Using text generation model to restrict vocabulary selection of NMT

テキスト生成モデルを用いて NMT モデルの出力を制限する研究

NGUYEN TRONG NGOC, 山本和英
長岡技術科学大学自然言語処理

Contents

1. Background
2. Introduction of Neural Machine Learning (NMT)
3. New method for restricting vocabulary selection
4. Future works

Background

Machine Translation (MT) is an important task in natural language processing. The dream of this task is creating an automatic high-quality translation.

Progress in MT:

Phrase-based statical MT(1993) → Syntax-based statical MT(2003) → Neural MT (2015)

Background

An Vietnamese-English example:

Vietnamese: 600 người Tây Ban Nha đã đổ bộ vào Mexico để chinh phục Đế quốc Aztec với dân số vài triệu người. Họ đã mất hai phần ba binh lính trong cuộc đụng độ đầu tiên.

2015 year google translation: 600 Spaniards landed in Mexico, millions of people to conquer the Aztec empire, the first two-thirds of soldiers against their loss.
(From ACL 2016 tutorial https://sites.google.com/site/acl16nmt/)

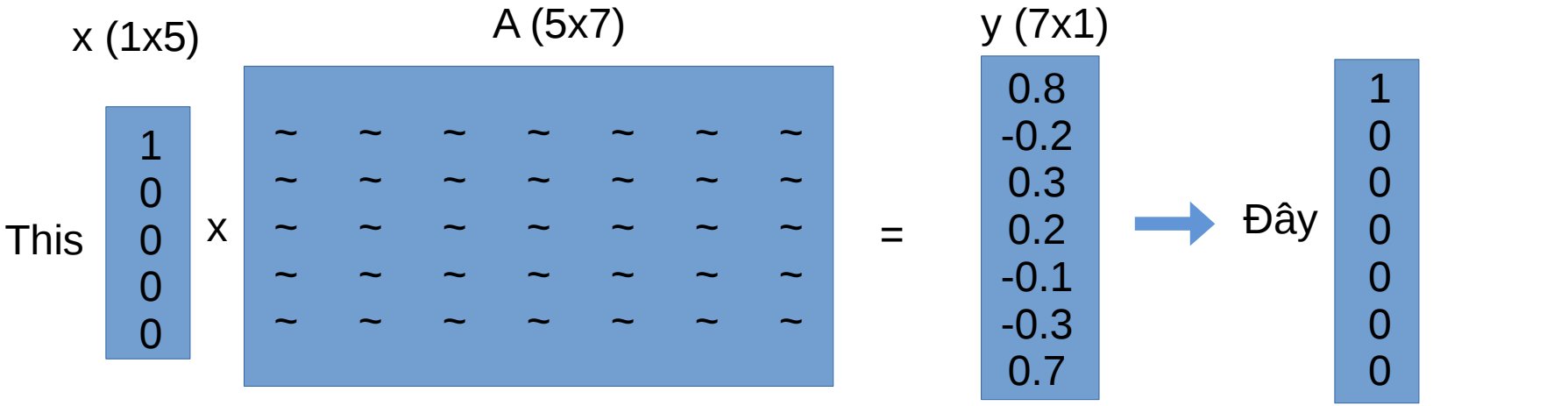
2018 year google translation: 600 Spaniards landed in Mexico to conquer the Aztec Empire with a population of several million. They lost two-thirds of the troops in the first clash.

Human translation: 600 Spaniards landed in Mexico to conquer the Aztec Empire with a population of a few million. They lost two thirds of their soldiers in the first clash.

