Introduction to Machine Translation

SIU SAI CHEONG
SCHOOL OF TRANSLATION
HANG SENG MANAGEMENT COLLEGE

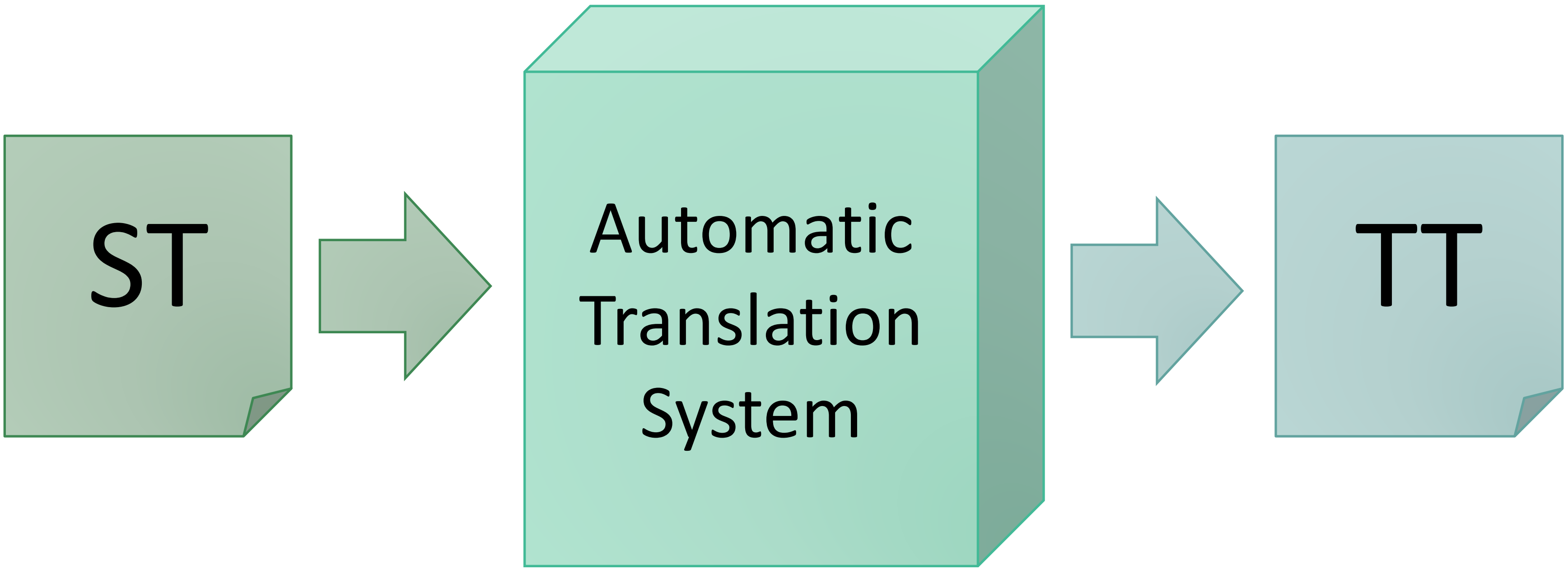
Outline

- Machine translation (MT) and translation technology
- Rule-based Machine Translation (RBMT)
- Example-based Machine Translation (EBMT)
- Statistical Machine Translation (SMT)
- Neural Machine Translation (NMT)
- MT Skills

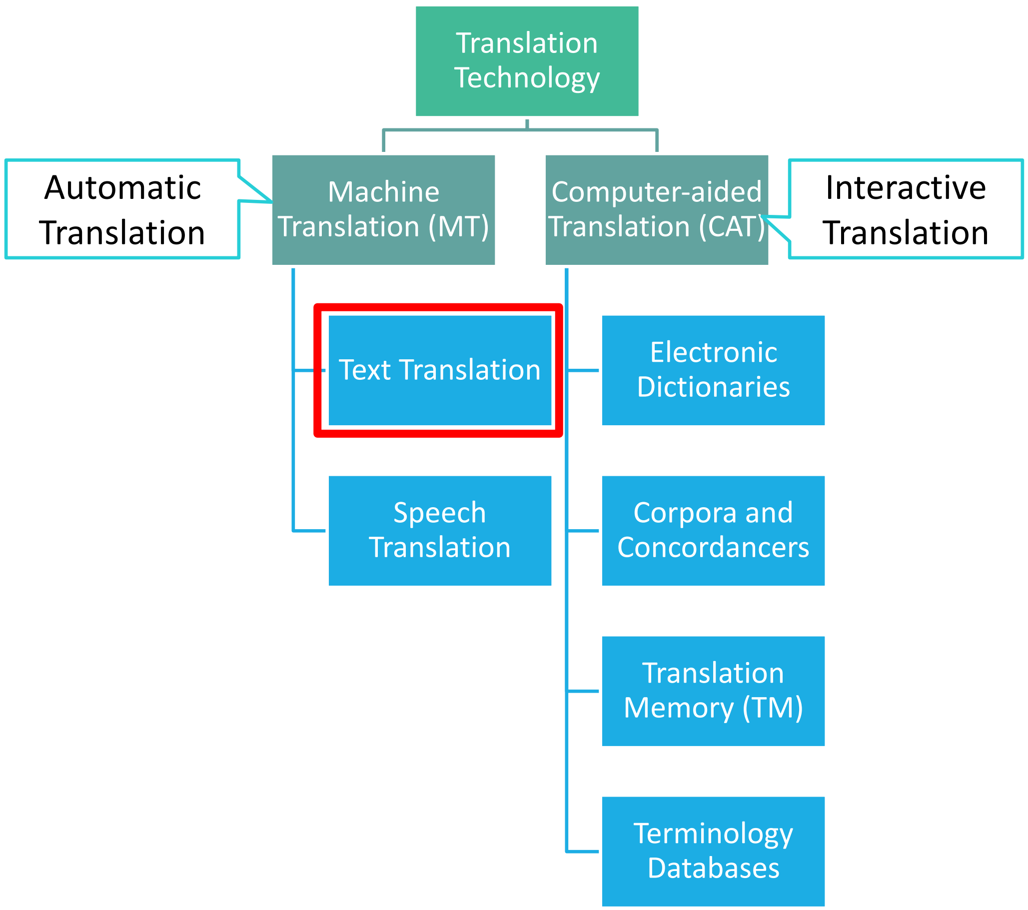
Topics

- Machine translation (MT) and translation technology
- Rule-based Machine Translation (RBMT)
- Example-based Machine Translation (EBMT)
- Statistical Machine Translation (SMT)
- Neural Machine Translation (NMT)
- MT Skills

Machine Translation (MT)



Translation Technology



Google Translate and Instant Camera Translation
https://www.youtube.com/watch?v=06olHmcJjS0

iTranslate Voice
http://vimeo.com/86562665

ili
https://www.youtube.com/watch?v=rIiGyn_Hfcl

Topics

- Machine translation (MT) and translation technology
- Rule-based Machine Translation (RBMT)
- Example-based Machine Translation (EBMT)
- Statistical Machine Translation (SMT)
- Neural Machine Translation (NMT)
- MT Skills

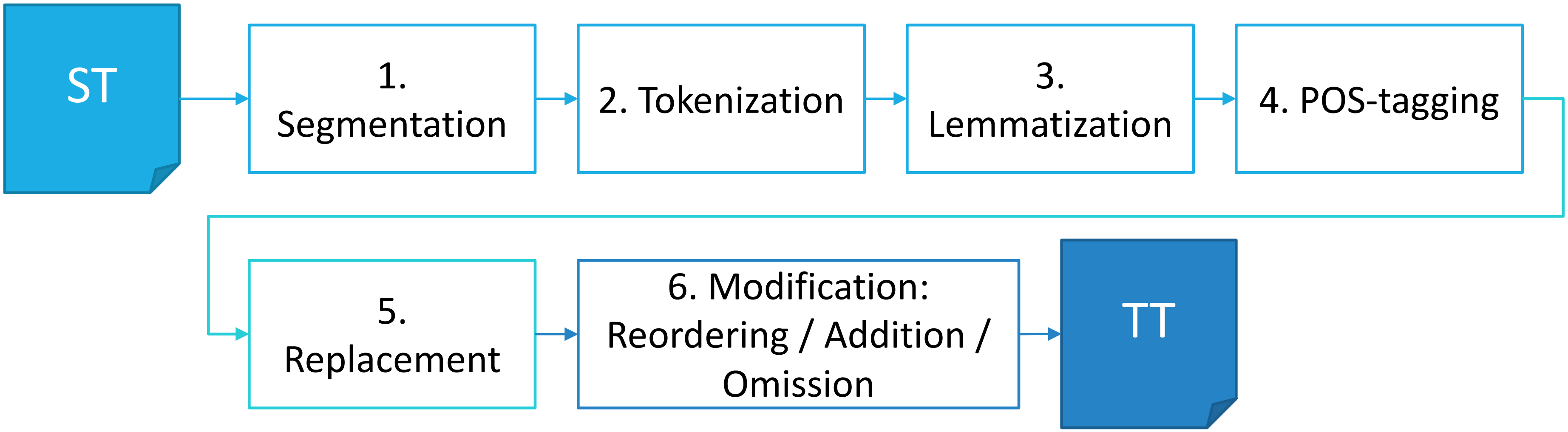
Rule-based Machine Translation (RBMT)

- Direct MT
- Transfer MT

Direct MT

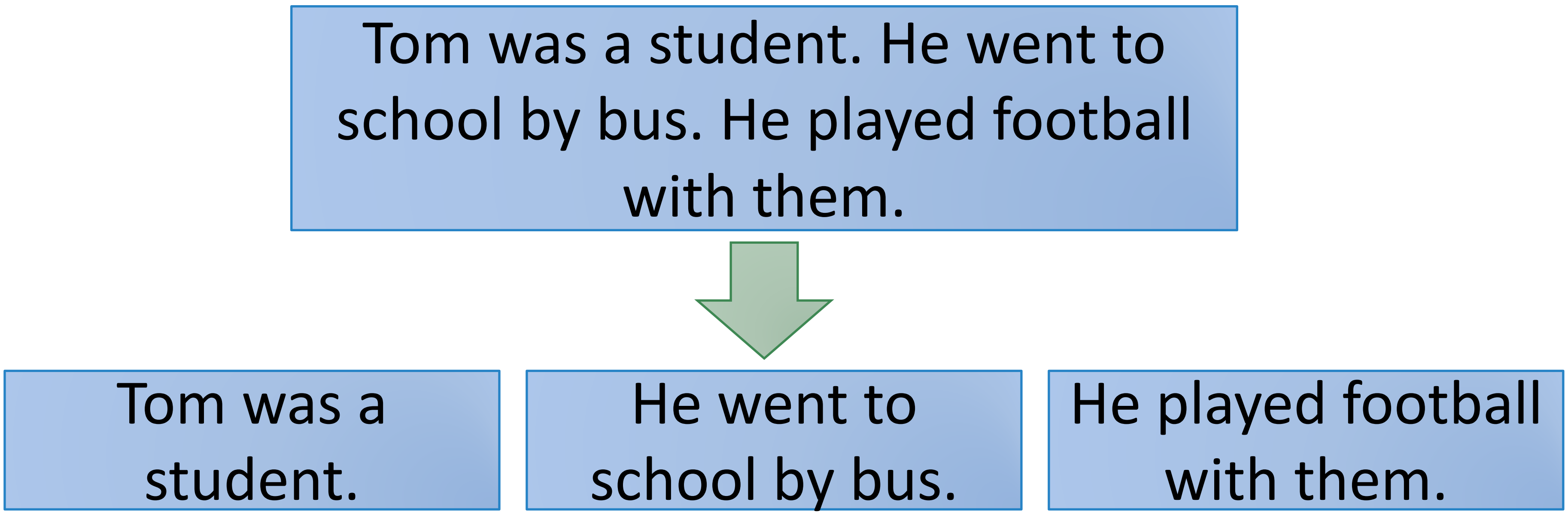
- This is considered to be the first generation of MT systems.
- The ST is treated as a string of words. The SL words are then replaced by TL words, and the TL words are reordered.

