Language Translation Using NLP

Mohammad Azam

Resources and Credits

- How Google Translate Works (https://youtu.be/AIpXjFwVdIE?si=pW-cwys8NhPFYV56)
- Language Translation with RNNs (https://towardsdatascience.com/language-translation-with-rnns-d84d43b40571)

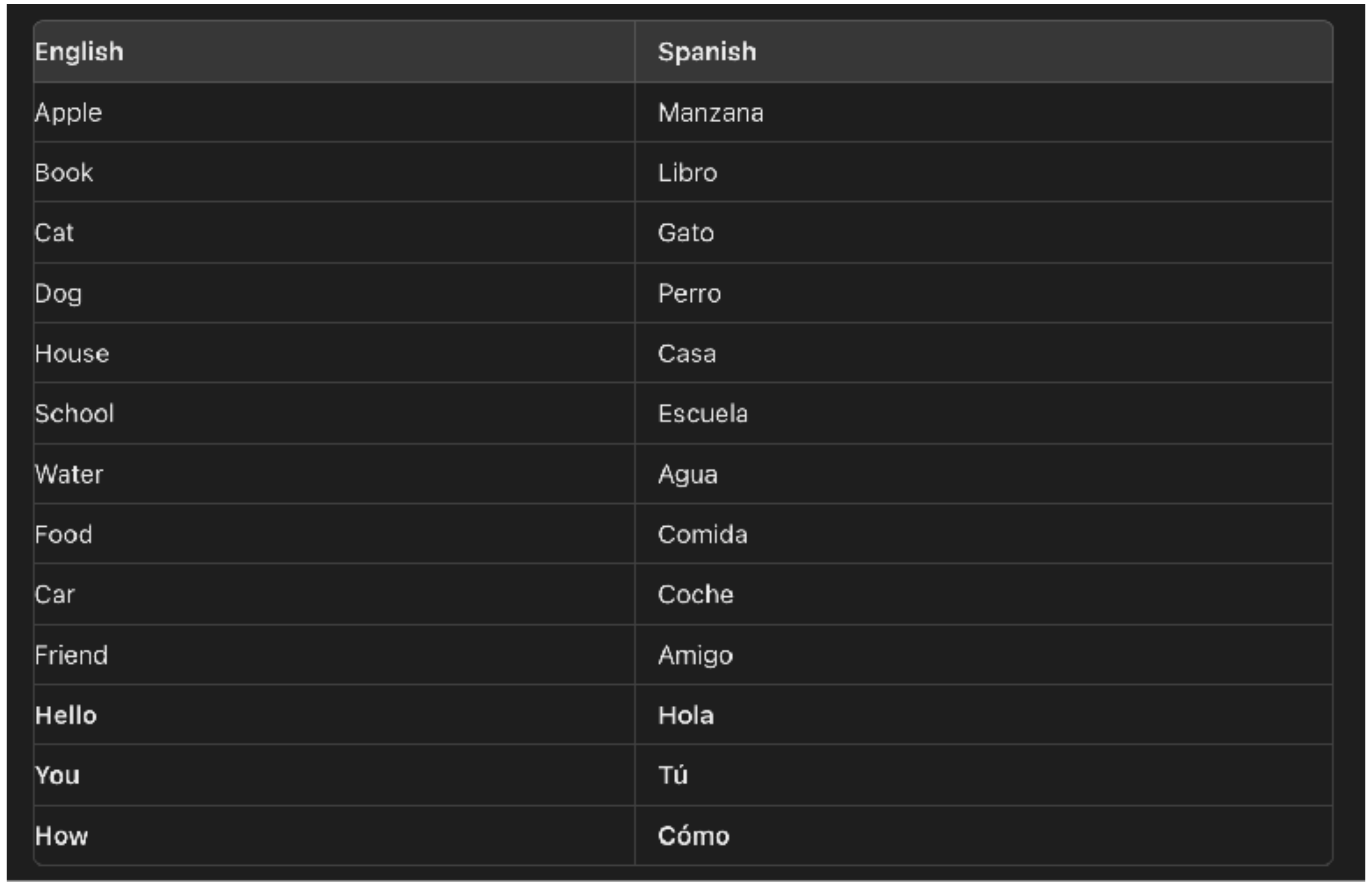
Introduction to Language Translation

- Language translation converts text from one language to another.
- Machine learning models can perform translations automatically.
- These models understand and generate language, just like humans!

Language Translation

English → Spanish

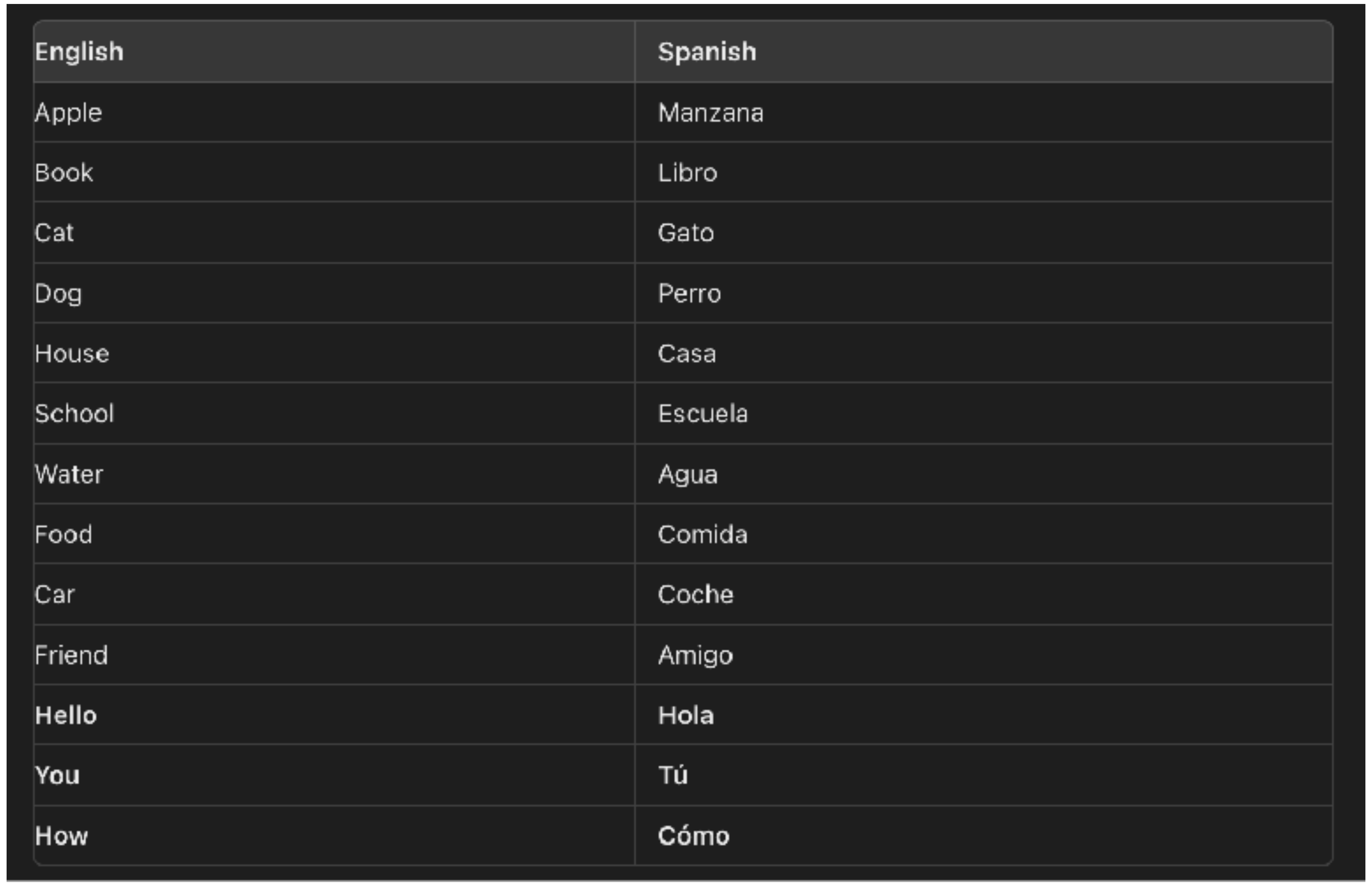
English	Spanish
Hello, how are you?	??????????



English	Spanish
Apple	Manzana
Book	Libro
Cat	Gato
Dog	Perro
House	Casa
School	Escuela
Water	Agua
Food	Comida
Car	Coche
Friend	Amigo
Hello	**Hola**
You	**Tú**
How	**Cómo**

Language Translation

English → Spanish



English	Spanish
Apple	Manzana
Book	Libro
Cat	Gato
Dog	Perro
House	Casa
School	Escuela
Water	Agua
Food	Comida
Car	Coche
Friend	Amigo
Hello	**Hola**
You	**Tú**
How	**Cómo**

English	**Spanish**
Hello, how are you?	Hola, ¿Cómo estás?