Neural machine translate

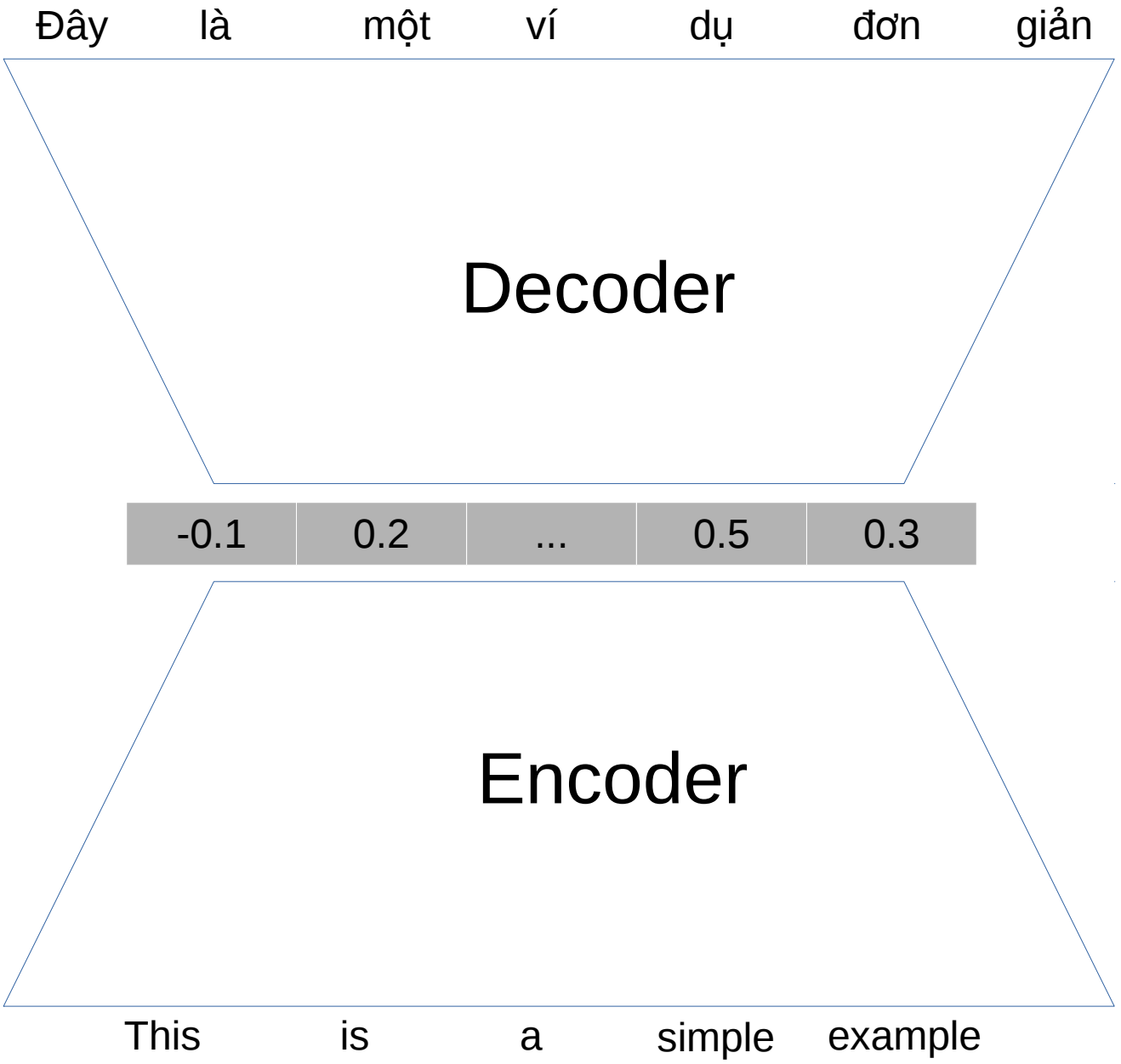
A neural network translation for a single word.