Direct MT


ST
1. Segmentation
2. Tokenization
3. Lemmatization
4. POS-tagging
5. Replacement
6. Modification: Reordering / Addition / Omission
TT

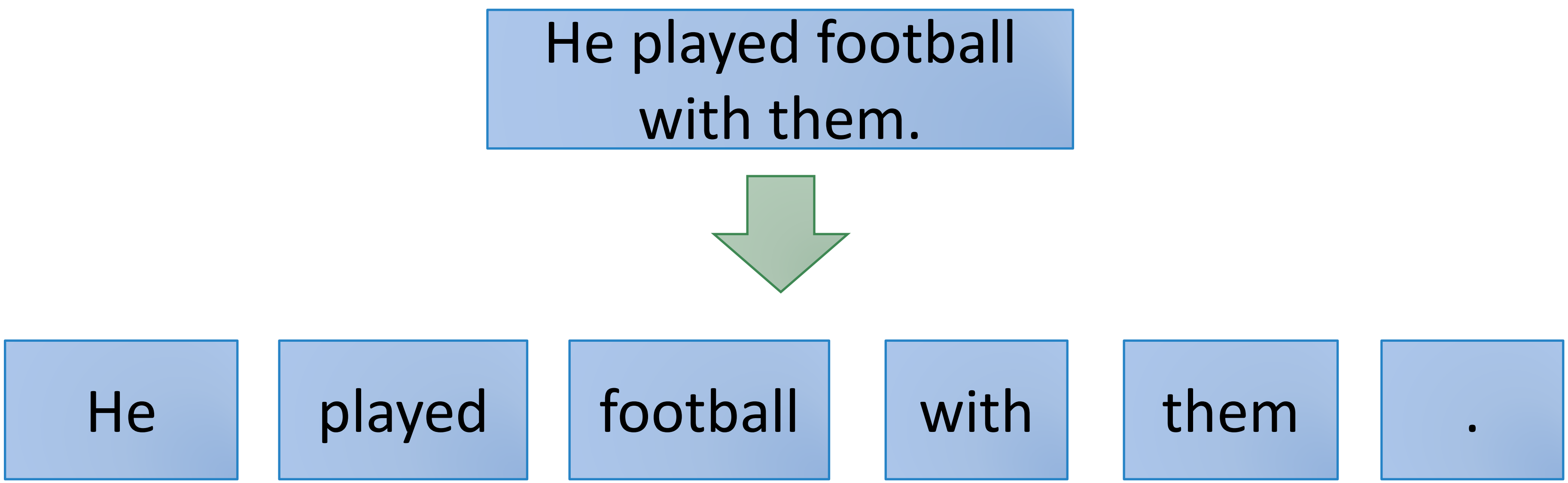
Segmentation

- the process of splitting a text into sentences



Tokenization

- the process of dividing a text into words



Lemmatization

- the process of grouping the identical, related or inflected forms of a word together as a single item (lemma)
- Example: do, does, doing, did → do

POS-tagging

- the process of assigning a part-of-speech (POS) to every lexical item in the input

He | play(ed) | football | with | them

Pronoun | Preposition | Noun | Verb

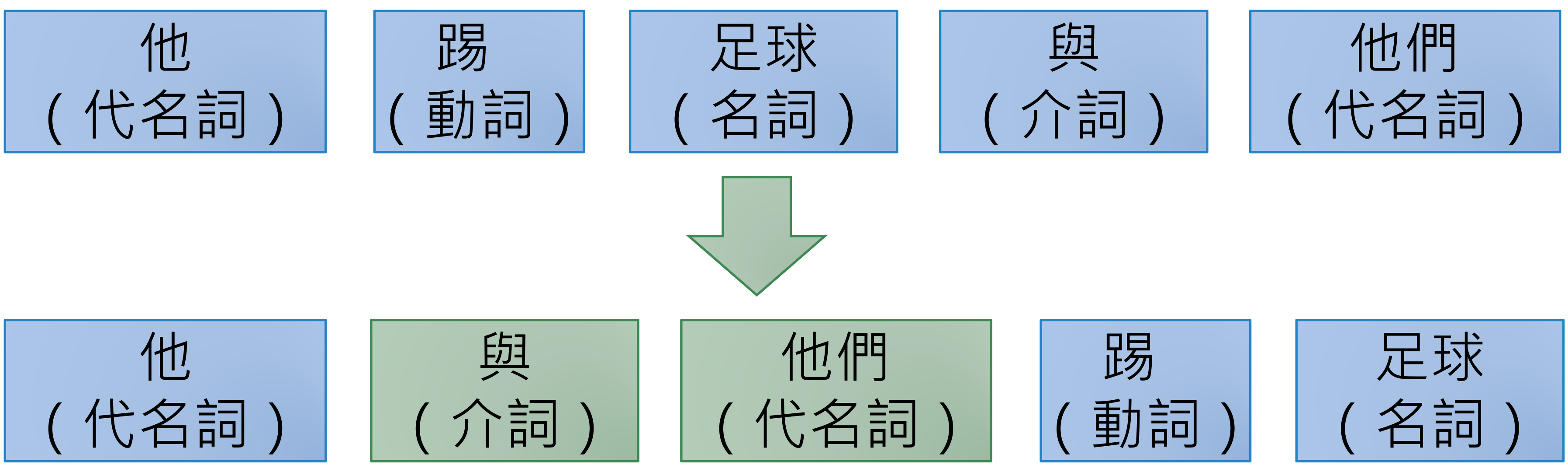
Replacement

- the process of replacing the SL words with TL words

He (pronoun)	play(ed) (verb)	football (noun)	with (prep)	them (pronoun)
他（代名詞）	踢（動詞）	足球（名詞）	與（介詞）	他們（代名詞）

Reordering and Modification

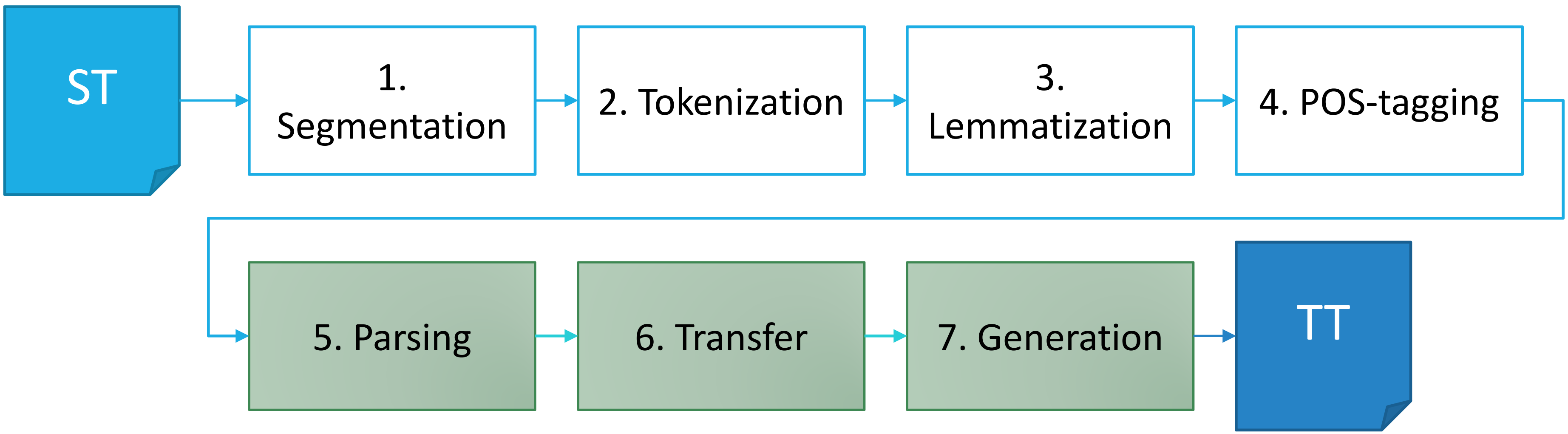
- the process of reordering the TL words and polishing the TT



Transfer MT

- The basic idea: **Analysis, Transfer and Generation**
- Input sentences are analyzed and translated into an abstract internal representation retaining their features in the SL.
- The representation is transferred to a representation with the features in TL and generates the TT.

Transfer MT



ST
1. Segmentation
2. Tokenization
3. Lemmatization
4. POS-tagging
5. Parsing
6. Transfer
7. Generation
TT

Parsing

- the process of analyzing the input and generating the SL representation (e.g., a parse tree) that identifies the function of each word/word group

Transfer

- the conversion of the SL representation into its equivalent TL representation

Generation

- the production of the TL text from the TL representation

An Example

- **ST**: The boy went to Didneyland.
- **Case conversion:** the boy went to didneyland.
- **Segmentation**

```
<1> the boy went to didneyland.
```

An Example

- **Tokenization**

```
<1> the
<2> boy
<3> went
<4> to
<5> didneyland
<6> .
```

An Example

- **Lemmatization**

```
<1> the
<2> boy
<3> go (past tense)
<4> to
<5> didneyland
<6> .
```

An Example

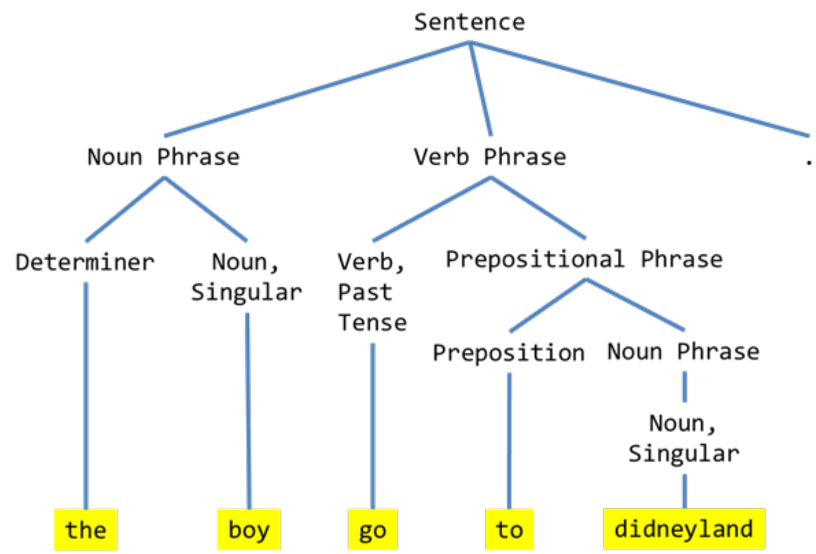
- **POS-tagging**

```
<1> the (definite article)
<2> boy (noun, singular)
<3> go (verb, past tense)
<4> to (preposition)
<5> didneyland (noun, singular)
<6> .
```

An Example

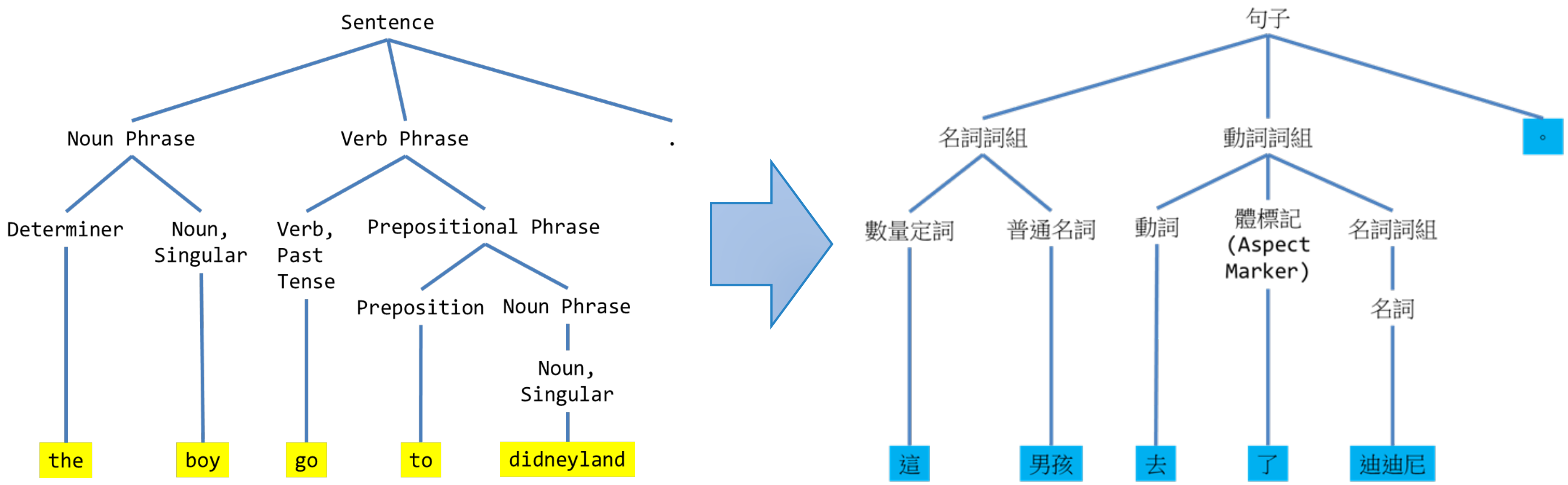
- **Parsing**

<1> the (definite article)
<2> boy (noun, singular)
<3> go (verb, past tense)
<4> to (preposition)
<5> didneyland (noun, singular)
<6> .