Hello → Hola

how → Como

are → estás

you → ??

Language Translation

Components

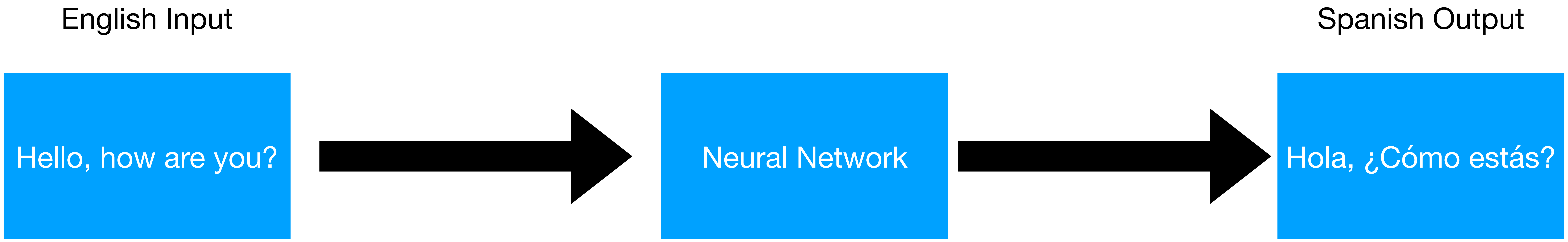
- **Tokens**: Smallest units of language for processing.
- **Grammer**: Defines the ordering of tokens in a language.

Tokens: Hello, how are you?

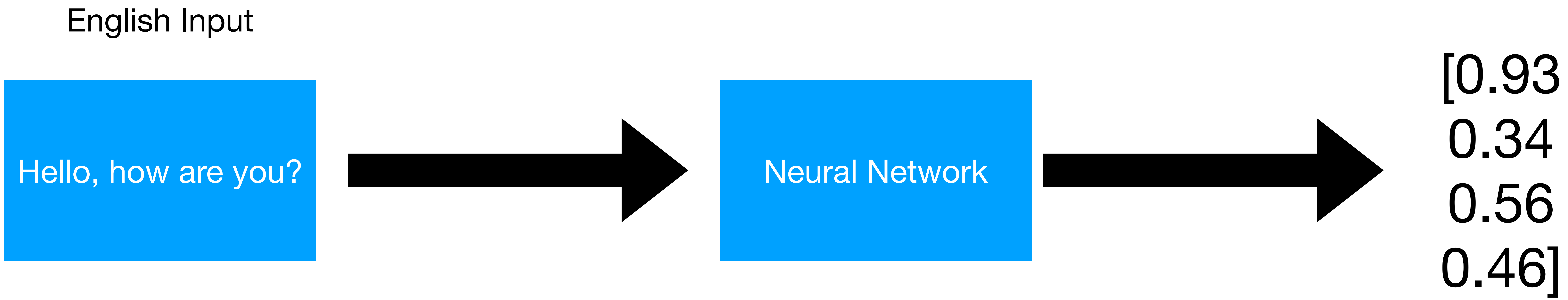
[“Hello”, “how”, “are”, “you”, “?”]

Grammer: Adjectives, Nouns, Verbs

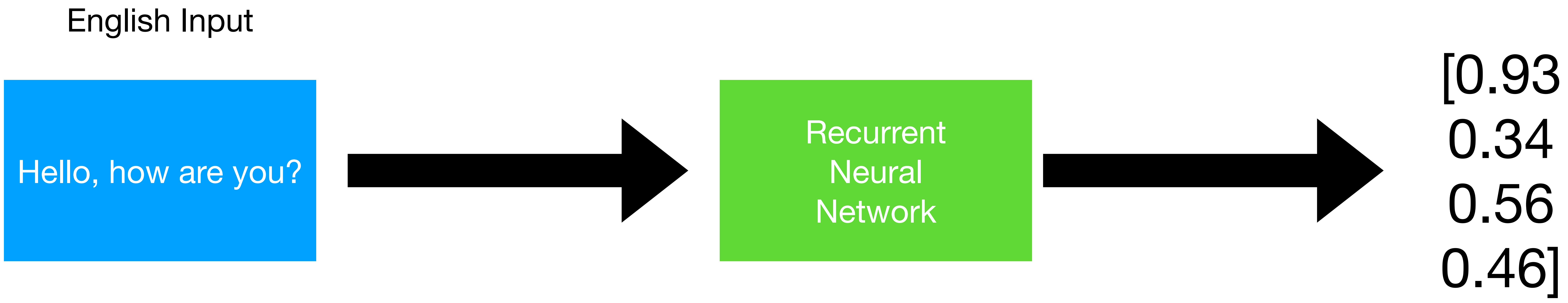
Language Translation



Language Translation



Language Translation



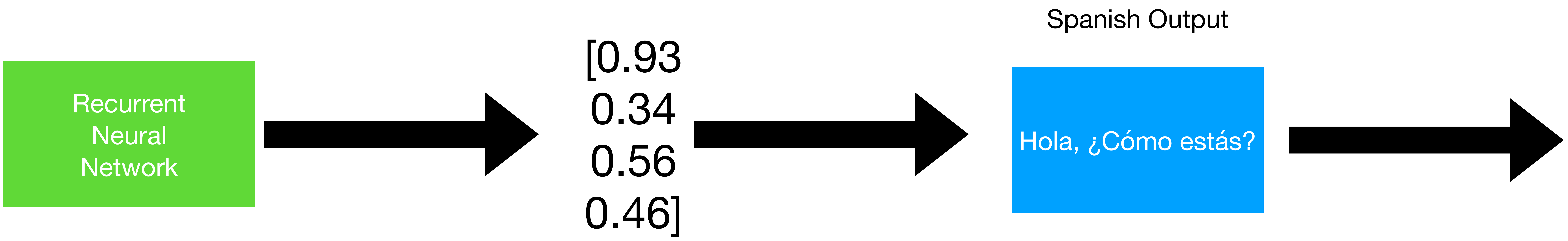
English Input

Hello, how are you?

Recurrent
Neural
Network

[0.93
0.34
0.56
0.46]

Language Translation



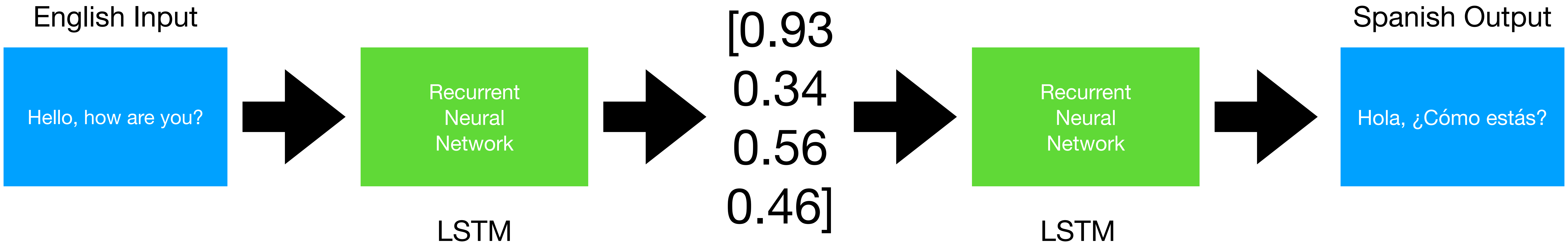
Recurrent Neural Network

[0.93 0.34 0.56 0.46]

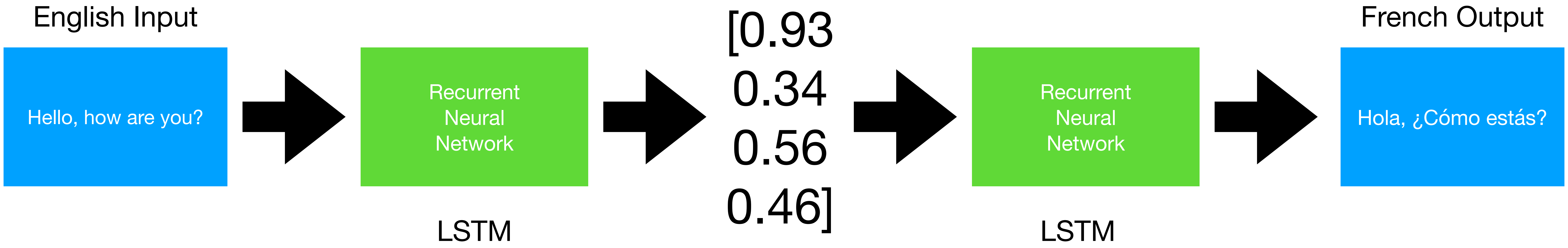
Spanish Output

Hola, ¿Cómo estás?