This	is	a	simple	example
Đây	là	một	đơn giản	ví dụ



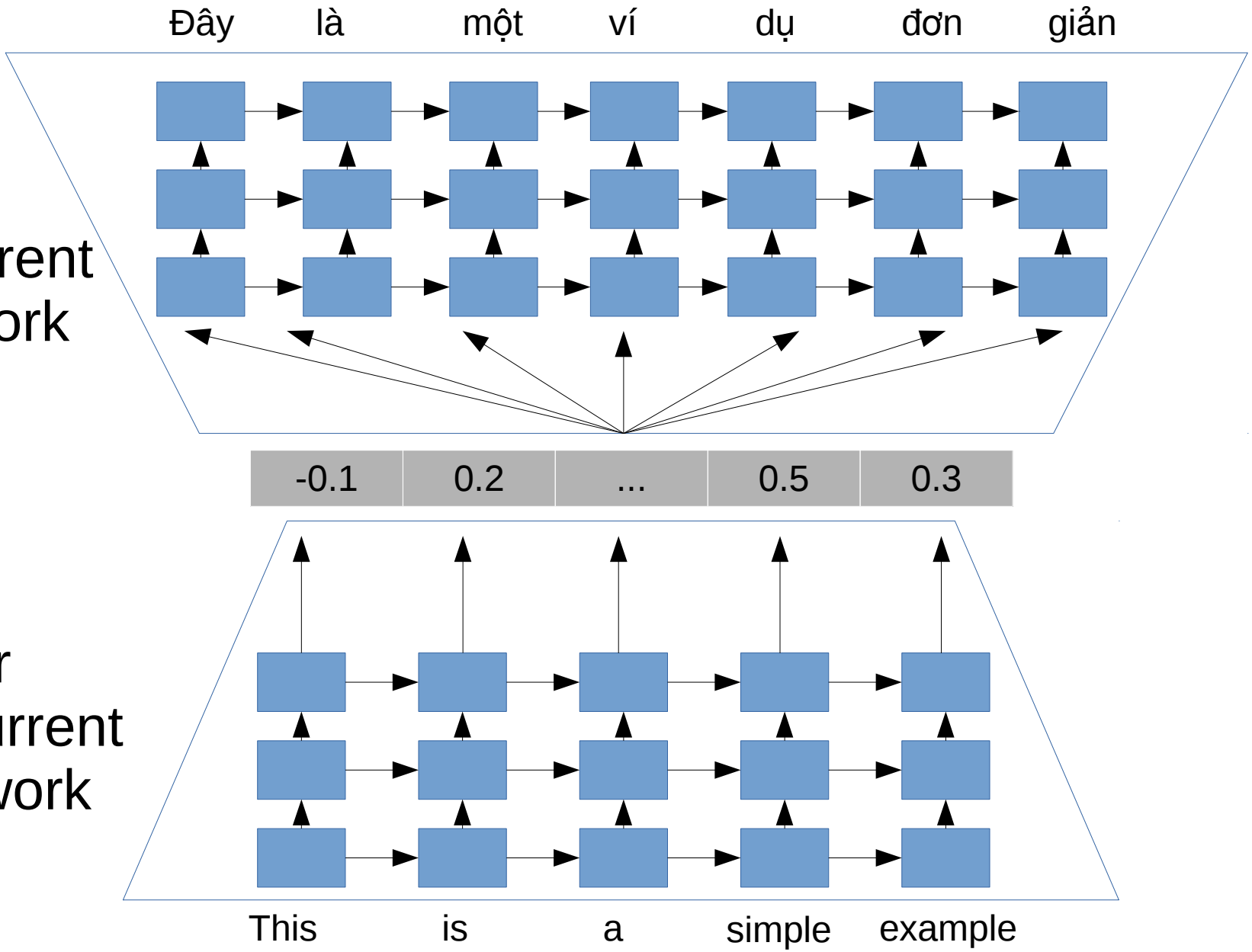
Neural machine translate

Neural encoder-decoder architecture:



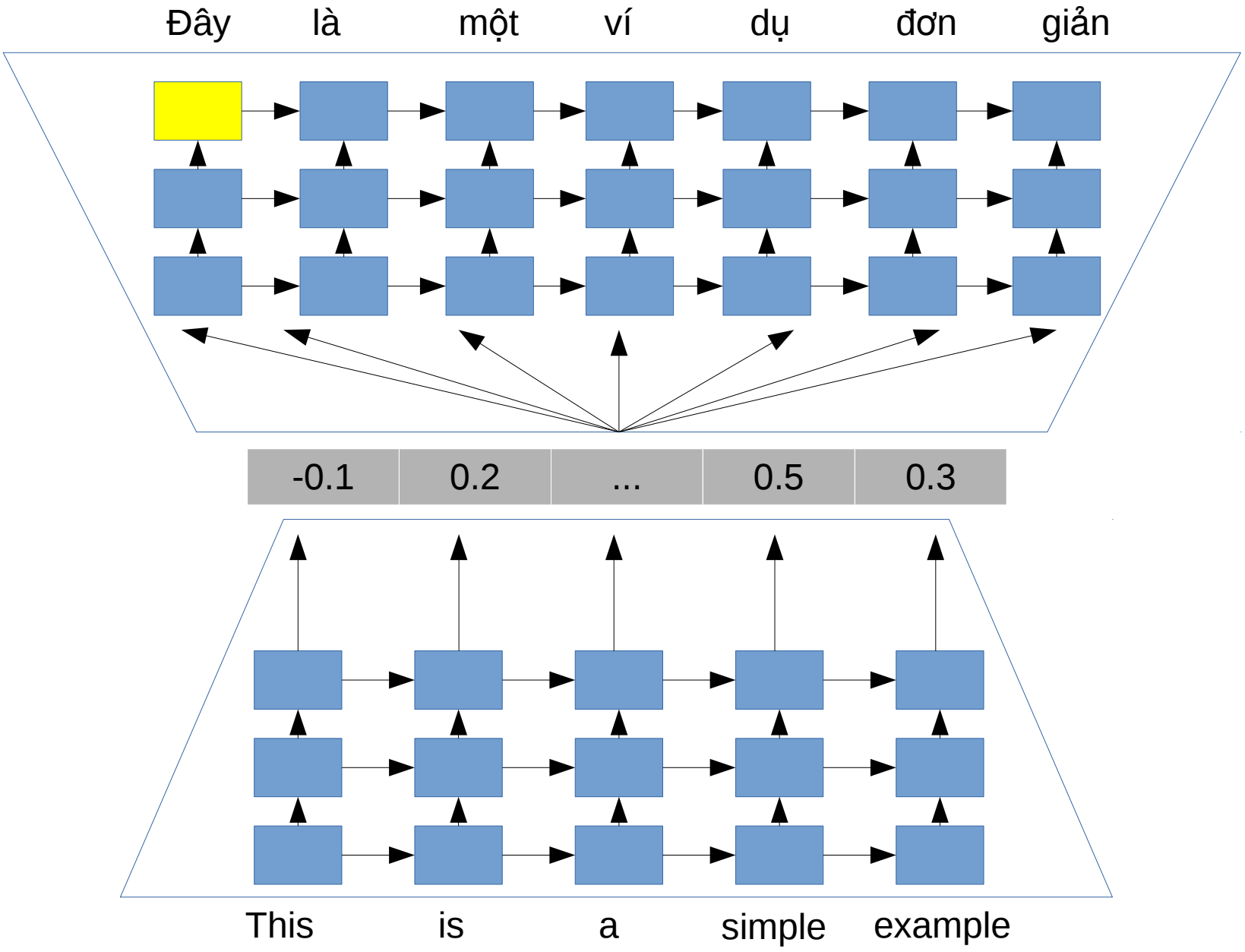
Neural machine translate

Sequence to sequence architectures



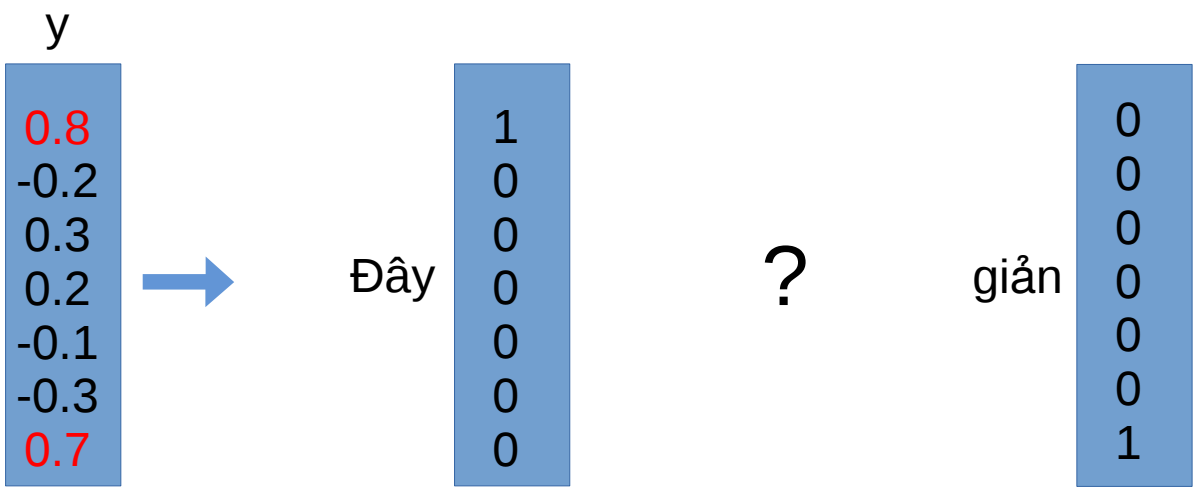
Neural machine translate

One of problems:



Neural machine translate

At the last layer (Projection layer), we need a large training corpus for the best selection.
For low-resource language such as Vietnamese, the problem is deep recurrent networks can not learn with high accuracy.
Can we solve low-resource problem by raw corpus.

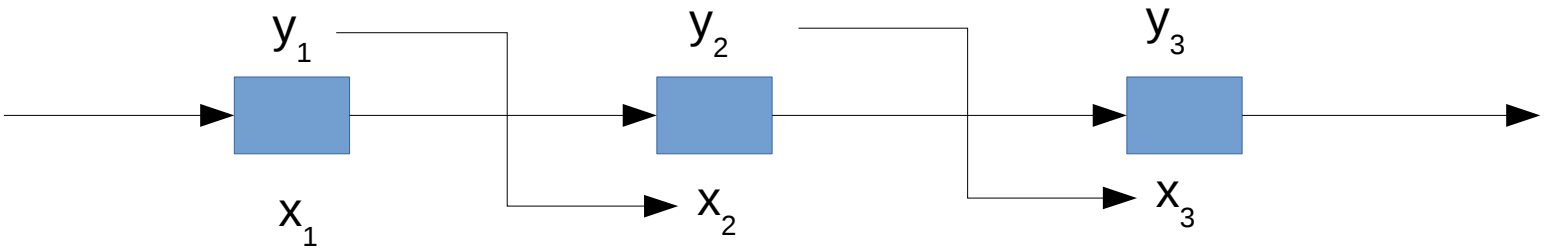


Proposed method

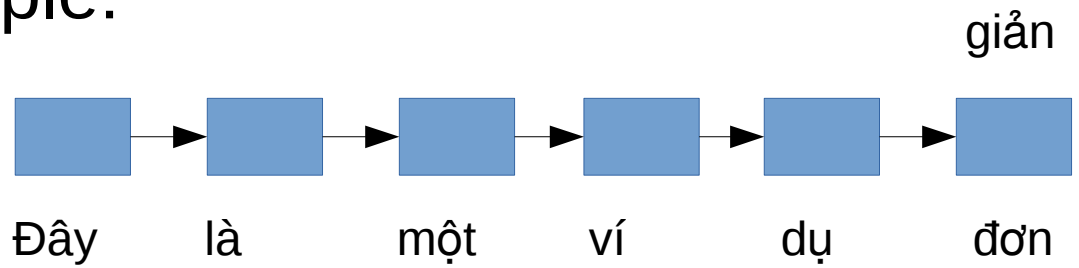
Using a text generation model to restrict vocabulary selection of NMT.

Text generation model

Text generation model predict the next word after reading context of preceding words.