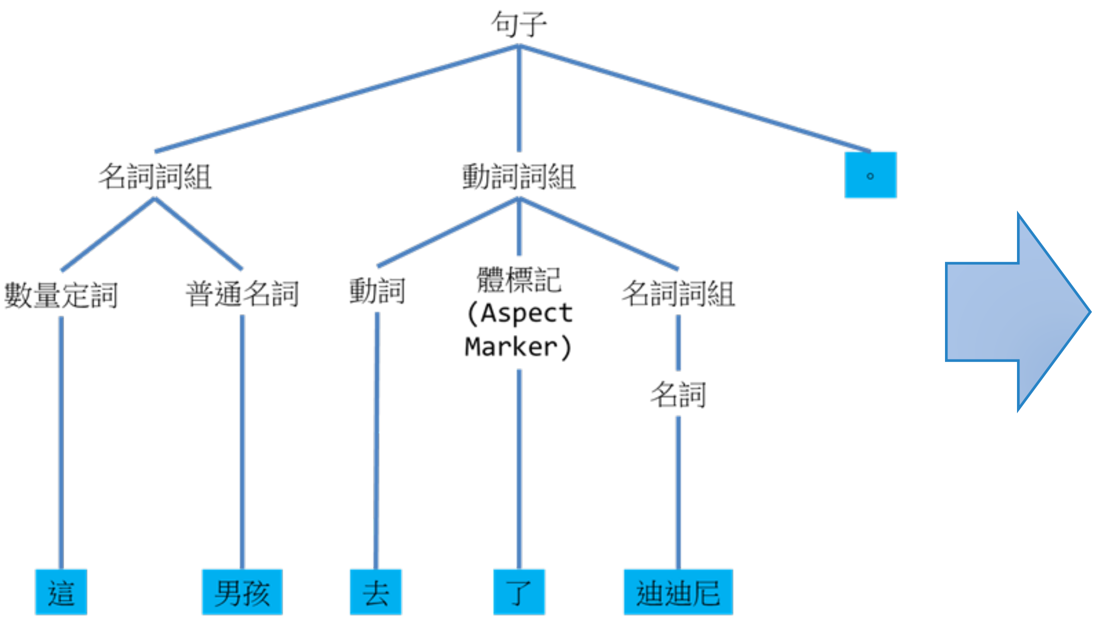
An Example

- **Transfer**



An Example

- **Generation**



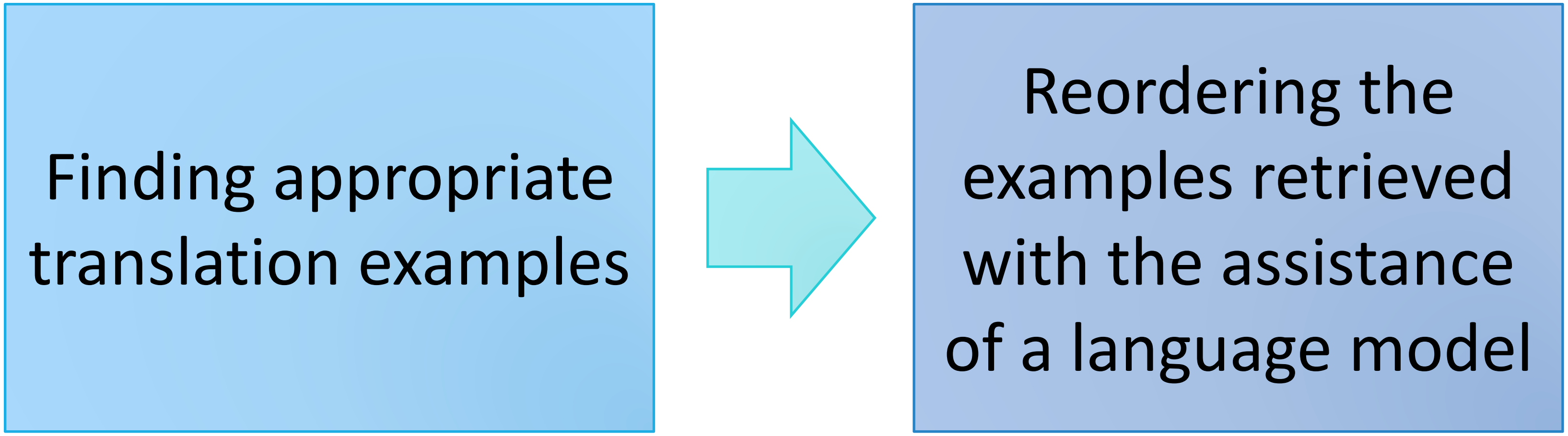
這男孩去了迪迪尼。

Topics

- Machine translation (MT) and translation technology
- Rule-based Machine Translation (RBMT)
- Example-based Machine Translation (EBMT)
- Statistical Machine Translation (SMT)
- Neural Machine Translation (NMT)
- MT Skills

Example-based Machine Translation

- An EBMT system translates by analogy. It requires a bilingual parallel corpus as its main knowledge base.
- To do translation, the system retrieves and modifies the closest translation example(s) from the corpus.


Finding appropriate translation examples
Reordering the examples retrieved with the assistance of a language model

The best example?

ST: The boy went to school.

Step 1: Retrieving the “closest” example

Example	English (Source Language)	Chinese (Target Language)
1	**The boy** is naughty.	那男孩很頑皮。
2	They **went to** the library.	他們到圖書館去。
3	He goes **to school** by train.	他乘火車上學去。
4	They **went to school.**	他們到學校去。

Step 2: Using the example

ST: The boy **went to school.**

Example: They **went to school.**

Useful words: **XXX went to school.**

Translation: The boy 到學校去。

Step 3: Modifying the example

The boy → 那男孩 (from dictionary/corpus)

Draft: 那男孩到學校去。

1. Dictionary/Rules:	2. Other examples in the database:	3. Translation Model:	4. Hybrid Approach:
The → 那 Boy → 男孩	*A. By comparison* The boy is naughty. →那男孩真頑皮。 He didn't see the boy. →他看不見那男孩。 *B. By Word Alignment* The/boy/is/tall. → 那/男孩/長得/高。	P(The|那), P(The|該), P(boy|男孩), P(boy|百厭星)…	1 + 2 + 3

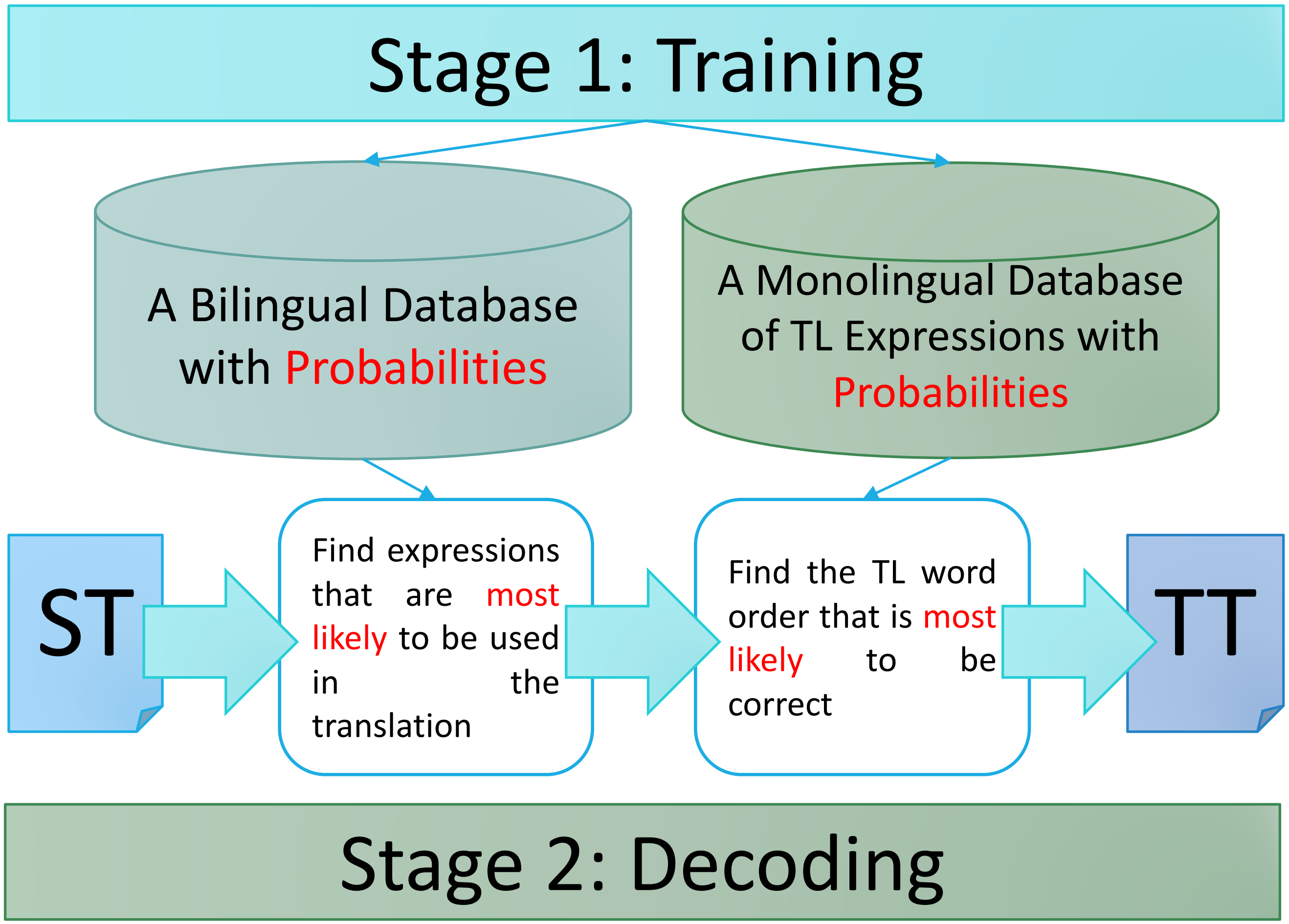
Step 4: Apply the language model and check the readability of the draft.

