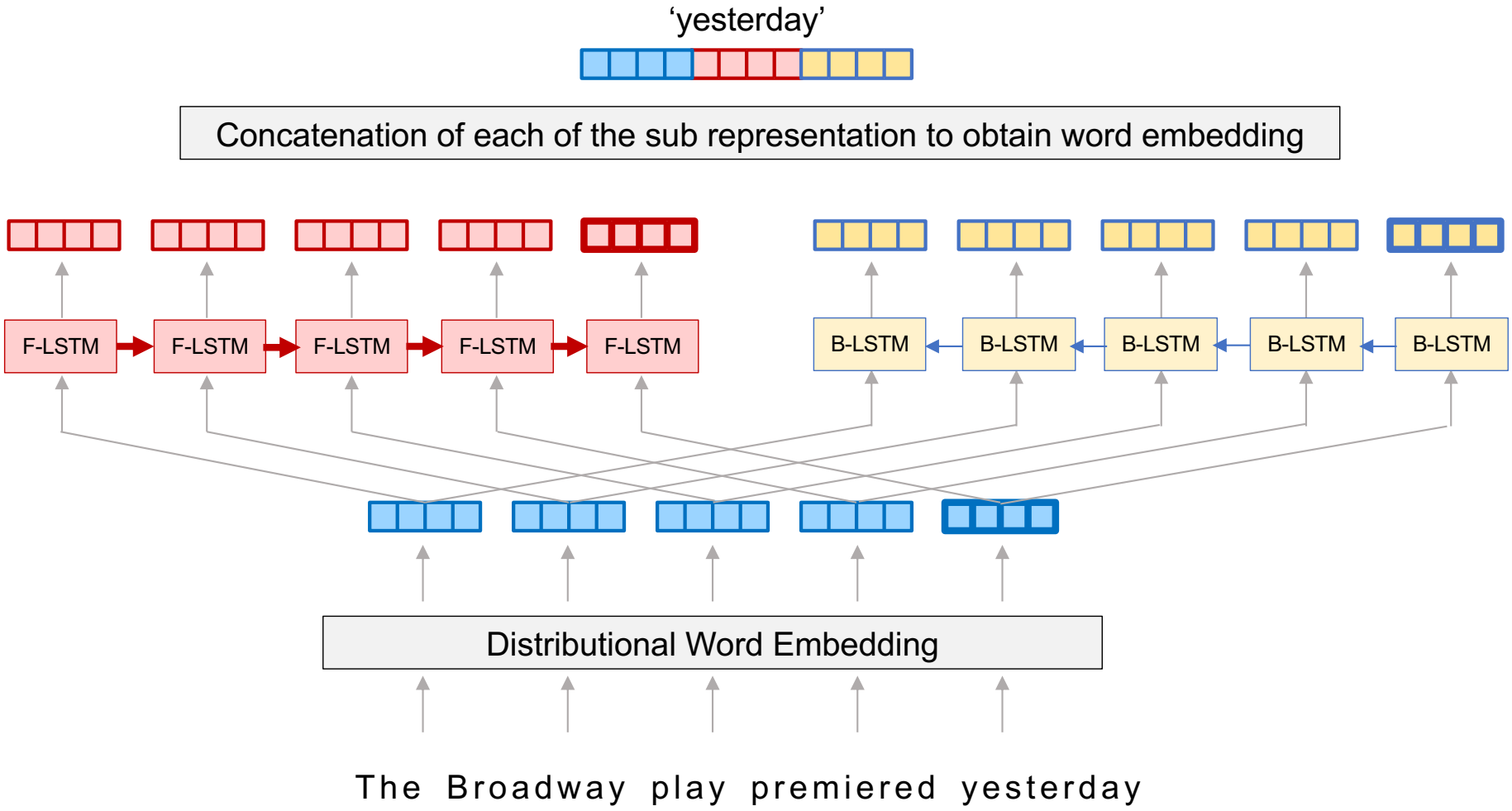
ELMo - Deep contextualized word representations