Language Translation



Language Translation



English Input

Hello, how are you?

Recurrent Neural Network

LSTM

[0.93 0.34 0.56 0.46]

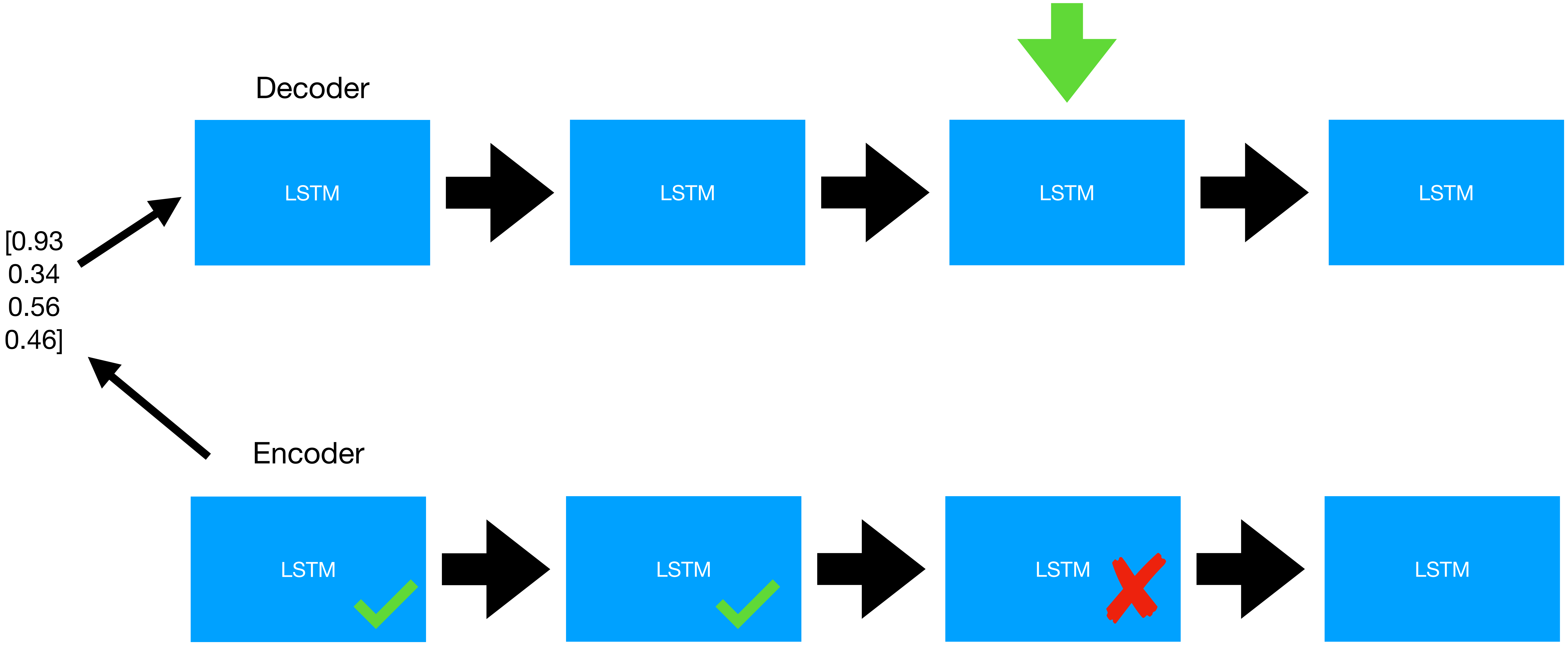
Recurrent Neural Network

LSTM

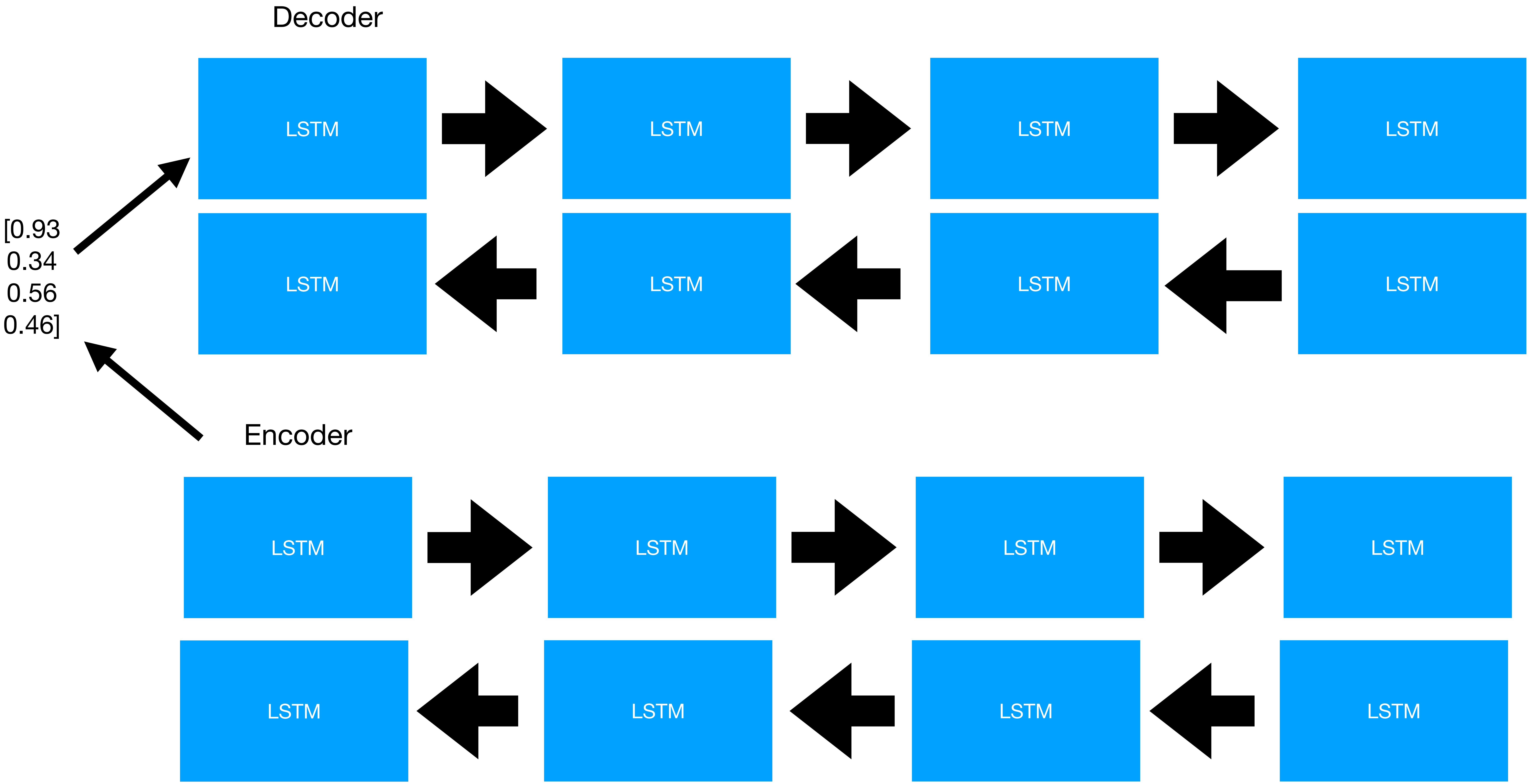
French Output

Hola, ¿Cómo estás?