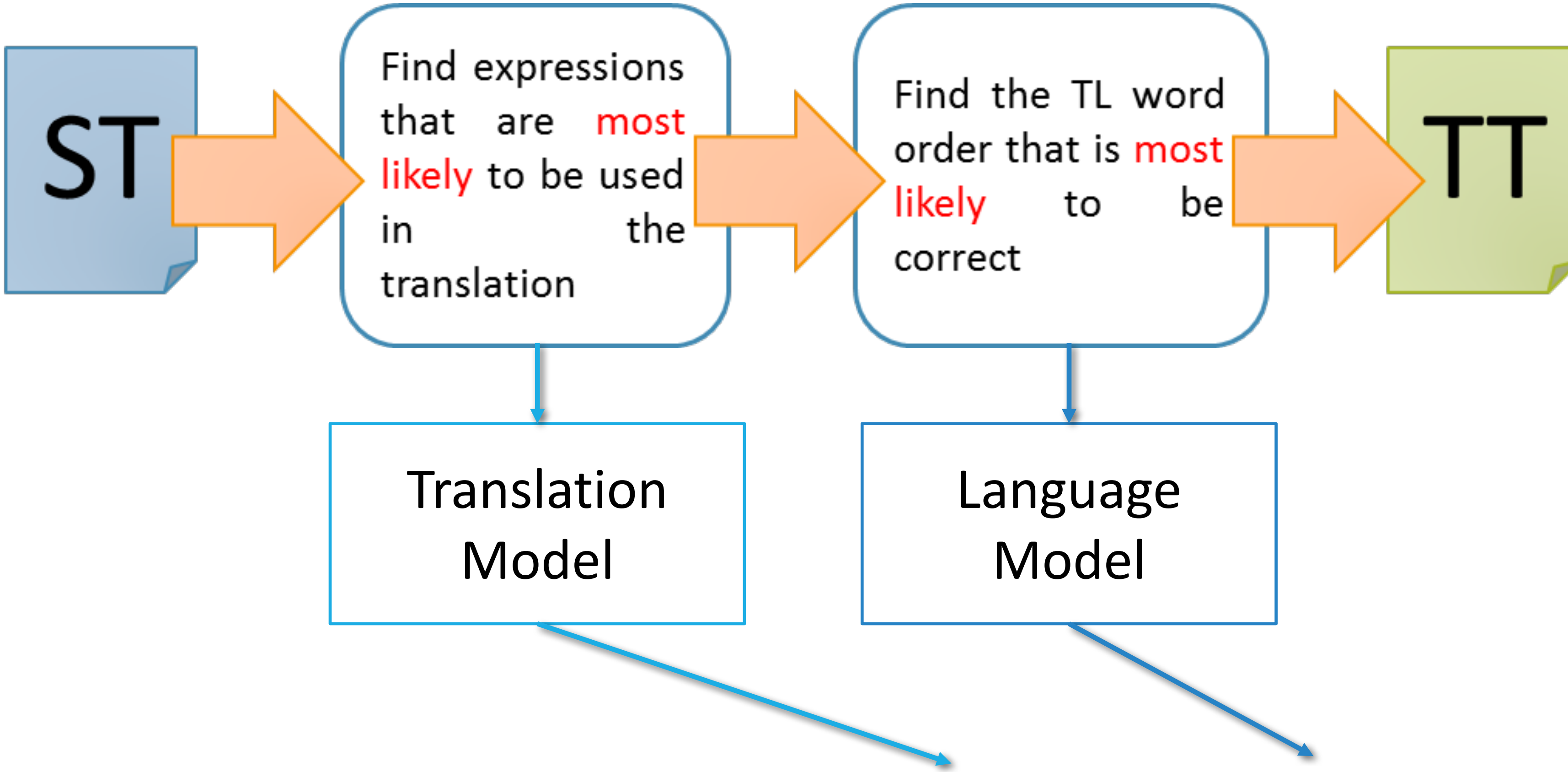
Output: 那男孩到學校去。

Topics

- Machine translation (MT) and translation technology
- Rule-based Machine Translation (RBMT)
- Example-based Machine Translation (EBMT)
- Statistical Machine Translation (SMT)
- Neural Machine Translation (NMT)
- MT Skills

Statistical Machine Translation





$$T' = \arg\max_{T} P(S|T) \times \mathrm{P}(T)$$

Language Model

```
Do you prefer fish ?
Elephants love her .
Give me fish .
Dogs love swimming .
We love dogs .
```

$$\Rightarrow \quad P(T)$$

Monolingual Corpus in the Target Language

Translation Model

Source Language
我喜歡貓。大象鍾意我。人們喜歡狗。他們嗜魚。我們喜歡猴子。你喜歡豬。

Target Language
I love cats. Elephants love me. People love dogs. They prefer fish. We love monkeys. You prefer pigs.

Bilingual Corpus

$$\Rightarrow \quad P(S|T)$$

Monolingual (in TL)

Bilingual

Data

Do you love dogs ?
Elephants loves her .
Give me fish .
They love elephants .
I go swimming .
We love dogs .

我喜歡貓。➔ I love cats.
大象喜歡我。➔ Elephants love me.
他喜歡狗。➔ He loves dogs.
他們喜歡魚。➔ They love fish.
我們喜歡猴子。➔ We love monkeys.
你喜歡豬。➔ You love pigs.

Tables

I	0.04
you	0.04
dogs	0.08
fish	0.04
elephants	0.08
love	0.12
loves	0.04
we	0.04
they	0.04

我	I	0.50
我	me	0.50
你	you	1.00
他	he	1.00
貓	cats	1.00
狗	dogs	1.00
魚	fish	1.00
豬	pigs	1.00
猴子	monkeys	1.00
大象	elephants	1.00
喜歡	love	0.83
喜歡	loves	0.17
我們	we	1.00
他們	they	1.00
。	.	1.00

User’s Input

狗喜歡大象。

Dogs love elephants.
Dogs elephants love.
Love dogs elephants.
Love elephants dogs.
Elephants love dogs.
Elephants dogs love.
Dogs loves elephants.
Dogs elephants loves.
Loves dogs elephants.
Loves elephants dogs.
Elephants loves dogs.
Elephants dogs loves.

MT Output

Dogs love elephants.

我	I	0.50
我	me	0.50
你	you	1.00
他	he	1.00
貓	cats	1.00
狗	dogs	1.00
魚	fish	1.00
豬	pigs	1.00
猴子	monkeys	1.00
大象	elephants	1.00
喜歡	love	0.83
喜歡	loves	0.17
我們	we	1.00
他們	they	1.00
。	.	1.00

狗	dogs	1.00
大象	elephants	1.00
喜歡	love	0.83
喜歡	loves	0.17
。	.	1.00

Bilingual Table

dogs	0.08
elephants	0.08
love	0.12
loves	0.04

Monolingual Table

Topics

- Machine translation (MT) and translation technology
- Rule-based Machine Translation (RBMT)
- Example-based Machine Translation (EBMT)
- Statistical Machine Translation (SMT)
- Neural Machine Translation (NMT)
- MT Skills

Traditional Approaches to Machine Translation

01	Rule-based Machine Translation	• I go to school by bus. • 我／去／學校／乘／公車／。 • 我乘公車去學校。
02	Example-based Machine Translation	• I go to school by bus. • I go to school by taxi. 我乘計程車去學校。 • 我乘公車去學校。
03	Statistical Machine Translation	• Translation Model P(“I” | “我”) x P(“go” | “去”) x ... • Language Model: P(“我乘公車去學校”)

New Approach: Neural Machine Translation

1. Deep Learning: the use of multiple layers of artificial neurons
2. Automatic feature engineering: Converting the input into an internal representation that captures the features of the input by using deep neural networks
3. Areas of Application: Speech recognition, handwriting recognition, image processing, and machine translation

Neural Machine Translation

Neural machine translation often adopts the encoder-decoder architecture with recurrent neural networks (RNN) to model the translation process. The bidirectional RNN encoder which consists of a forward RNN and a backward RNN reads a source sentence $\mathbf{x} = x_1, x_2, ..., x_{T_x}$ and transforms it into word annotations of the entire source sentence $\mathbf{h} = h_1, h_2, ..., h_{T_x}$. The decoder uses the annotations to emit a target sentence $\mathbf{y} = y_1, y_2, ..., y_{T_y}$ in a word-by-word manner.

In the training phase, given a parallel sentence $(\mathbf{x}, \mathbf{y})$, NMT models the conditional probability as follows,

$$P(\mathbf{y}|\mathbf{x}) = \prod_{i=1}^{T_y} P(y_i|\mathbf{y_{<i}}, \mathbf{x}) \quad (1)$$

where y_i is the target word emitted by the decoder at step i and $\mathbf{y_{<i}} = y_1, y_2, ..., y_{i-1}$. The conditional probability $P(y_i|\mathbf{y_{<i}}, \mathbf{x})$ is computed as

$$P(y_i|\mathbf{y_{<i}}, \mathbf{x}) = softmax(f(s_i, y_{i-1}, c_i)) \quad (2)$$

where $f(\cdot)$ is a non-linear function and s_i is the hidden state of the decoder at step i:

$$s_i = g(s_{i-1}, y_{i-1}, c_i) \quad (3)$$

where $g(\cdot)$ is a non-linear function. Here we adopt Gated Recurrent Unit (Cho et al., 2014) as the recurrent unit for the encoder and decoder. c_i is the context vector, computed as a weighted sum of the annotations $\mathbf{h}$:

$$c_i = \sum_{j=1}^{T_x} \alpha_{t,j} h_j \quad (4)$$

where h_j is the annotation of source word x_j and its weight $\alpha_{t,j}$ is computed by the attention model.

We train the attention-based NMT model by maximizing the log-likelihood:

$$C(\theta) = \sum_{n=1}^{N} \sum_{i=1}^{T_y} \log P(y_i^n|\mathbf{y}_{<i}^n, \mathbf{x}^n) \quad (5)$$

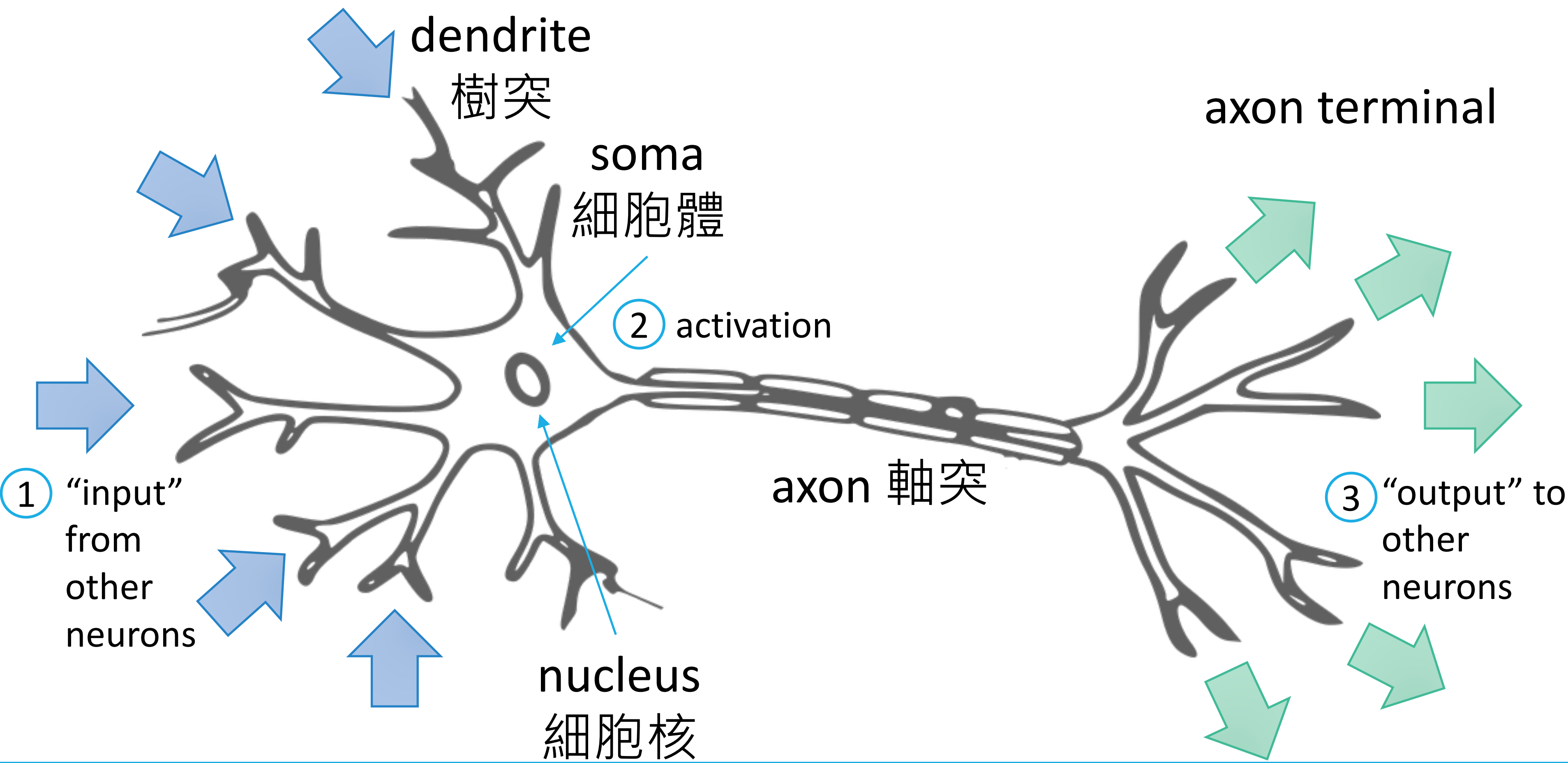
given the training data with N bilingual sentences (Cho, 2015).

