Machine Translation

How does Seq2seq behave
with multi-head attention?

Illustrations from Storyset

1.

Introduction

EUROPEAN
PARLIAMENT

This project
Seq2Seq
Multi-head Attention

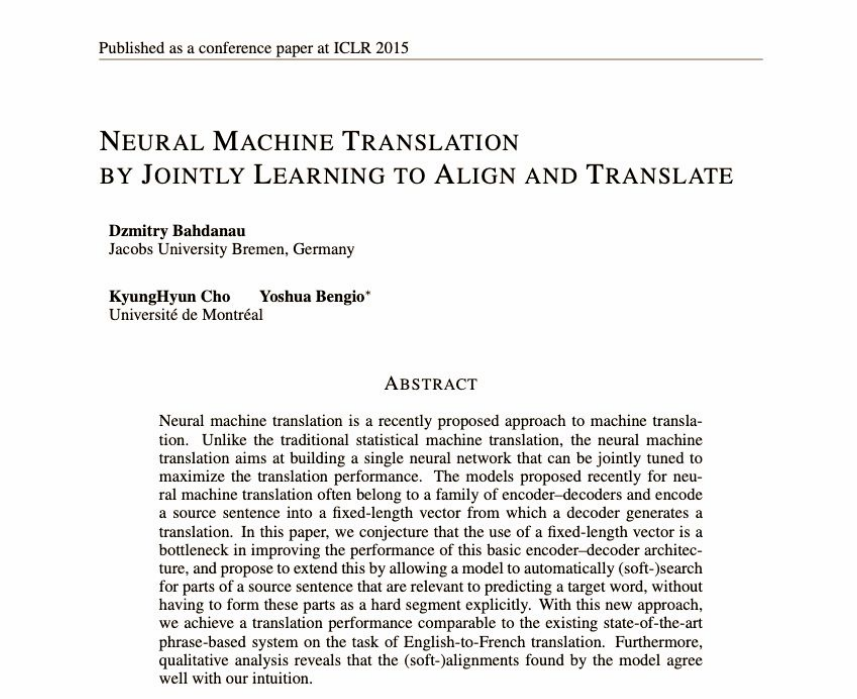
Tao, C., Gao, S., Shang, M., Wu, W., Zhao, D., & Yan, R. (2018, July). Get The Point of My Utterance! Learning Towards Effective Responses with Multi-Head Attention Mechanism. In IJCAI (pp. 4418-4424).

1.

Background

Bahdanau Attention

- The OG attention
- Single head



Published as a conference paper at ICLR 2015

NEURAL MACHINE TRANSLATION BY JOINTLY LEARNING TO ALIGN AND TRANSLATE

Dzmitry Bahdanau
Jacobs University Bremen, Germany

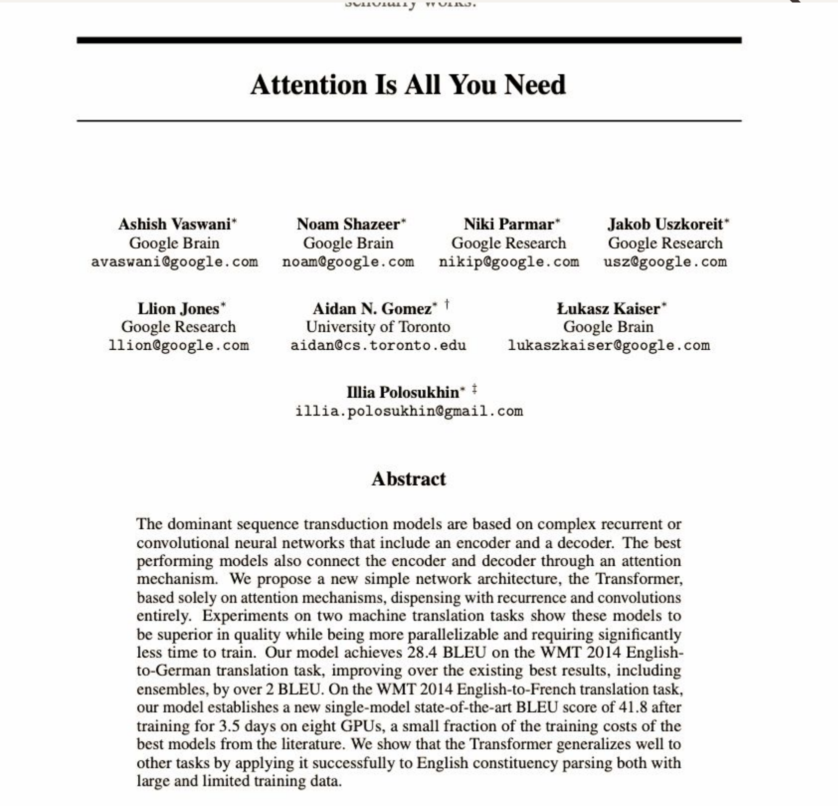
KyungHyun Cho **Yoshua Bengio**[*]
Université de Montréal

ABSTRACT

Neural machine translation is a recently proposed approach to machine translation. Unlike the traditional statistical machine translation, the neural machine translation aims at building a single neural network that can be jointly tuned to maximize the translation performance. The models proposed recently for neural machine translation often belong to a family of encoder–decoders and encode a source sentence into a fixed-length vector from which a decoder generates a translation. In this paper, we conjecture that the use of a fixed-length vector is a bottleneck in improving the performance of this basic encoder–decoder architecture, and propose to extend this by allowing a model to automatically (soft-)search for parts of a source sentence that are relevant to predicting a target word, without having to form these parts as a hard segment explicitly. With this new approach, we achieve a translation performance comparable to the existing state-of-the-art phrase-based system on the task of English-to-French translation. Furthermore, qualitative analysis reveals that the (soft-)alignments found by the model agree well with our intuition.