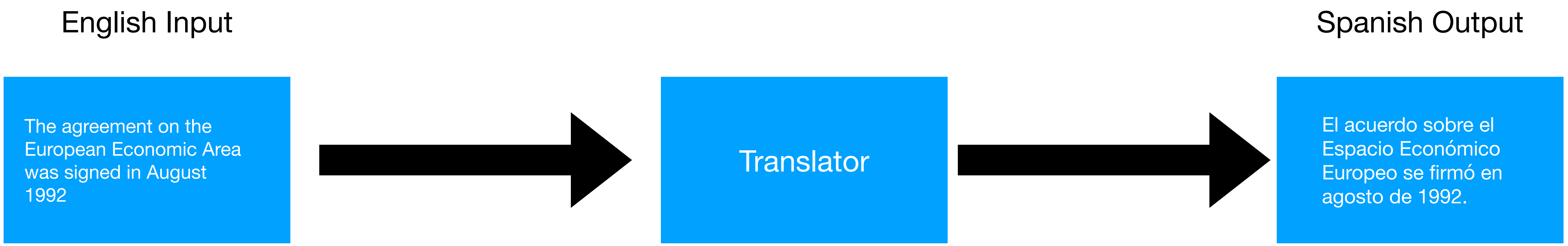
Language Translation



Language Translation



Language Translation



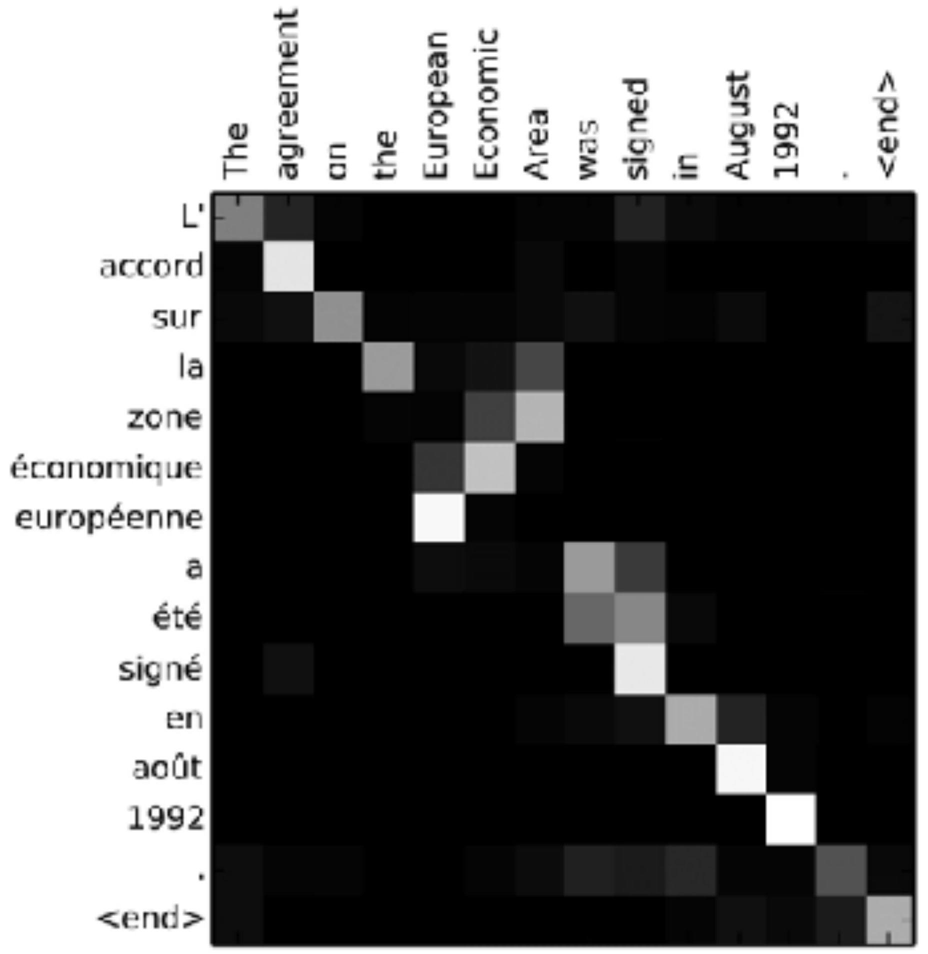
English Input

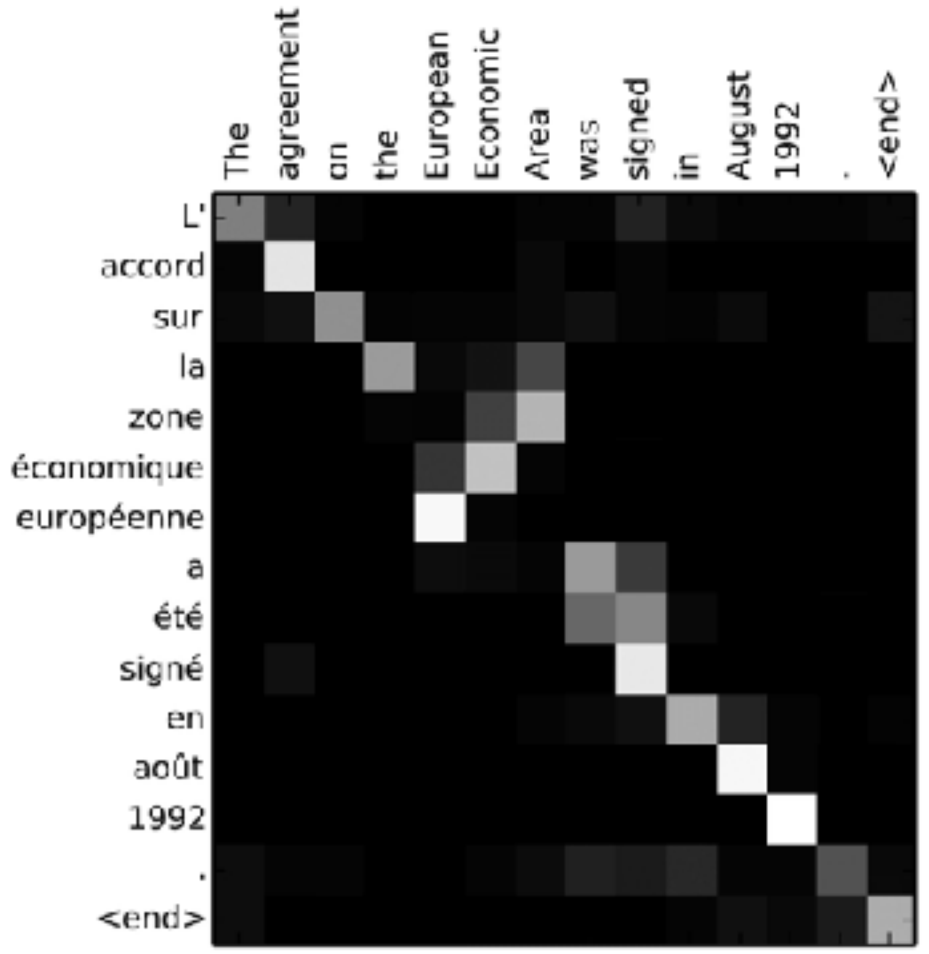
The agreement on the European Economic Area was signed in August 1992

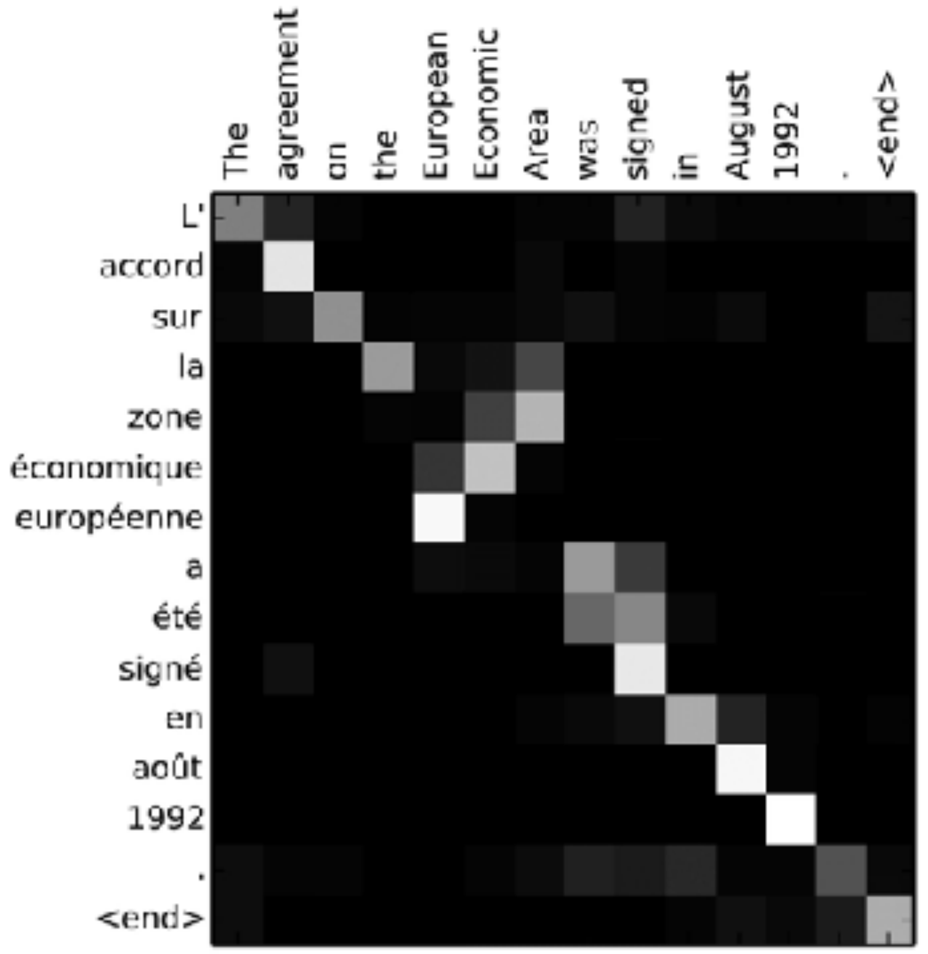
Translator

Spanish Output

El acuerdo sobre el Espacio Económico Europeo se firmó en agosto de 1992.