In the testing phase, given a source sentence $\mathbf{x}$, we use beam search strategy to search a target sentence $\hat{\mathbf{y}}$ that approximately maximizes the conditional probability $P(\mathbf{y}|\mathbf{x})$

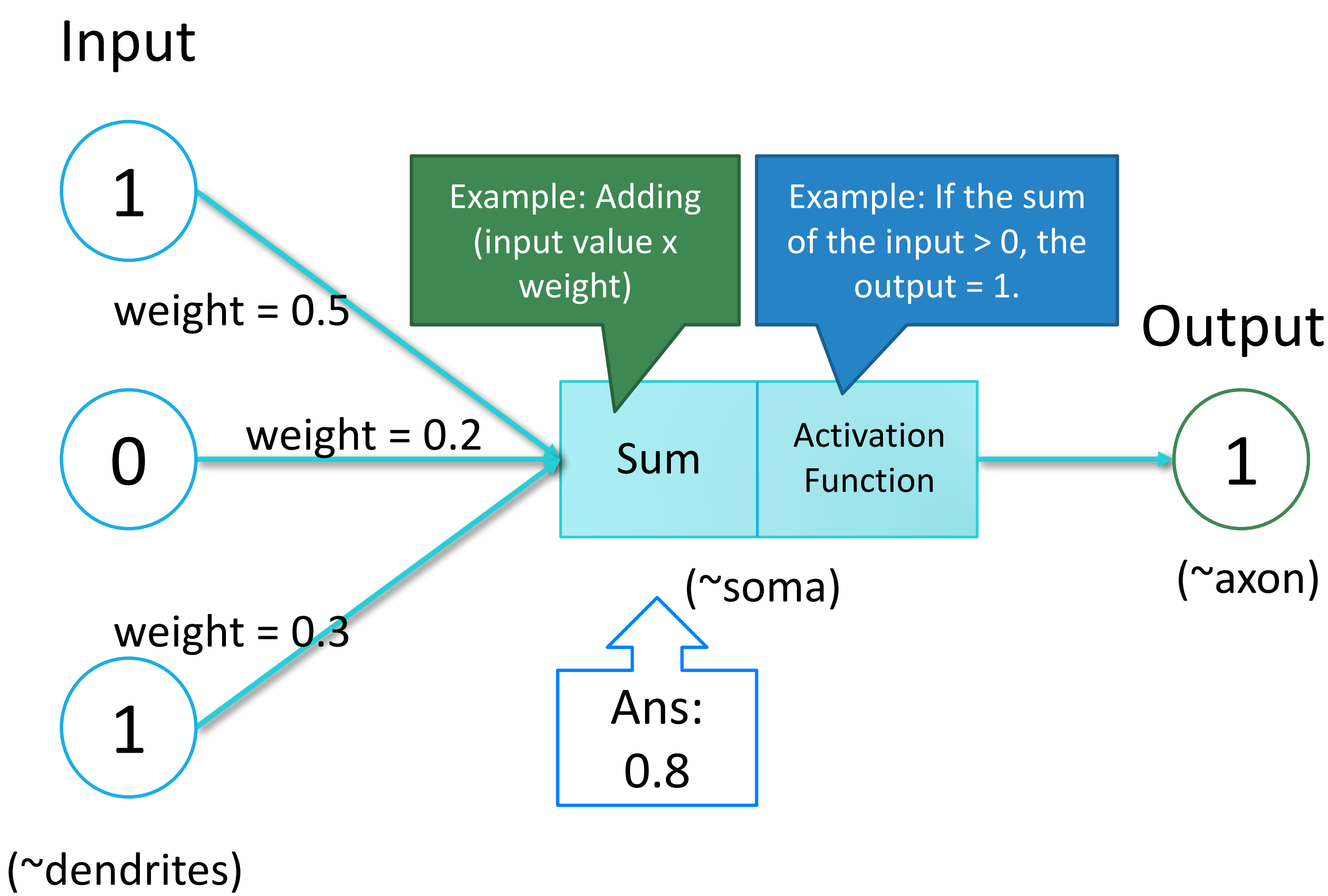
$$\hat{\mathbf{y}} = \operatorname*{argmax}_{\mathbf{y}} P(\mathbf{y}|\mathbf{x}) \quad (6)$$

(http://aclweb.org/anthology/D17-1149)