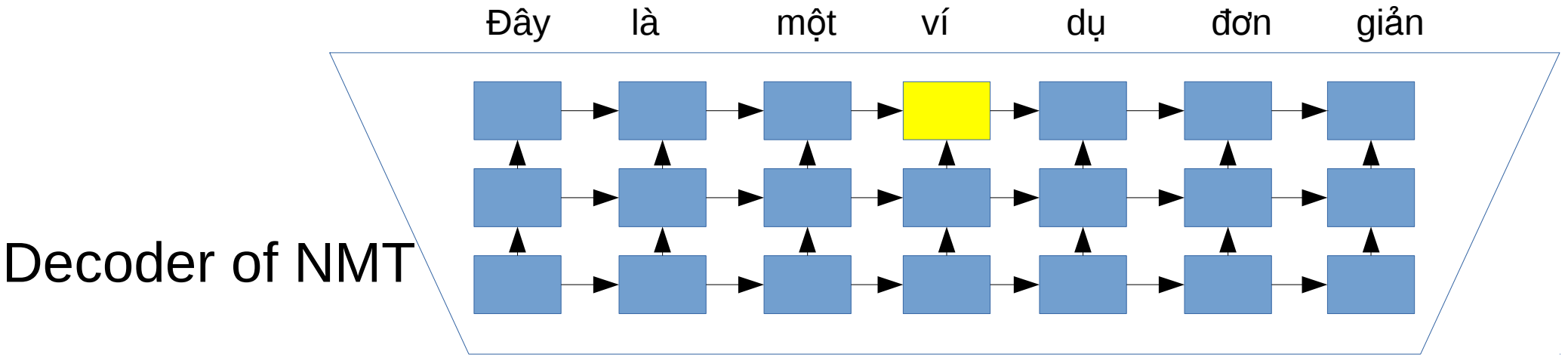


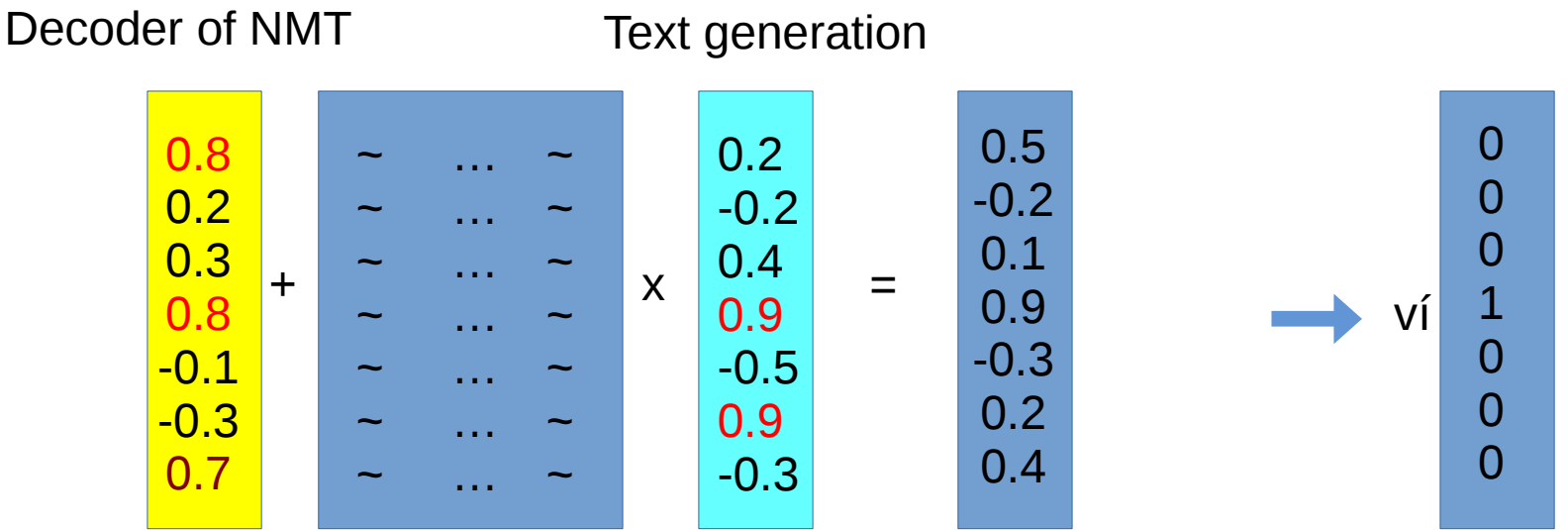
Example:



Proposed method

Combining 2 models:





Evaluation

Usually Machine Translation systems are evaluated in BLEU score.
BLEU (bilingual evaluation understudy) is an algorithm for evaluating the quality of text which has been translated.

System	En-Vi Translation	Vi-En Translation
SOTA	27.01	24.61
Stanford-NMT	26.90	24.38
Our approach	26.84	23.90

SOTA machine translation:The iwslt 2015 evaluation campaign.
Stanford NMT:Stanford neural machine translation systems for spoken language domains

Future work

Upgrade Text generation model:

- Classify training corpus for specific translation (Ex: Daily conversation, historical document, science paper, ...)
- Use high-quality text generation model (Ex: Neural text generation)

Upgrade NMT:

- Incorporating with other algorithm such as dropout, beam search, …
- Attaching attribution as input to improve accuracy. (Ex: POS tag,)