The full model explained



ELMo - Deep contextualized word representations

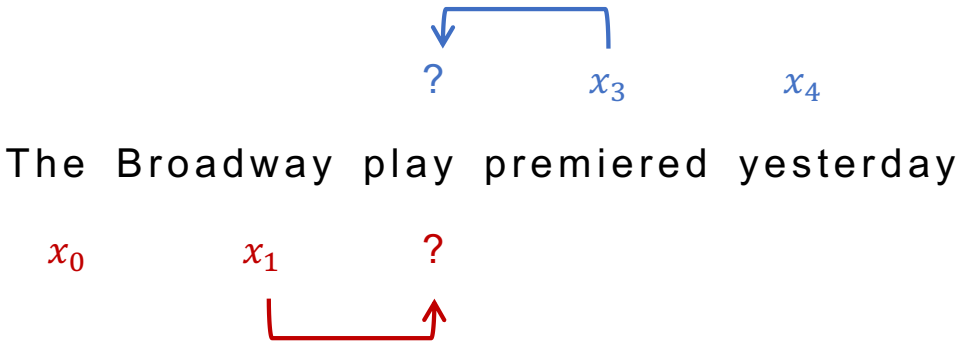
The full model explained



ELMo - Deep contextualized word representations

How do we train each sub representation?

We use a language model. We read the sequence in both directions and use the LSTM to predict the next word of the sentence:

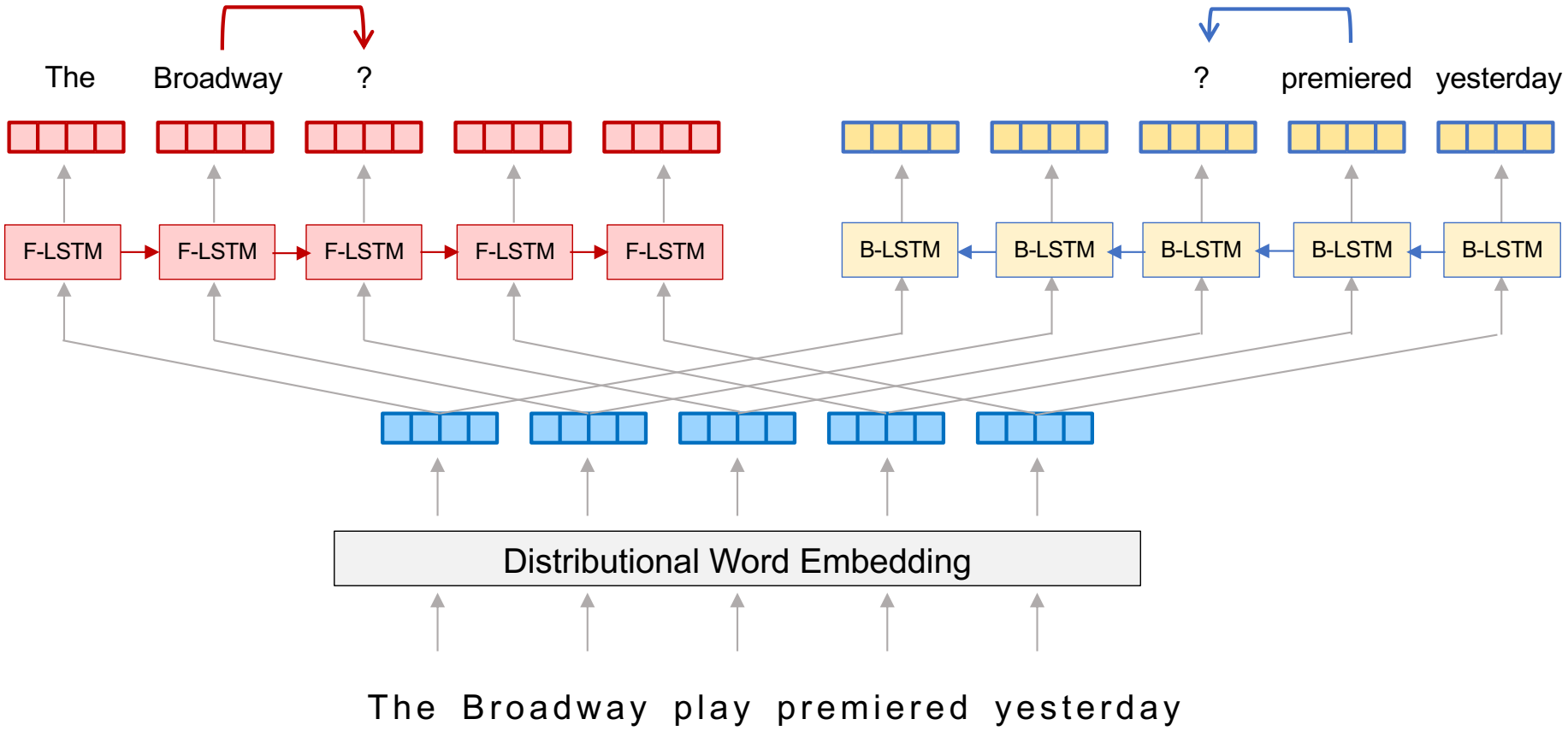


The **forward** LSTM reads the sequence **from left to right** and tries to predict the word 'play' **given** the history ['The', 'Broadway'] to learn the **forward** word representation:

The **backward** LSTM reads the sequence **from right to left** and tries to predict the word 'play' **given** the history ['yesterday', 'premiered'] to learn the **backward** word representation:

ELMo - Deep contextualized word representations

How do we train each sub representation?



ELMo - Deep contextualized word representations

What does the LSTMs learn?

We repeat this prediction task over **billions of sequences**.

By showing many sequences to the two LSTMs and predicting many next words inside their context, we guide the **internal state** of the LSTMs to **forget and remember** the correct information.

From a grammatical point of view, it means that we **convey the correct information about the grammar** (gender, subjects and other contextual information).

For instance when we see a new subject in a sequence, we want to **forget the gender** of the old subject.

To sum up:

The ELMo word vector is a composition of three sub representation

Context independent values

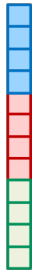
Left context values

Right context values

ELMo - Deep contextualized word representations

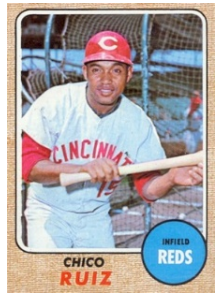
Example of ELMo capabilities

In this example we try to find the nearest neighbors of the word ‘play’ using non-contextualized word embedding (part 1) and using contextualized word embedding:



	Source	Nearest Neighbors
Word2vec	play	playing, game, games, played, players, plays, player, Play, football, multiplayer
ELMo	Chico Ruiz made a spectacular play on Alusik ’s grounder {...}	Kieffer , the only junior in the group , was commended for his ability to hit in the clutch , as well as his all-round excellent play .
	Olivia De Havilland signed to do a Broadway play for Garson {...}	{...} they were actors who had been handed fat roles in a successful play , and had talent enough to fill the roles competently , with nice understatement .

Paragraph from wikipedia



Baseball context correctly caught

Acting context correctly caught

ELMo - Deep contextualized word representations

When should we use ELMo?

The LSTMs of the ELMo model capture **local context** and learn how words are related to each other inside a sentence in order to influence their meanings.

Text classification

The class of a text (for instance the type of a news) is a **general property** of the input.

Adding a local context analysis is not necessarily useful

TF-IDF and BoW give higher level information about the text and are usually sufficient to tackle such task