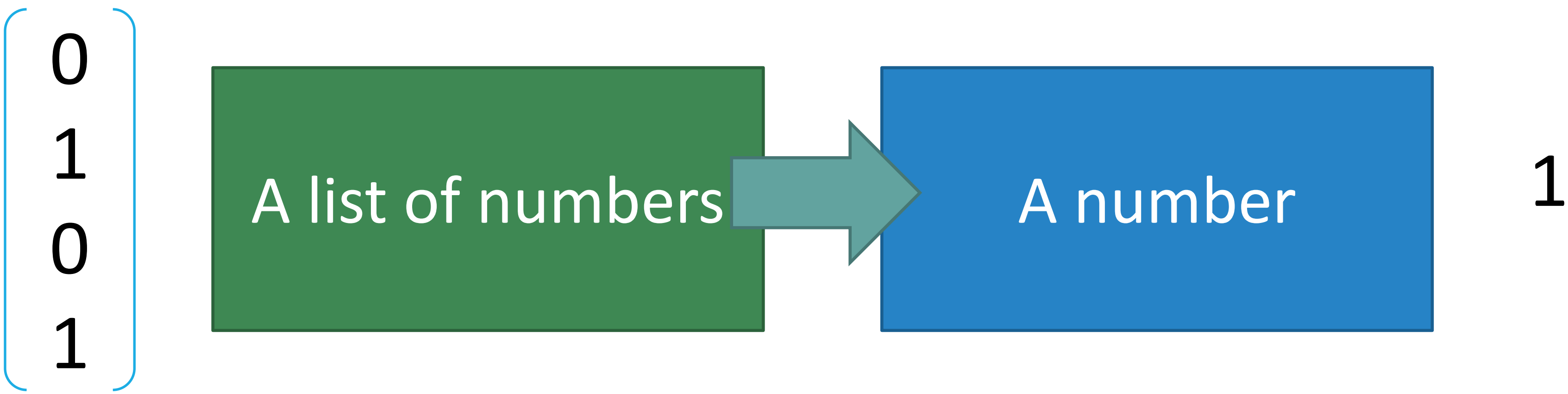
Biological Neuron



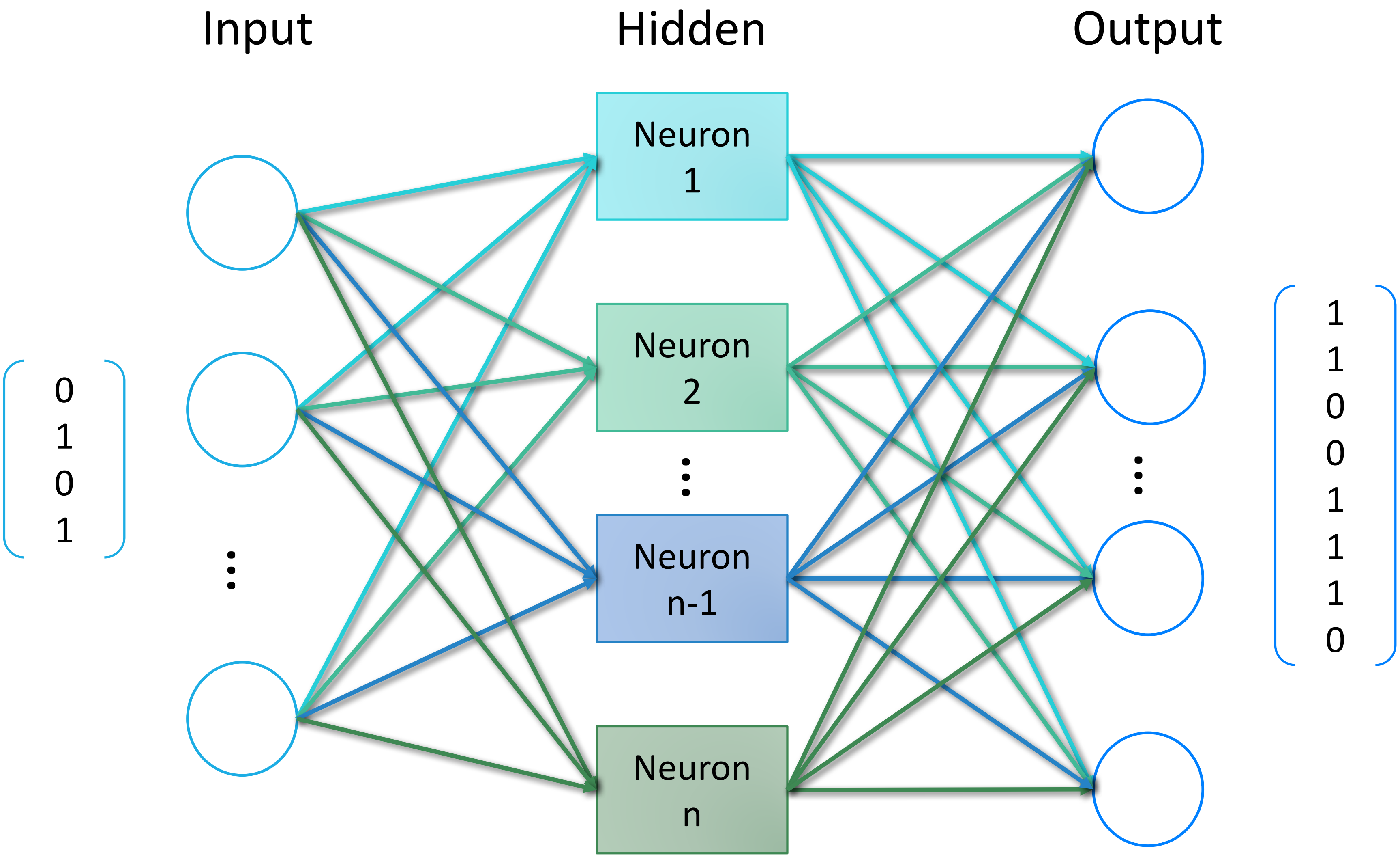
Artificial Neuron



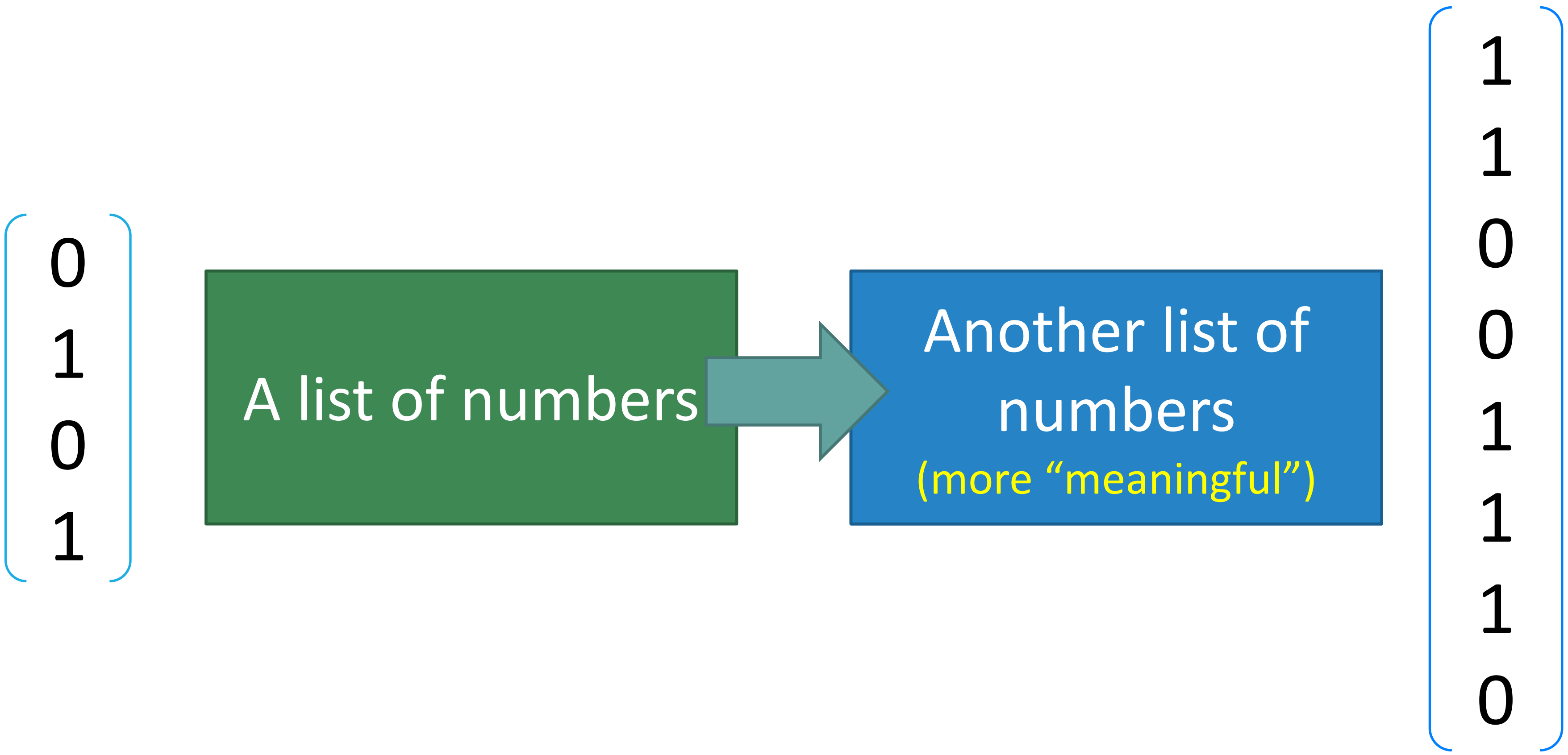
Artificial Neuron



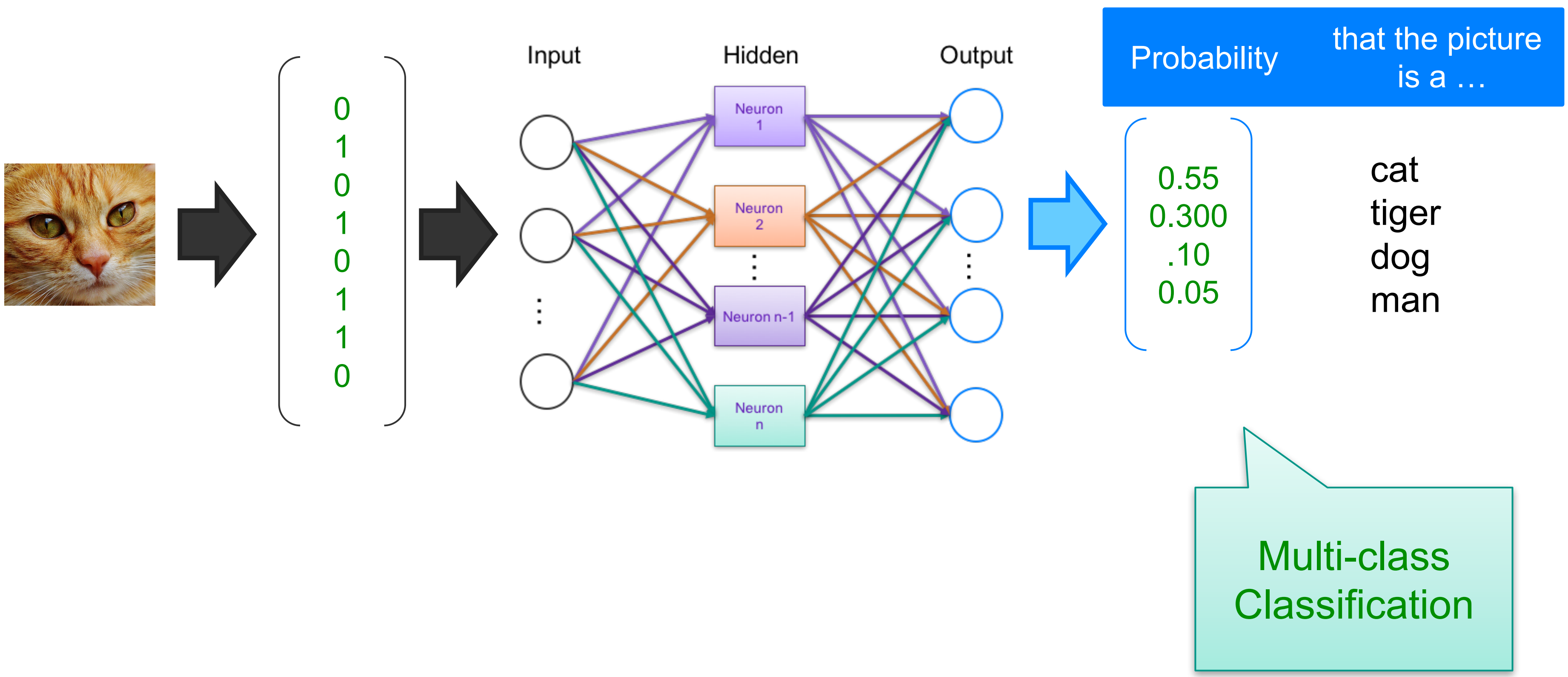
Artificial Neural Network: Many Neurons



Artificial Neural Network



Artificial Neural Network and Image Processing



NMT as a Prediction Problem: Given X, what is the next word?

Given the following:

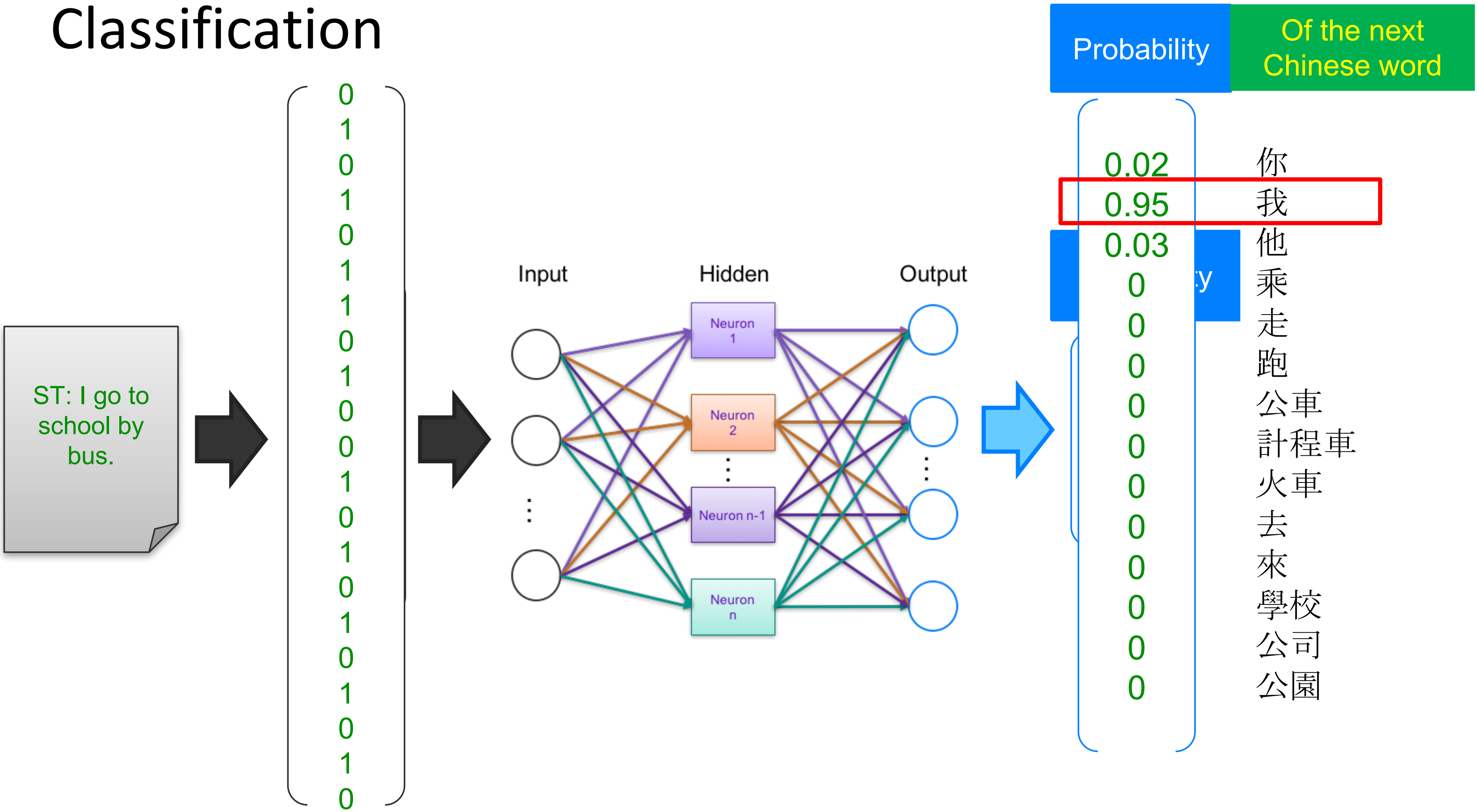
ST: I / go / to / school / by / bus / . /

TT: 我 / 乘

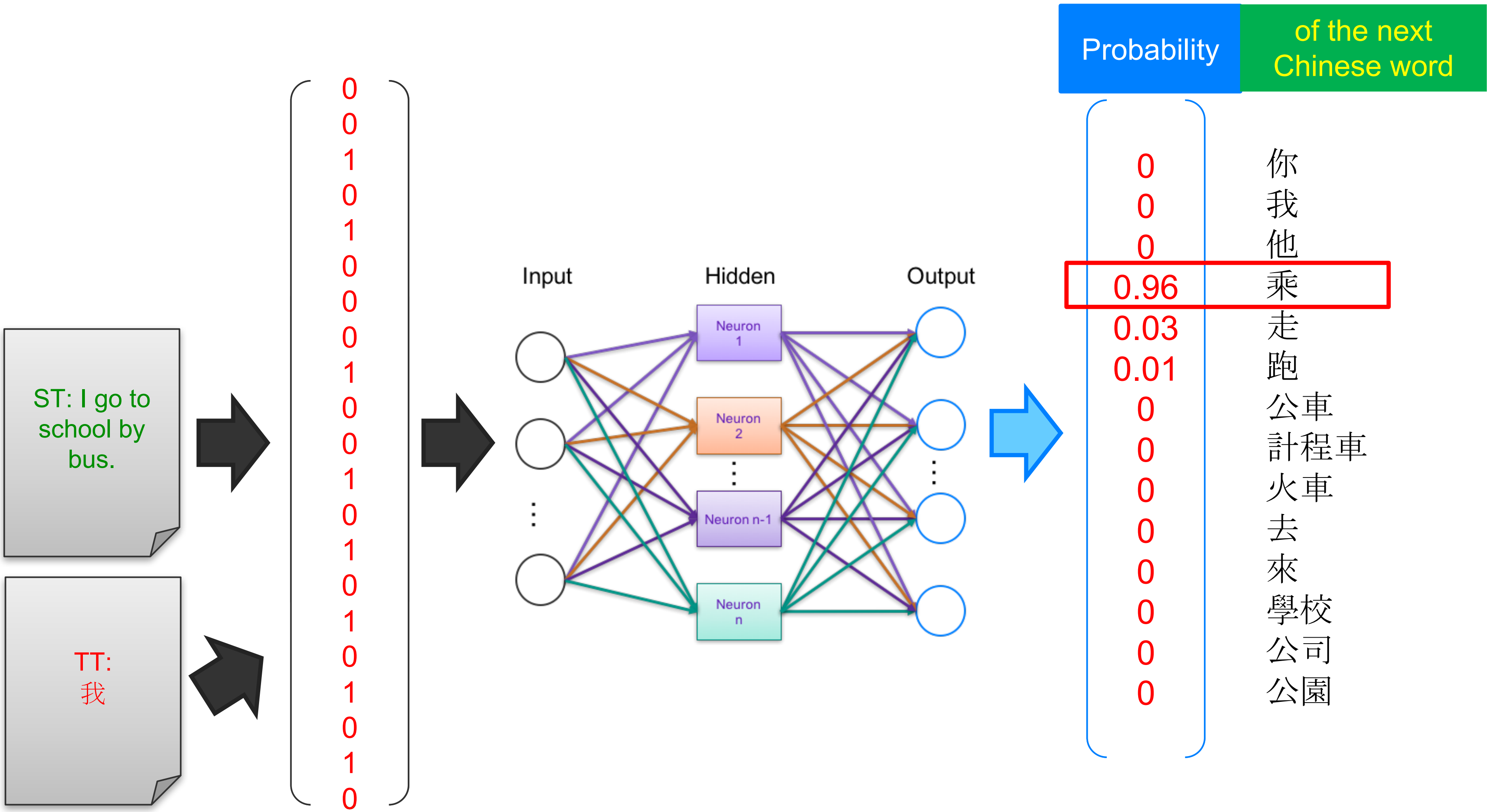
Which of the following is most likely to be the next word?