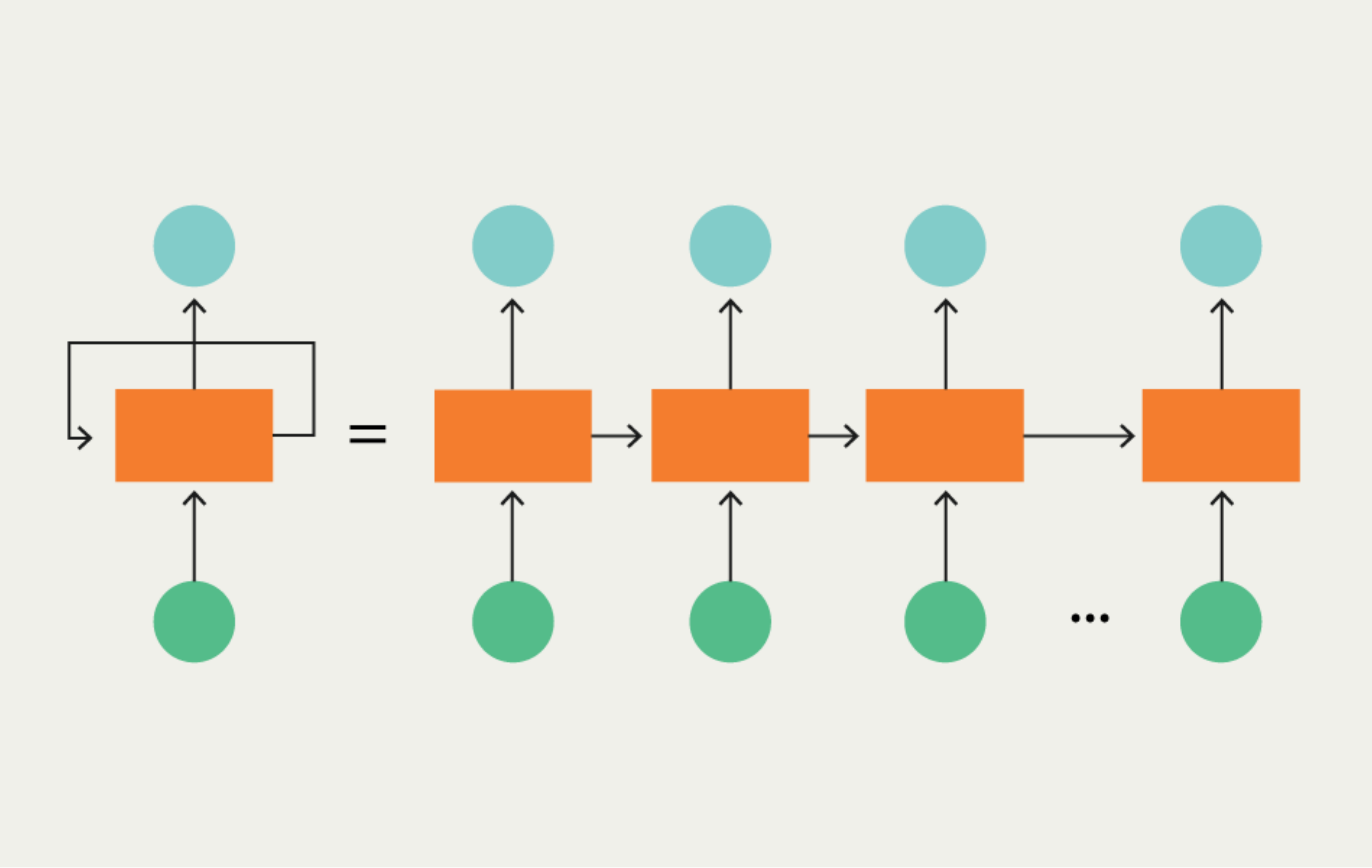


Behind the Scenes

Recurrent Neural Network

RNNs are designed to take sequences of text as inputs or return sequences of text as outputs, or both.

Recurrent Neural Network



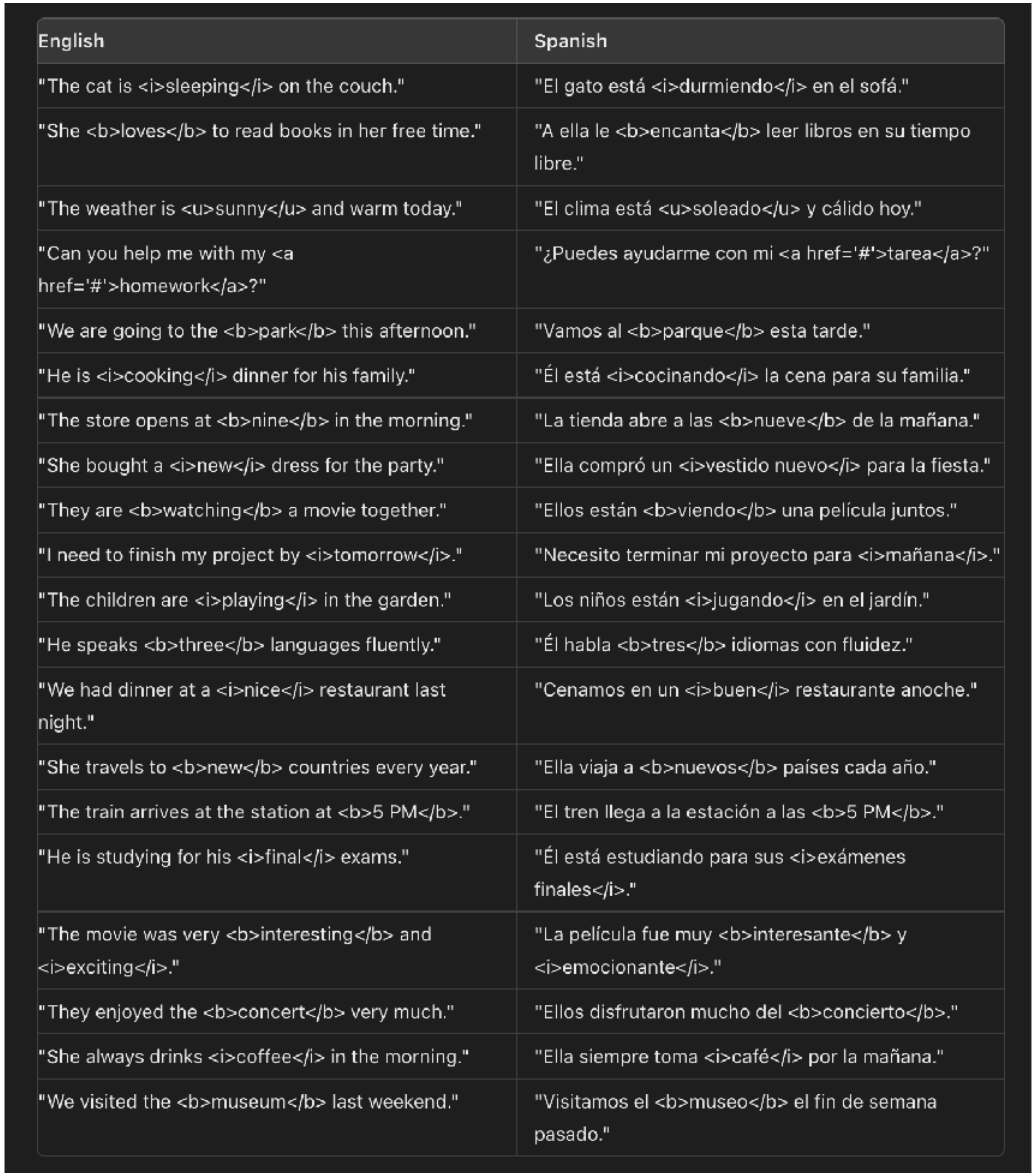
https://www.telusinternational.com/insights/ai-data/article/difference-between-cnn-and-rnn

Building the Pipeline

- **Preprocessing:** load and examine data, cleaning, tokenization, padding
- **Modeling:** build, train, and test the model
- **Prediction**: generate specific translations of English to French, and compare the output translations to the ground truth translations
- **Iteration**: iterate on the model, experimenting with different architecture