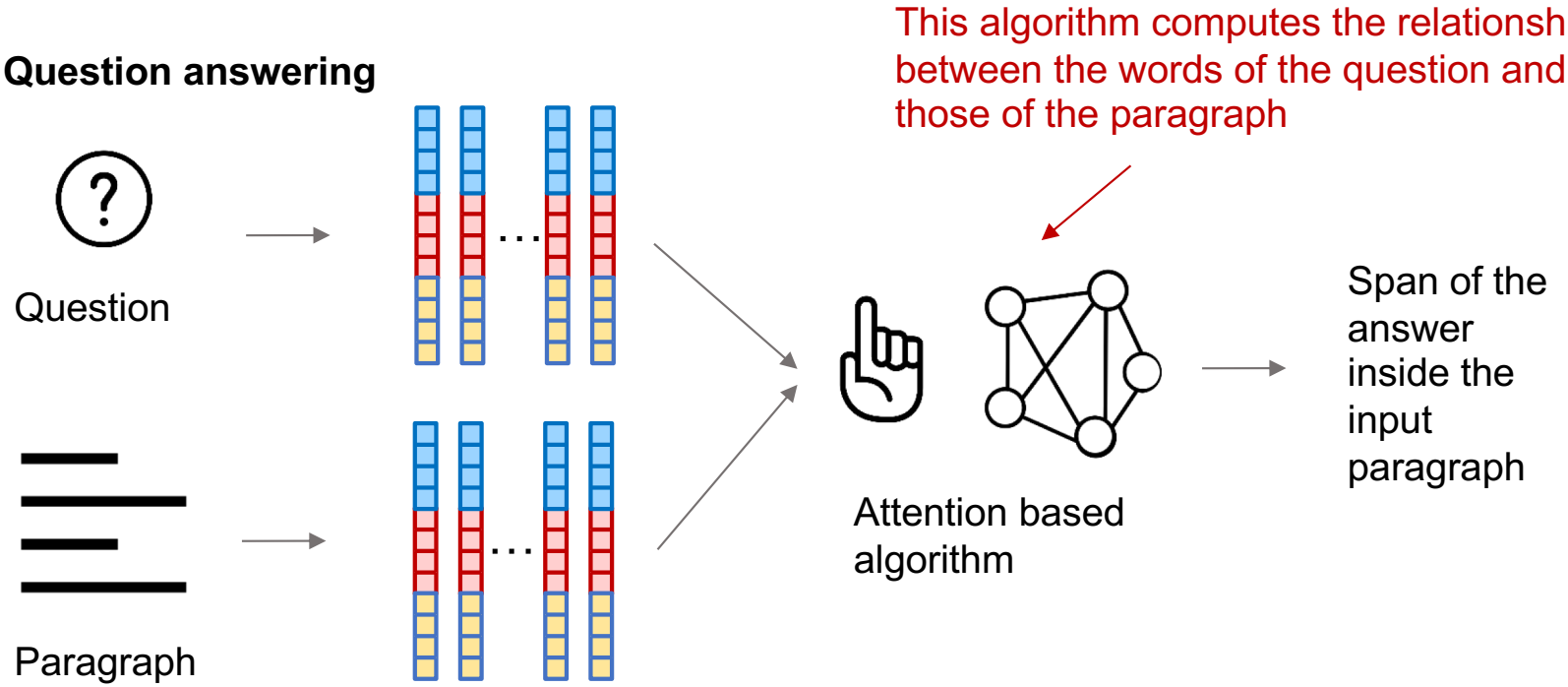
Question answering

To answer question Elmo allows to focus on local part of text leveraging the context and the grammar thanks to the LSTMs.

ELMo - Deep contextualized word representations

Using ELMo for NLP tasks

Question answering



The analysis of the context of the question compared to the different local context of the paragraph allows to pay a particular **attention** to certain words to find the answer

ELMo - Deep contextualized word representations

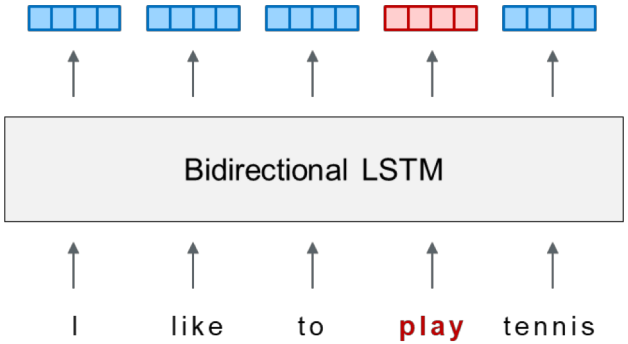
Pros	Cons
• The ELMo embedding is able to catch different meanings of the same word • It allows to **solve harder NLP task** such as **question answering** because the word vectors carry the information about the context.	• Computationally intensive to train • Need for attention based algorithm on top of word representation to correctly catch the relationship between the question and the paragraph → This complication is solved with BERT

BERT Bidirectional Encoder Representations from Transformers

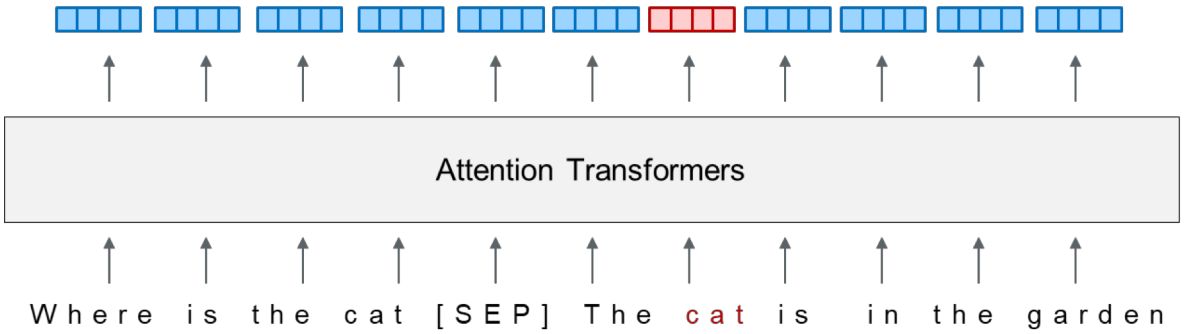
Introducing BERT

- BERT is the most advance contextualized word representation (released by google in Oct 2018)
- It also construct a word representation that varies with the context in which the word is taken into consideration.