公車
火車
學校
咖啡
電腦
翻譯

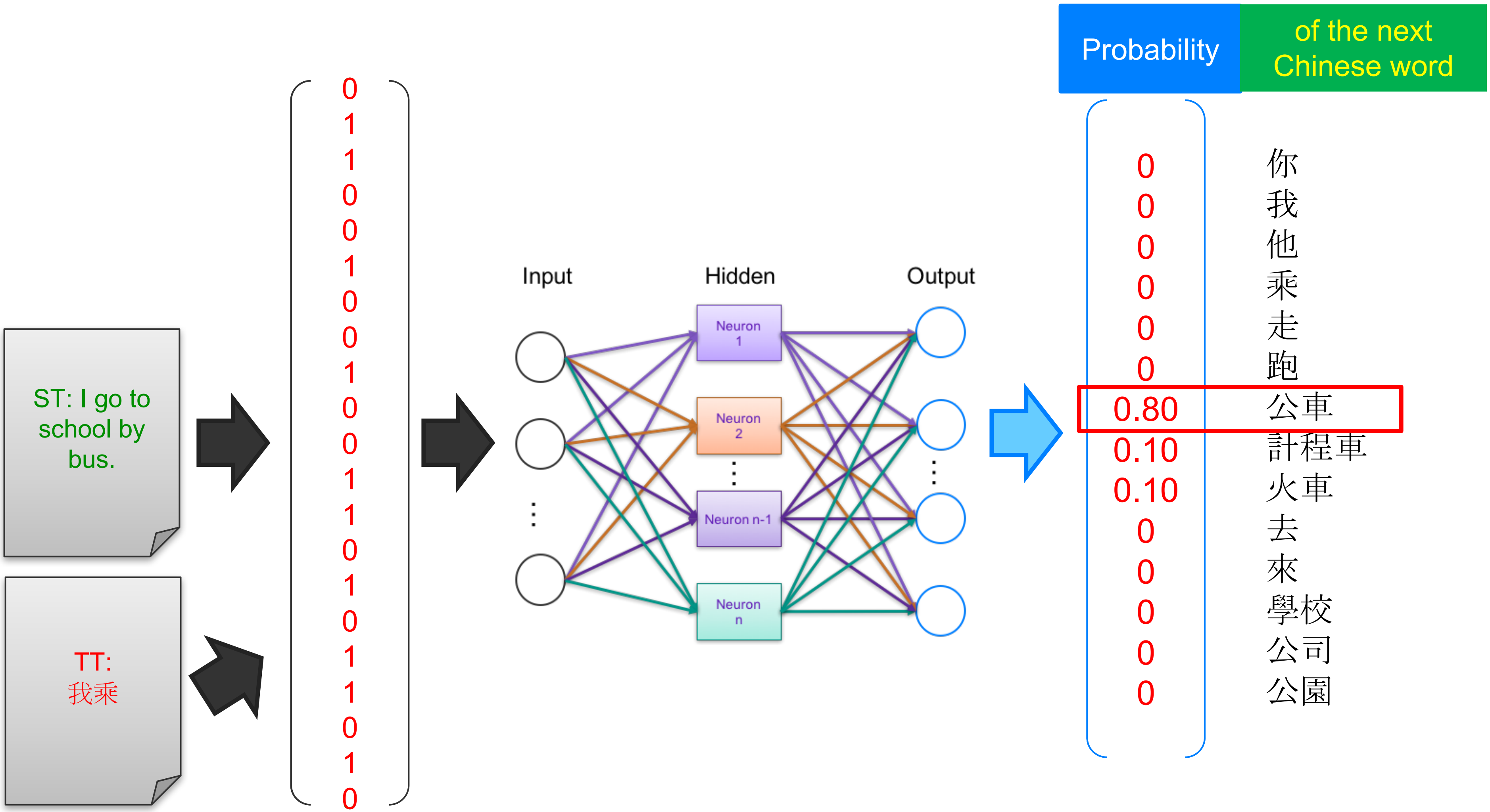
Neural Machine Translation as Multi-class Classification



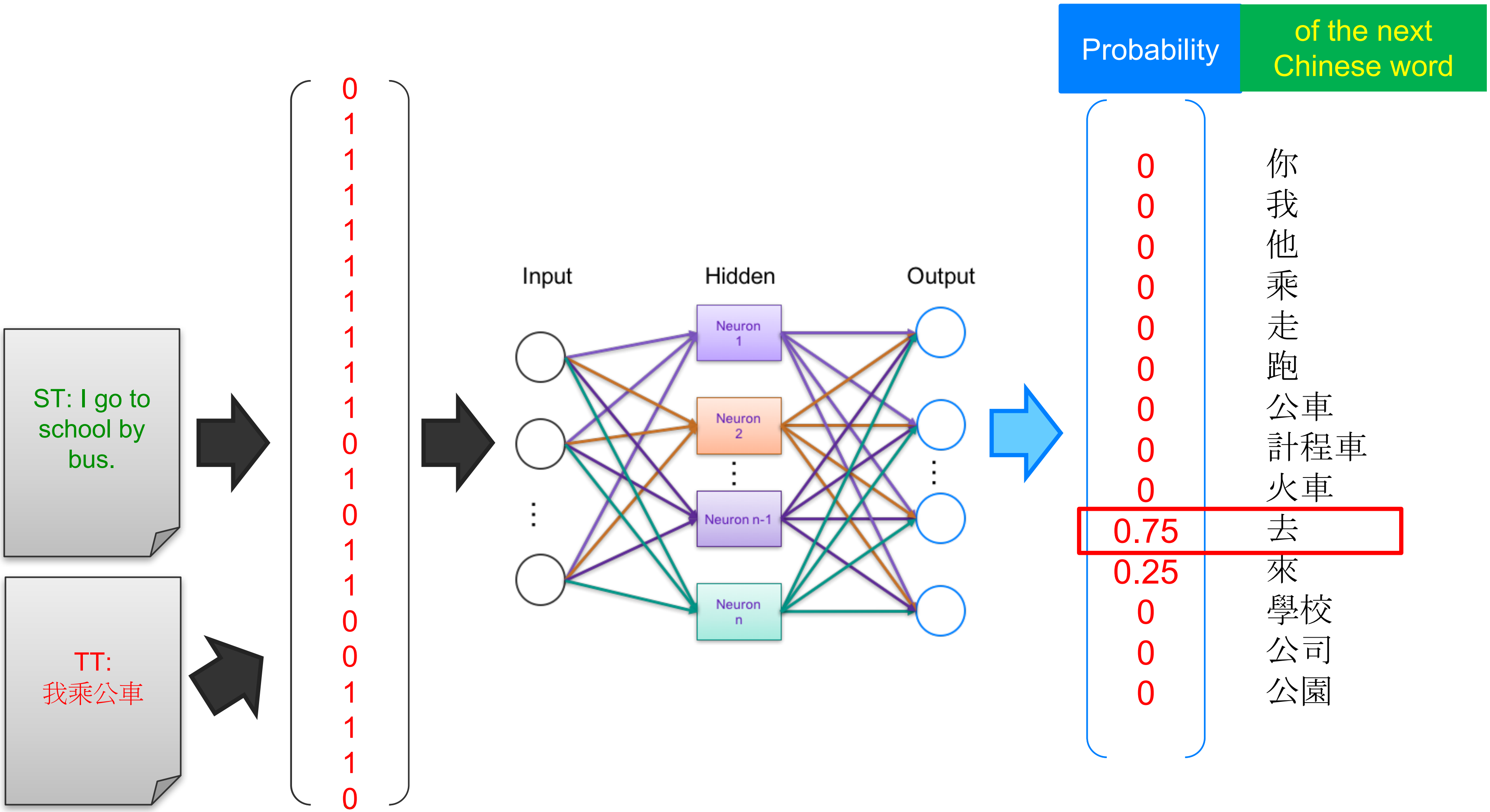
Neural Machine Translation



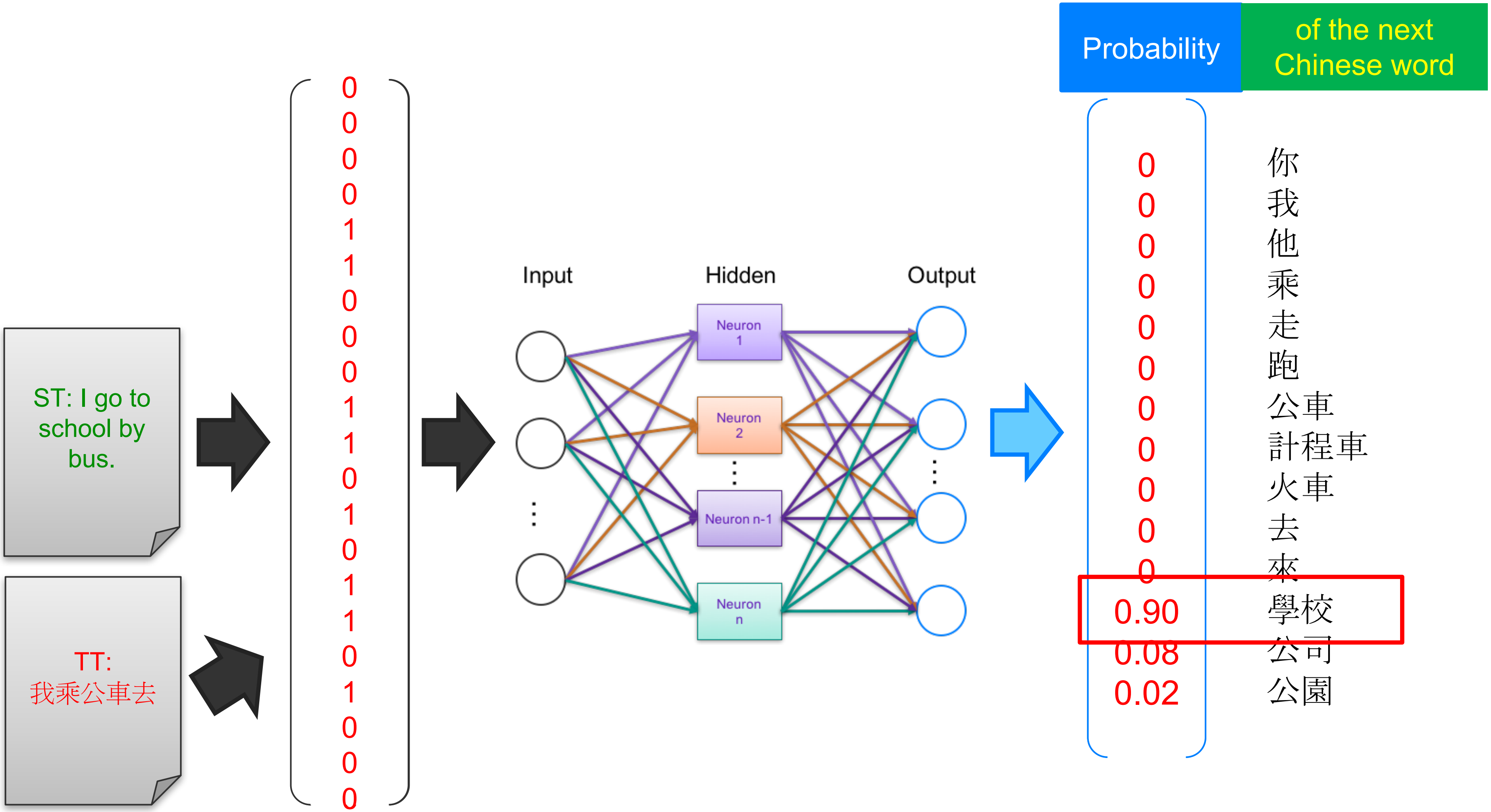
Neural Machine Translation



Neural Machine Translation



Neural Machine Translation



Types of Neural Machine Translation

1. Encoder-decoder (Recurrent Neural Networks using Long-short Term Memory Cells or Gated Recurrent Units) with Attention (Bahdanau, Cho, & Bengio, 2014; Luong, Pham, & Manning, 2015; Wu et al., 2016)
2. Convolutional Neural Networks (Gehring et al., 2017)
3. Self-attentional Transformers (Vaswani et al., 2017)

Examples (since 2016)

1. Google Neural Machine Translation
2. Microsoft Translator: AI Chinese-English machine translation system (news translation) that can translate with the same accuracy as a human
3. Tencent: Conference Interpreting (Simultaneous transcription and translation)
4. Sogou: WMT 2017 News Translation Task (1st in Chinese-English Translation and 3rd in English-Chinese Translation)