Preprocessing

Load & Examine Data



English	Spanish
"The cat is <i>sleeping</i> on the couch."	"El gato está <i>durmiendo</i> en el sofá."
"She <b>loves</b> to read books in her free time."	"A ella le <b>encanta</b> leer libros en su tiempo libre."
"The weather is <u>sunny</u> and warm today."	"El clima está <u>soleado</u> y cálido hoy."
"Can you help me with my <a href='#'>homework</a>?"	"¿Puedes ayudarme con mi <a href='#'>tarea</a>?"
"We are going to the <b>park</b> this afternoon."	"Vamos al <b>parque</b> esta tarde."
"He is <i>cooking</i> dinner for his family."	"Él está <i>cocinando</i> la cena para su familia."
"The store opens at <b>nine</b> in the morning."	"La tienda abre a las <b>nueve</b> de la mañana."
"She bought a <i>new</i> dress for the party."	"Ella compró un <i>vestido nuevo</i> para la fiesta."
"They are <b>watching</b> a movie together."	"Ellos están <b>viendo</b> una película juntos."
"I need to finish my project by <i>tomorrow</i>."	"Necesito terminar mi proyecto para <i>mañana</i>."
"The children are <i>playing</i> in the garden."	"Los niños están <i>jugando</i> en el jardín."
"He speaks <b>three</b> languages fluently."	"Él habla <b>tres</b> idiomas con fluidez."
"We had dinner at a <i>nice</i> restaurant last night."	"Cenamos en un <i>buen</i> restaurante anoche."
"She travels to <b>new</b> countries every year."	"Ella viaja a <b>nuevos</b> países cada año."
"The train arrives at the station at <b>5 PM</b>."	"El tren llega a la estación a las <b>5 PM</b>."
"He is studying for his <i>final</i> exams."	"Él está estudiando para sus <i>exámenes finales</i>."
"The movie was very <b>interesting</b> and <i>exciting</i>."	"La película fue muy <b>interesante</b> y <i>emocionante</i>."
"They enjoyed the <b>concert</b> very much."	"Ellos disfrutaron mucho del <b>concierto</b>."
"She always drinks <i>coffee</i> in the morning."	"Ella siempre toma <i>café</i> por la mañana."
"We visited the <b>museum</b> last weekend."	"Visitamos el <b>museo</b> el fin de semana pasado."

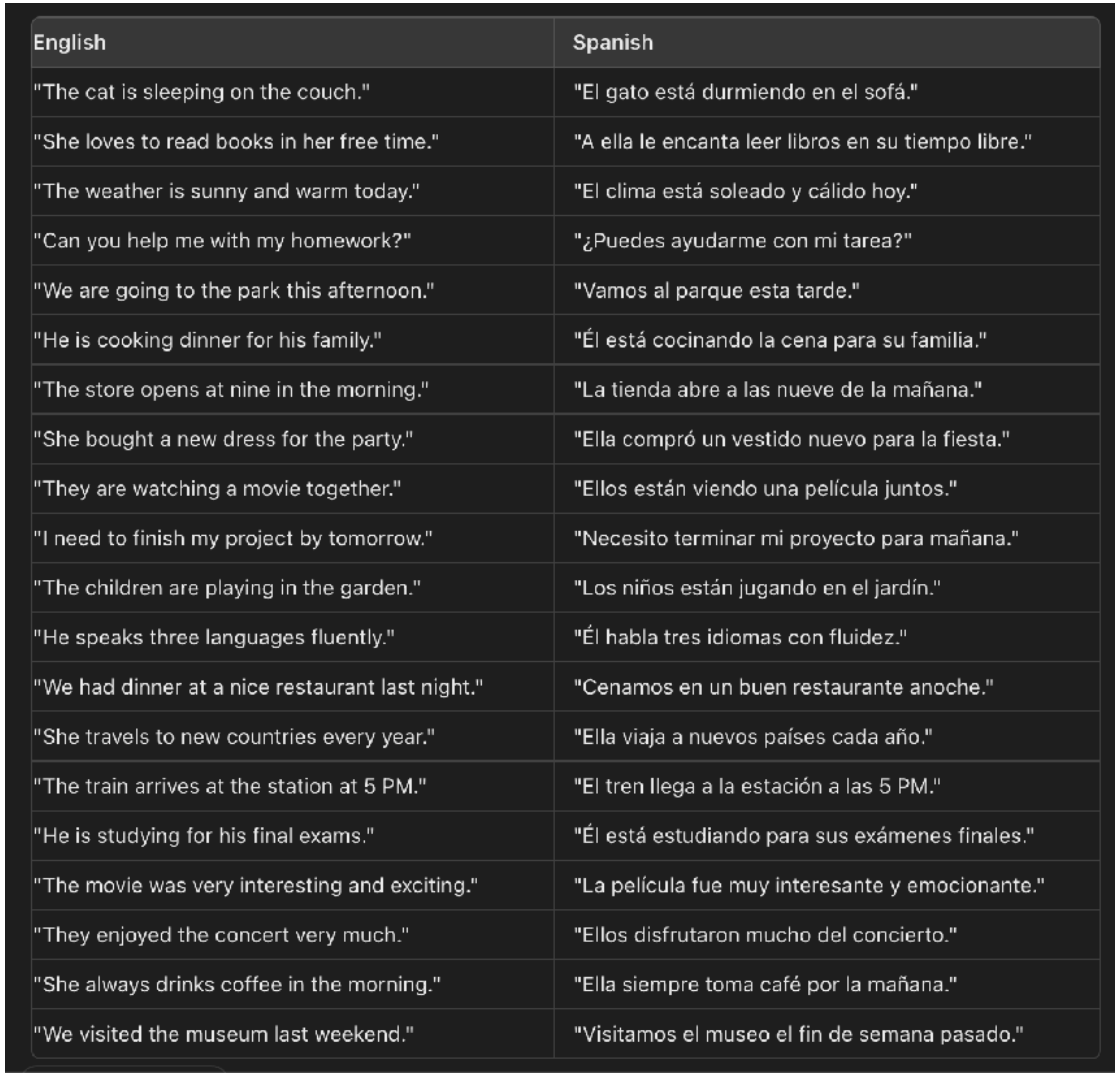
Preprocessing

Cleaning

Remove HTML Tags

Remove Stop Words

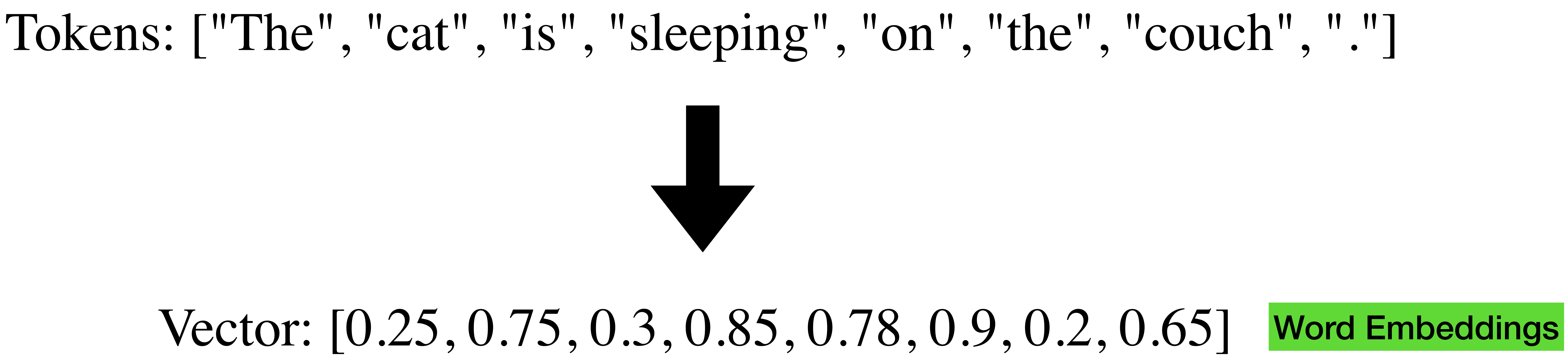
Remove Punctuation



English	Spanish
"The cat is sleeping on the couch."	"El gato está durmiendo en el sofá."
"She loves to read books in her free time."	"A ella le encanta leer libros en su tiempo libre."
"The weather is sunny and warm today."	"El clima está soleado y cálido hoy."
"Can you help me with my homework?"	"¿Puedes ayudarme con mi tarea?"
"We are going to the park this afternoon."	"Vamos al parque esta tarde."
"He is cooking dinner for his family."	"Él está cocinando la cena para su familia."
"The store opens at nine in the morning."	"La tienda abre a las nueve de la mañana."
"She bought a new dress for the party."	"Ella compró un vestido nuevo para la fiesta."
"They are watching a movie together."	"Ellos están viendo una película juntos."
"I need to finish my project by tomorrow."	"Necesito terminar mi proyecto para mañana."
"The children are playing in the garden."	"Los niños están jugando en el jardín."
"He speaks three languages fluently."	"Él habla tres idiomas con fluidez."
"We had dinner at a nice restaurant last night."	"Cenamos en un buen restaurante anoche."
"She travels to new countries every year."	"Ella viaja a nuevos países cada año."
"The train arrives at the station at 5 PM."	"El tren llega a la estación a las 5 PM."
"He is studying for his final exams."	"Él está estudiando para sus exámenes finales."
"The movie was very interesting and exciting."	"La película fue muy interesante y emocionante."
"They enjoyed the concert very much."	"Ellos disfrutaron mucho del concierto."
"She always drinks coffee in the morning."	"Ella siempre toma café por la mañana."
"We visited the museum last weekend."	"Visitamos el museo el fin de semana pasado."