ELMo vs BERT



- Take a **single** sentence as input
- Model context with LSTMs



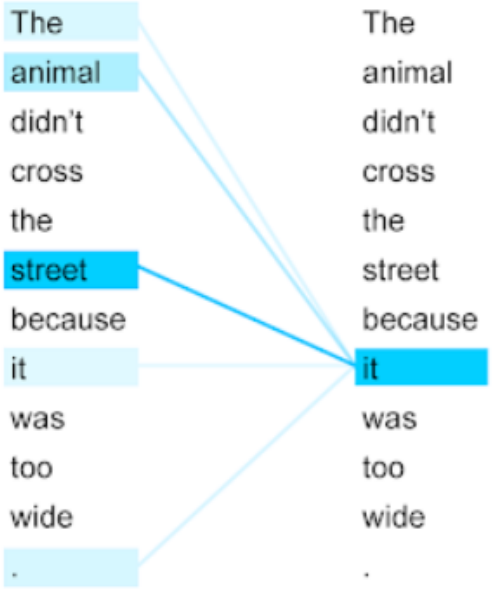
- Take **two** sentence as input
- Model context with **Attention Transformers**

BERT Bidirectional Encoder Representations from Transformers

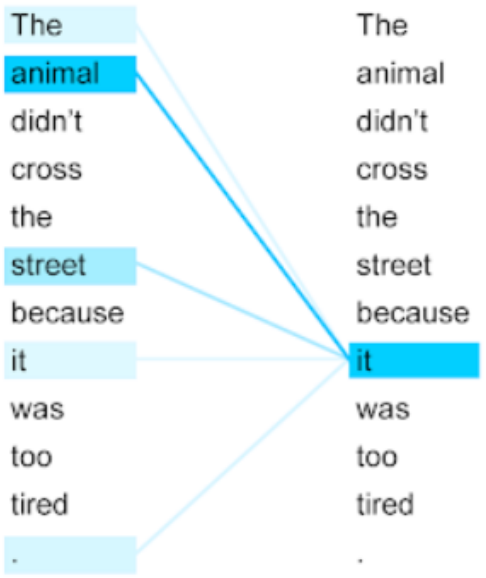
The attention mechanism

Attention and Transformer are more complex than LSTM. They also allow to compute how a word interact with the sequence at a grammatical and semantic level. They seem to be more robust than LSTM

The animal didn't cross the street because it was too wide.



The animal didn't cross the street because it was too tired.