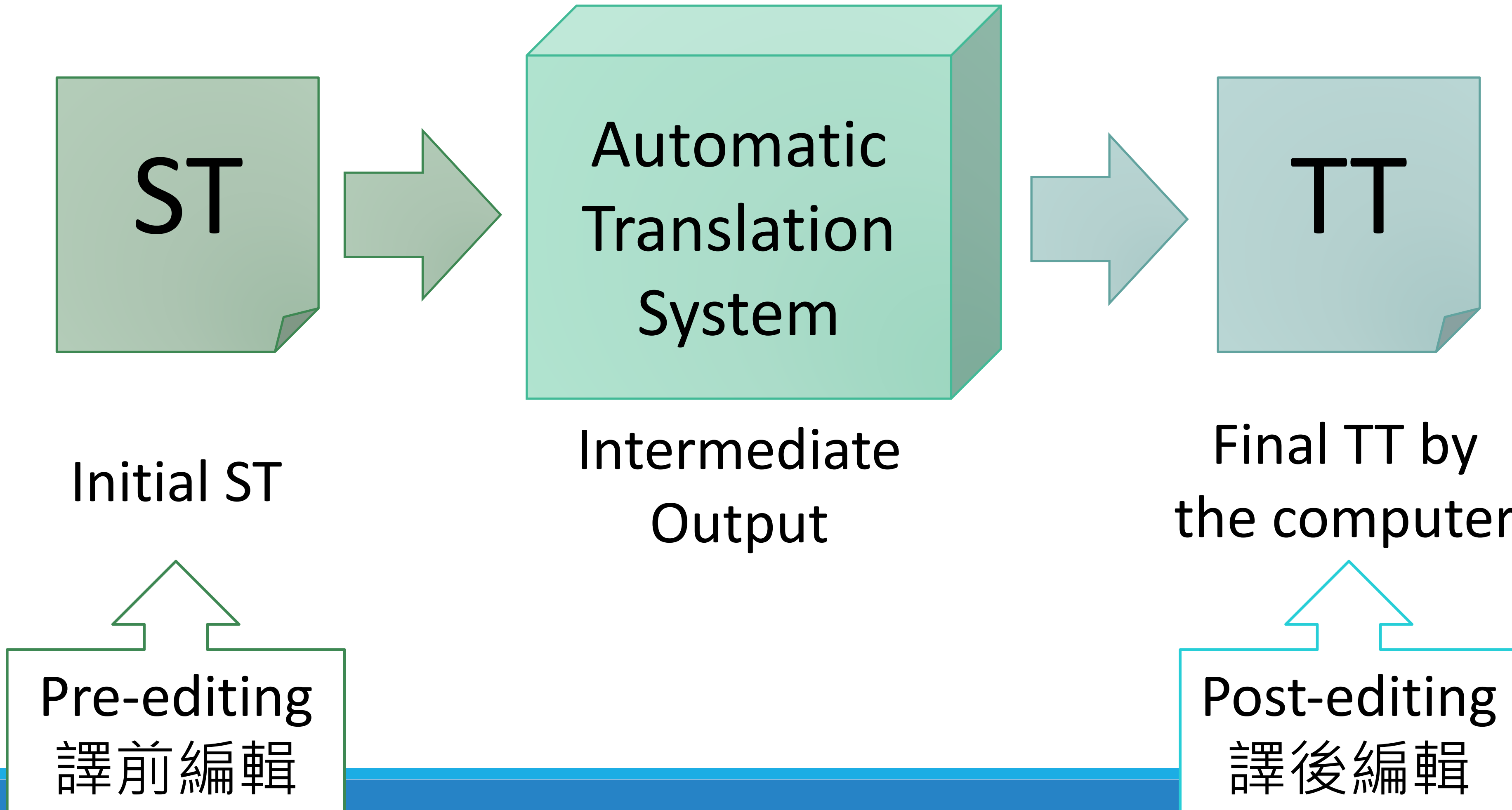
Topics

- Machine translation (MT) and translation technology
- Rule-based Machine Translation (RBMT)
- Example-based Machine Translation (EBMT)
- Statistical Machine Translation (SMT)
- Neural Machine Translation (NMT)
- MT Skills

MT Skills

- Editing: Pre-editing and Post-editing
- Evaluation

Modes of Editing



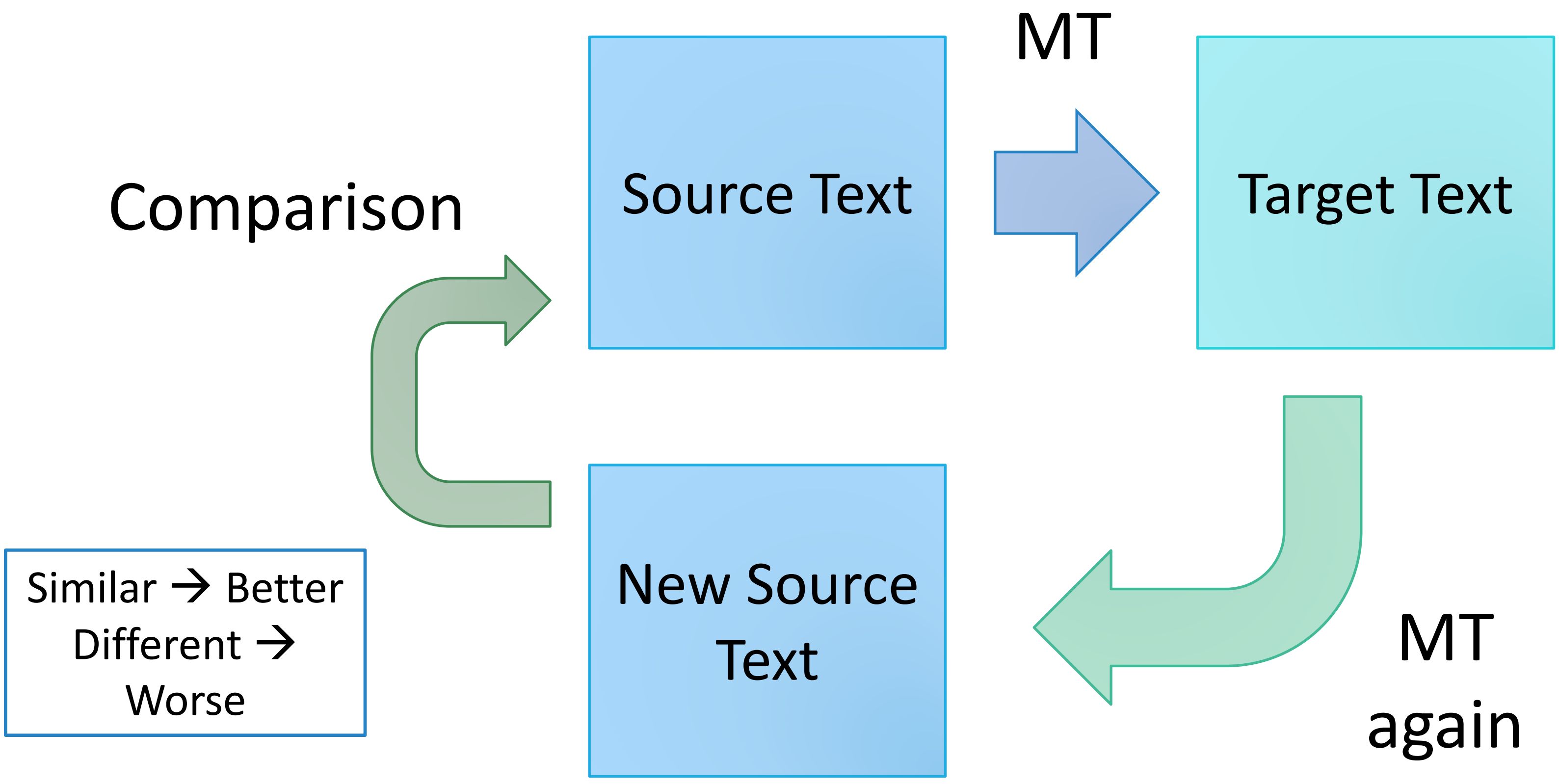
Example

申請期於十月二十日開始，至十一月二日晚上七時截止，在申請期開始前或截止時間後所遞交的申請表將不獲處理。

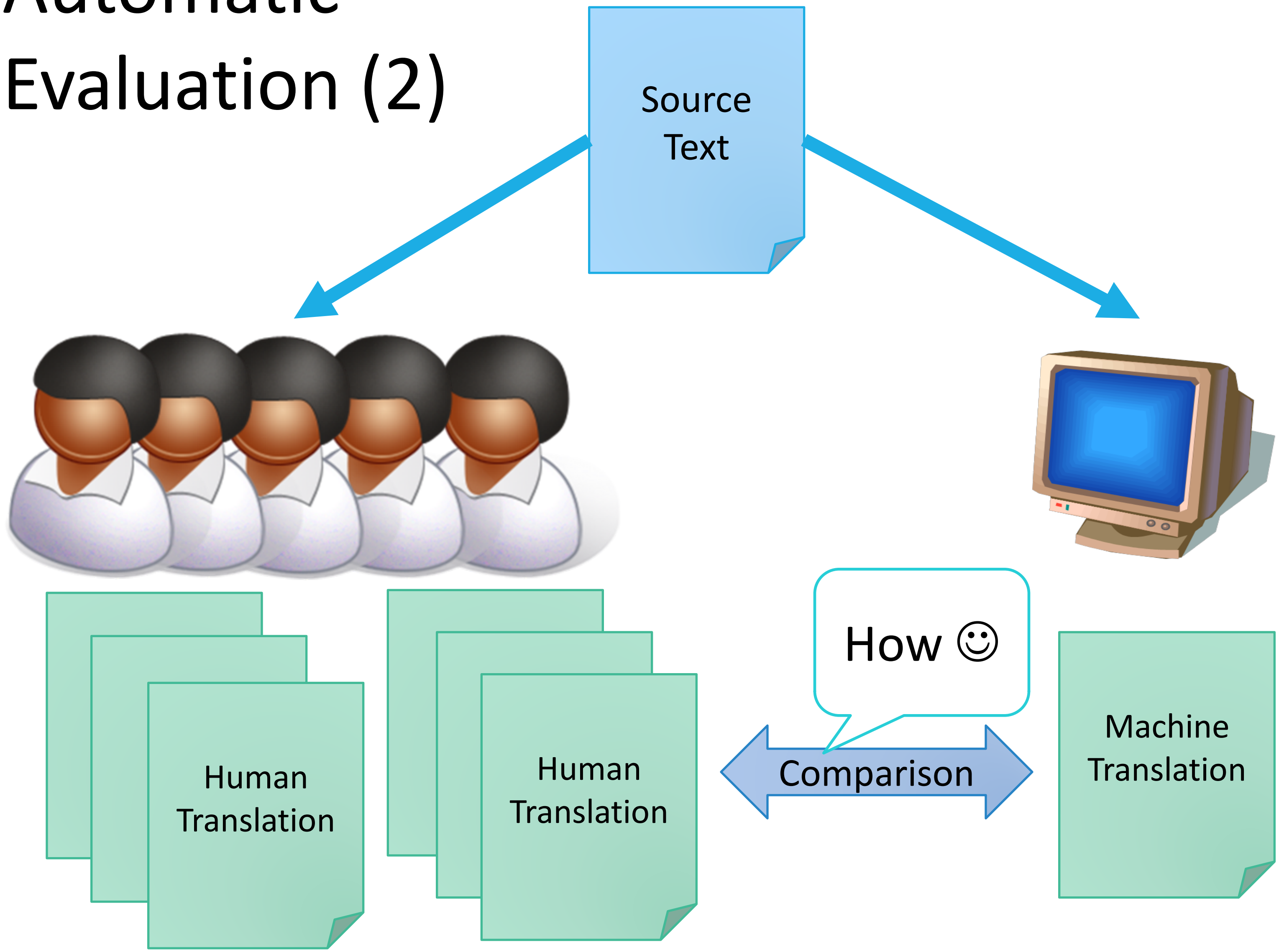
Example

- Application forms will be closed from 20 October to 7 pm on November 2 and applications submitted before or after the application period will not be processed.
- The application period begins on October 20 and ends at 7 pm on November 2. Applications submitted before or after the application period will not be processed.

Automatic Evaluation (1): Round-trip Translation



Automatic Evaluation (2)



Thank you!