Preprocessing

Tokenization (Text to Number)



Tokens: ["The", "cat", "is", "sleeping", "on", "the", "couch", "."]

Vector: $[0.25, 0.75, 0.3, 0.85, 0.78, 0.9, 0.2, 0.65]$

Word Embeddings

Preprocessing

Padding

The cat is sleeping on the couch

I like to travel

Preprocessing

Padding

The cat is sleeping on the couch

$[0.25, 0.75, 0.3, 0.85, 0.78, 0.9, 0.2, 0.65]$

I like to travel

$[0.45, 0.15, 0.33]$

Preprocessing

Padding

The cat is sleeping on the couch

$[0.25, 0.75, 0.3, 0.85, 0.78, 0.9, 0.2, 0.65]$

I like to travel

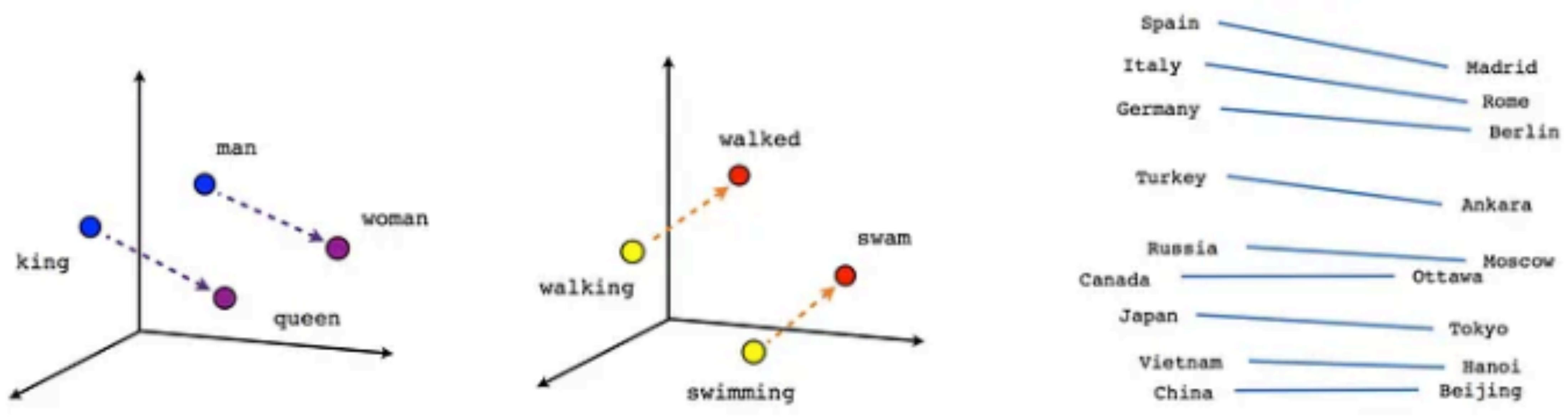
$[0.45, 0.15, 0.33, 0, 0, 0, 0, 0]$ Padding

Modeling

- **Inputs.** Input sequences are fed into the model with one word for every time step. Each word is encoded as a unique integer or one-hot encoded vector that maps to the English dataset vocabulary.
- **Embedding Layers.** Embeddings are used to convert each word to a vector. The size of the vector depends on the complexity of the vocabulary.
- **Recurrent Layers (Encoder).** This is where the context from word vectors in previous time steps is applied to the current word vector.
- **Dense Layers (Decoder).** These are typical fully connected layers used to decode the encoded input into the correct translation sequence.
- **Outputs.** The outputs are returned as a sequence of integers or one-hot encoded vectors which can then be mapped to the French dataset vocabulary.

Embeddings

- Technique to capture syntactic and semantic relationships between words.
- Projects words into an n-dimensional space.
- Similar words are positioned close to each other.



Male-Female

Verb tense

Country-Capital

Photo credit: Chris Bail

Benefits of Word Embeddings

- Understand word similarities and differences.
- Identify relationships like gender, verb tense, and geopolitical context.
- Enhances natural language processing tasks.

Pre-trained Embeddings

- Require large datasets and extensive computation.
- Commonly used packages: GloVe, word2vec.
- Saves time and resources, providing robust, ready-to-use models.

Encoder & Decoder

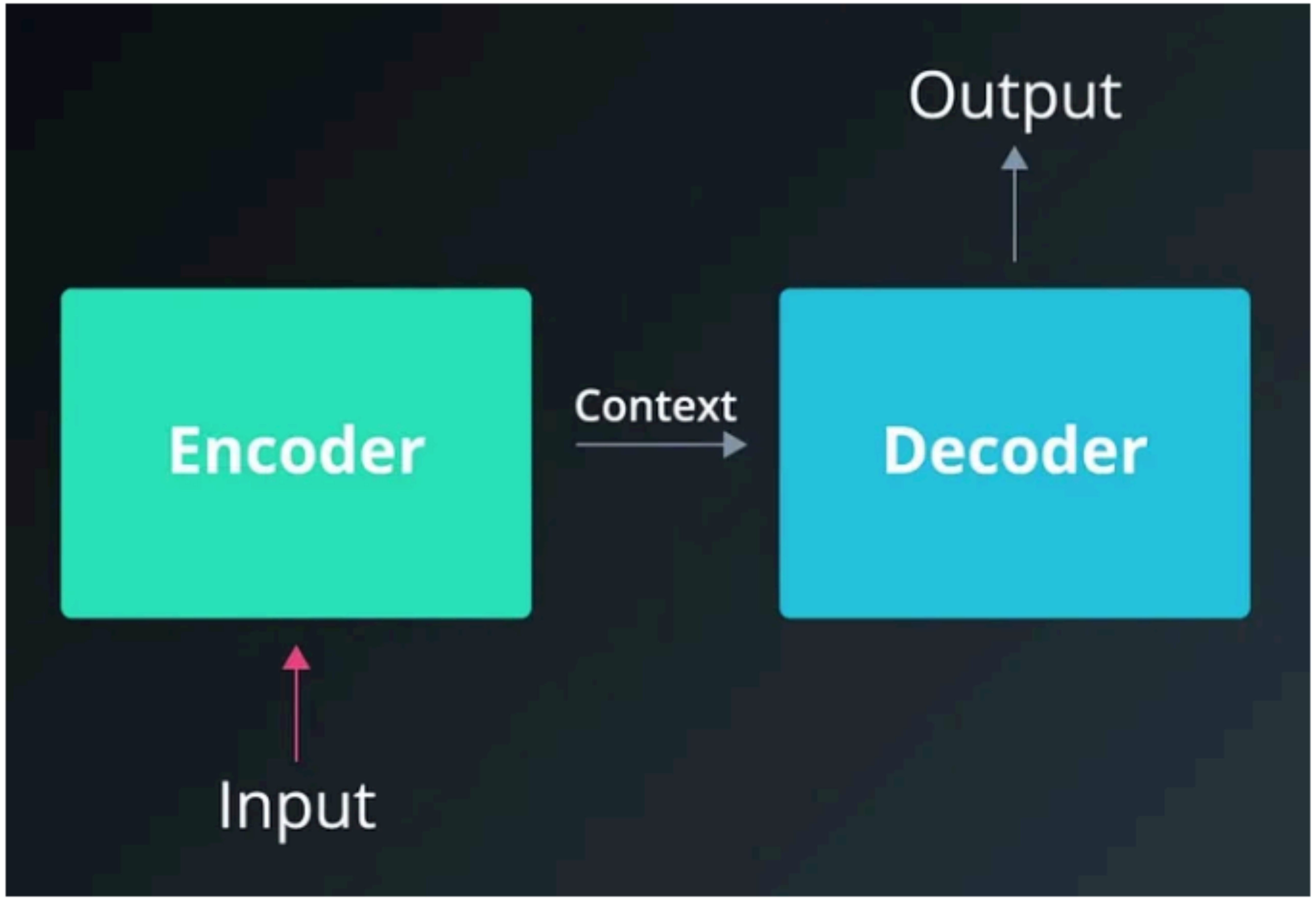


Image credit: Udacity

How Does the Encoder Work?

- The encoder reads the sentence word by word.
- It remembers important information (context) as it reads.
- This information is stored in something called a hidden state.

Time Steps in the Encoder

- The encoder processes one word at a time.
- Each word helps update the hidden state.
- After reading all words, the hidden state has all the information needed.

The Decoder’s Job

- The decoder uses the hidden state to start translating.
- It produces the translation word by word.
- Each word it generates is based on the hidden state and the previous word.