BERT Bidirectional Encoder Representations from Transformers

A word representation designed for question answering

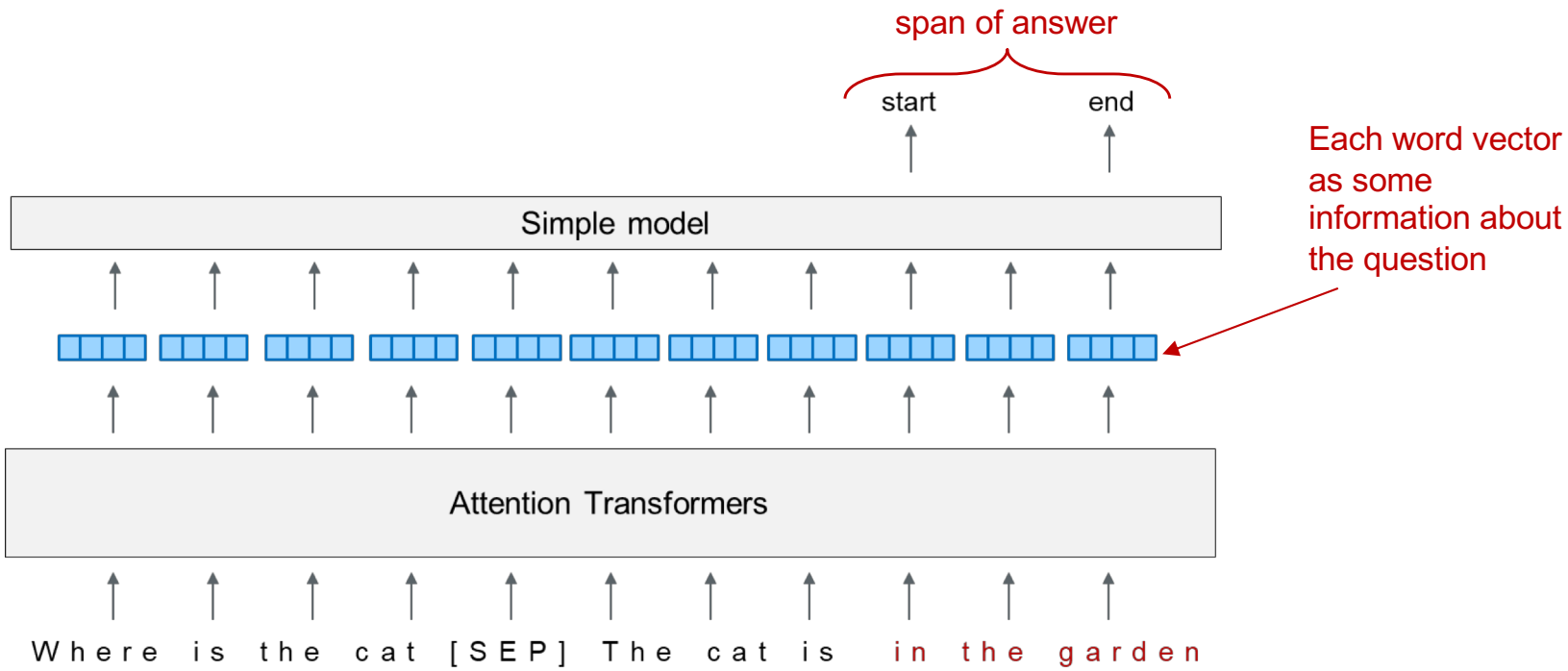
The concept of attention used to replace ELMo's LSTMs is similar to the algorithm used to solve the question answering task with the ELMo's vectors

We can give the question and the paragraph to the BERT model

BERT's word representation **embeds** the ability to **compute the interaction of the question with the paragraph**

BERT Bidirectional Encoder Representations from Transformers

A word representation designed for question answering



By feeding such model with enough data of the form (Question, paragraph, span of answer). The model can learn the **correct association of words** inside the paragraph that answers the question

→ Show training set https://rajpurkar.github.io/SQuAD-explorer/explore/1.1/dev/

BERT Bidirectional Encoder Representations from Transformers

Demo of question answering

Paragraph

BNP Paribas has been a major sponsor of tennis. In 1973 it became the major sponsor of the French Open, one of the four prestigious Grand Slam tournaments in the sport. In 2001 the bank began to sponsor the Davis Cup before becoming the title sponsor in 2002. Also in 2002 it became the sponsor of the Paris Masters, one of the nine high-profile tournaments of the ATP World Tour Masters 1000 series. It also expanded to the United States in 2009 when it became the title sponsor of the Indian Wells Masters, a two-week tournament in California which is also one of the nine Masters 1000 series. The Stanford Classic, since 1992, is instead directly sponsored by the Bank of the West subsidiary.

Question

Which competition BNP Paribas decided to sponsor in the US?

→ Let's run the code!!!

BERT Bidirectional Encoder Representations from Transformers

Demo of question answering

Paragraph

Retail banking is BNP Paribas' largest business unit representing 72% of its 2015 revenues. Its operations are concentrated in Europe, especially in the group's three domestic markets of France, Italy (where it operates as Banca Nazionale del Lavoro (BNL)), and Belgium (as BNP Paribas Fortis). The group also owns an American subsidiary BancWest which operates as Bank of the West in the western United States and First Hawaiian Bank in Hawaii. BNP Paribas's Europe Mediterranean group also runs large retail banks in Poland, Turkey, Ukraine, and northern Africa. BNP Paribas is the largest bank in the Eurozone by total assets and second largest by market capitalization according to The Banker magazine, just behind Banco Santander. It employs over 189,000 people, according to the bank as of 31 December 2015, of which 147,000 work in Europe, and maintains a presence in 75 countries

Question

How many employees is there at BNP Paribas ?

→ Let’s run the code!!!