Why Use a Hidden State?

- The hidden state holds the “memory” of the sentence.
- The bigger the hidden state, the more information it can store.
- More memory means better translations, but it also needs more computing power.

Word Representation

- When we say “word,” we actually mean its vector.
- Vectors are numbers that represent the word’s meaning.
- These vectors come from something called an embedding layer.

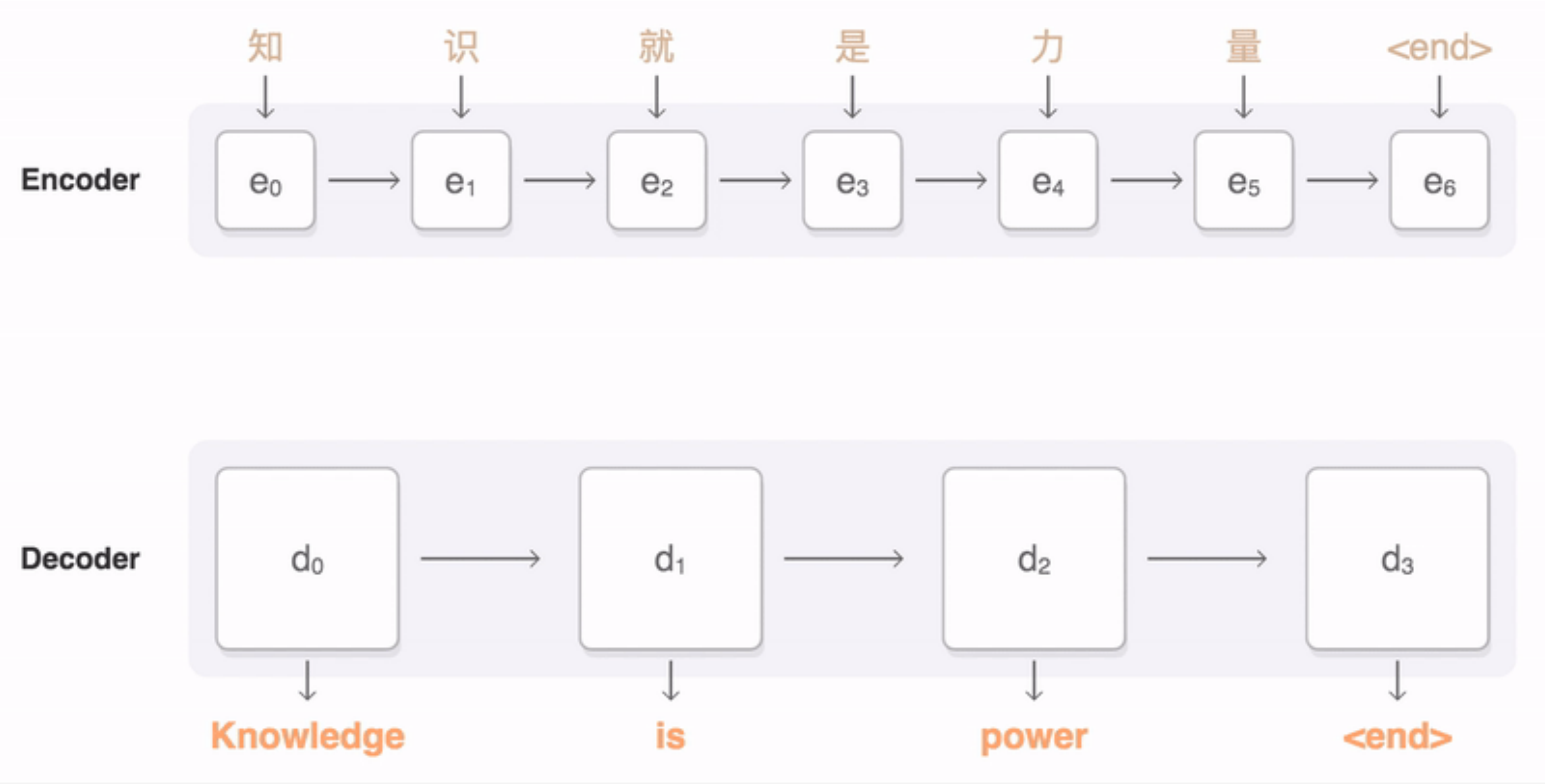


Image credit: xiandong79.github.io

Bidirectional Layer

- Traditional RNNs process information in one direction.
- Bidirectional layers process information in both forward and backward directions.
- This helps the model understand the entire context of the input.

Why Use Bidirectional Layers?

- Provides additional context by looking at future words.
- Can enhance model performance, especially in tasks requiring full context.
- Example: Understanding sentences better, like when interpreting Yoda's speech.

How Bidirectional Layers Work

- One RNN processes the sequence from start to end.
- Another RNN processes the reversed sequence from end to start.
- The outputs from both RNNs are combined for a richer context.

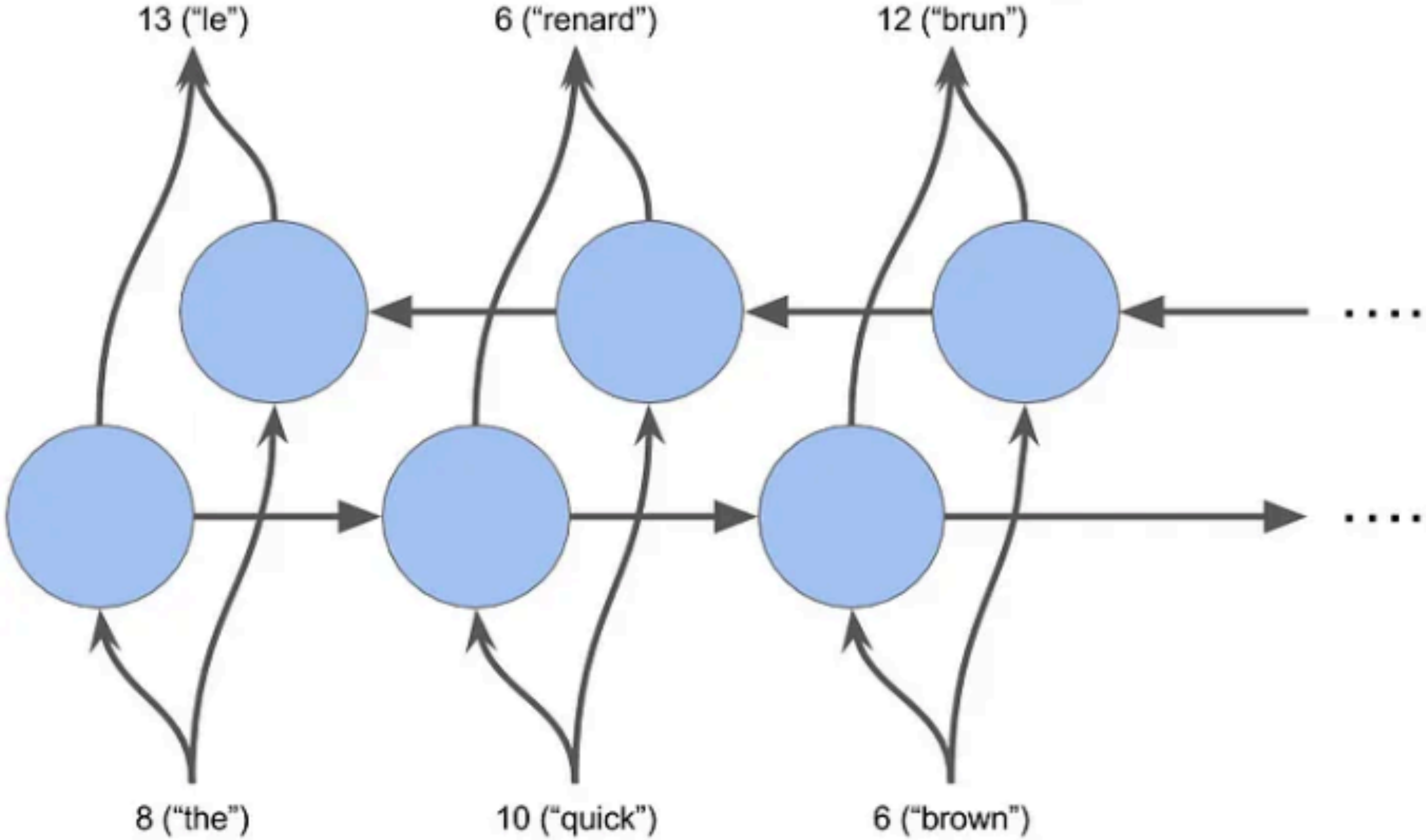


Image credit: Udacity

Benefits of Bidirectional Layers

- Better understanding of context and meaning.
- Improved performance in tasks like text classification and translation.
- Useful in any task where future context is important.