BERT Bidirectional Encoder Representations from Transformers

Demo of question answering

Paragraph

Retail banking is BNP Paribas' largest business unit representing 72% of its 2015 revenues. Its operations are concentrated in Europe, especially in the group's three domestic markets of France, Italy (where it operates as Banca Nazionale del Lavoro (BNL)), and Belgium (as BNP Paribas Fortis). The group also owns an American subsidiary BancWest which operates as Bank of the West in the western United States and First Hawaiian Bank in Hawaii. BNP Paribas's Europe Mediterranean group also runs large retail banks in Poland, Turkey, Ukraine, and northern Africa. BNP Paribas is the largest bank in the Eurozone by total assets and second largest by market capitalization according to The Banker magazine, just behind Banco Santander. It employs over 189,000 people, according to the bank as of 31 December 2015, of which 147,000 work in Europe, and maintains a presence in 75 countries

Question

How many employees is there at BNP Paribas in Europe ?

→ Let’s run the code!!!

BERT Bidirectional Encoder Representations from Transformers

Demo of question answering

Paragraph

In 2009, BNP Paribas reorganized its retail banking divisions renaming its Emerging Markets group the Europe Mediterranean group. This change was made because after the integration of Fortis Bank's Polish and Turkish subsidiaries, BNP Paribas's emerging market activities are now heavily concentrated in Eastern Europe and the southern half of the Mediterranean basin.

Question

"Where are BNP emerging markets

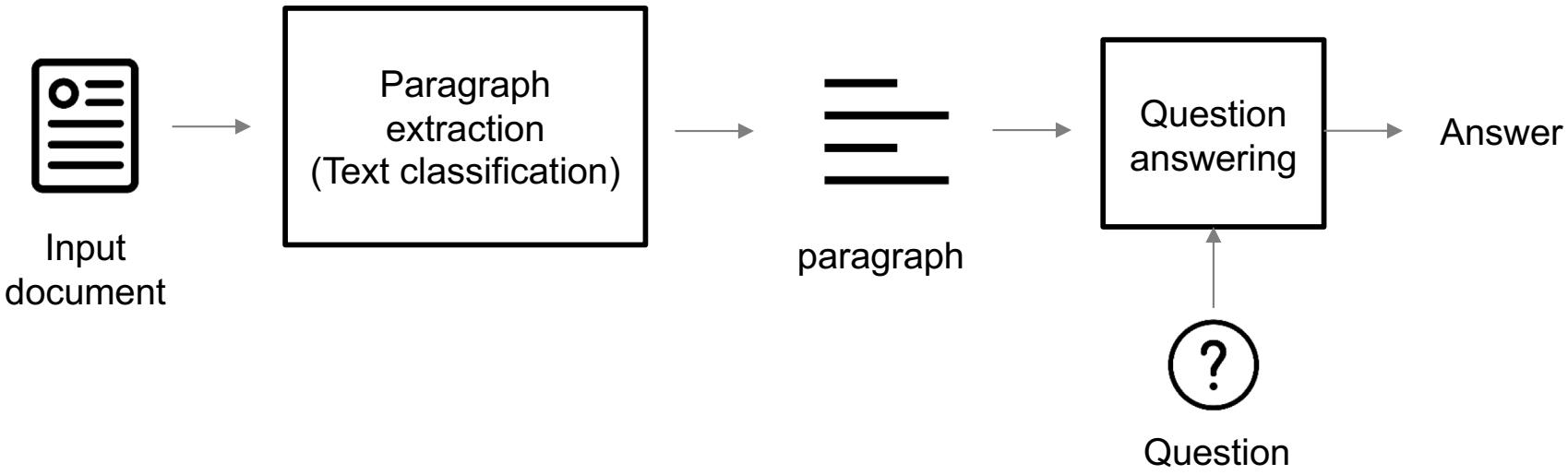
→ Let’s run the code!!!

Conclusion

Limitation of Question Answering models

The biggest limitation of QA models is the fact that we cannot give huge paragraphs as input to the model.

→ To answer questions in big document we have to split the problem into two steps.



References

LSTM:

- http://colah.github.io/posts/2015-08-Understanding-LSTMs/

ELMo:

- https://allennlp.org/elmo
- https://towardsdatascience.com/beyond-word-embeddings-part-2-word-vectors-nlp-modeling-from-bow-to-bert-4ebd4711d0ec

BERT:

- http://jalammar.github.io/illustrated-transformer/
- http://jalammar.github.io/illustrated-bert/