Multi-Head Attention

- Commonly used for Transformers
- Each head using Scaled Dot-Product Attention



scholarly works.

Attention Is All You Need

Ashish Vaswani[*]
Google Brain
avaswani@google.com

Noam Shazeer[*]
Google Brain
noam@google.com

Niki Parmar[*]
Google Research
nikip@google.com

Jakob Uszkoreit[*]
Google Research
usz@google.com

Llion Jones[*]
Google Research
llion@google.com

Aidan N. Gomez[*] [†]
University of Toronto
aidan@cs.toronto.edu

Łukasz Kaiser[*]
Google Brain
lukaszkaiser@google.com

Illia Polosukhin[*] [‡]
illia.polosukhin@gmail.com

Abstract

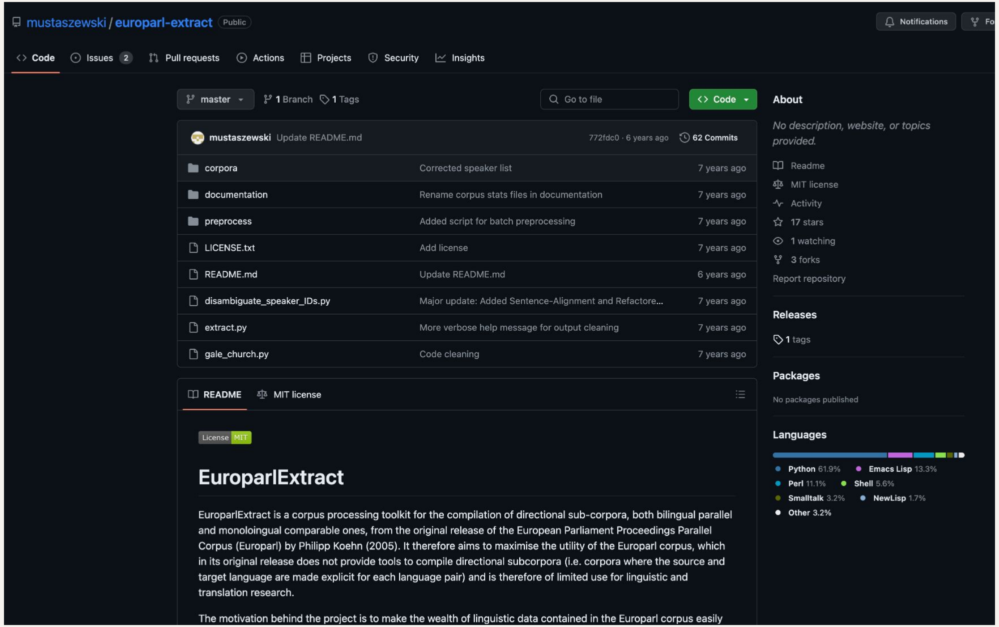
The dominant sequence transduction models are based on complex recurrent or convolutional neural networks that include an encoder and a decoder. The best performing models also connect the encoder and decoder through an attention mechanism. We propose a new simple network architecture, the Transformer, based solely on attention mechanisms, dispensing with recurrence and convolutions entirely. Experiments on two machine translation tasks show these models to be superior in quality while being more parallelizable and requiring significantly less time to train. Our model achieves 28.4 BLEU on the WMT 2014 English-to-German translation task, improving over the existing best results, including ensembles, by over 2 BLEU. On the WMT 2014 English-to-French translation task, our model establishes a new single-model state-of-the-art BLEU score of 41.8 after training for 3.5 days on eight GPUs, a small fraction of the training costs of the best models from the literature. We show that the Transformer generalizes well to other tasks by applying it successfully to English constituency parsing both with large and limited training data.

2.

Methods

Preprocessing

- Identify speakers



mustaszewski / europarl-extract Public

mustaszewski Update README.md 772fdc0 · 6 years ago 62 Commits

corpora	Corrected speaker list	7 years ago
documentation	Rename corpus stats files in documentation	7 years ago
preprocess	Added script for batch preprocessing	7 years ago
LICENSE.txt	Add license	7 years ago
README.md	Update README.md	6 years ago
disambiguate_speaker_IDs.py	Major update: Added Sentence-Alignment and Refactore...	7 years ago
extract.py	More verbose help message for output cleaning	7 years ago
gale_church.py	Code cleaning	7 years ago

About

No description, website, or topics provided.

Readme
MIT license
Activity
17 stars
1 watching
3 forks
Report repository

Releases

1 tags

Packages

No packages published

Languages

Python 61.9%	Emacs Lisp 13.3%
Perl 11.1%	Shell 5.6%
Smalltalk 3.2%	NewLisp 1.7%
Other 3.2%	

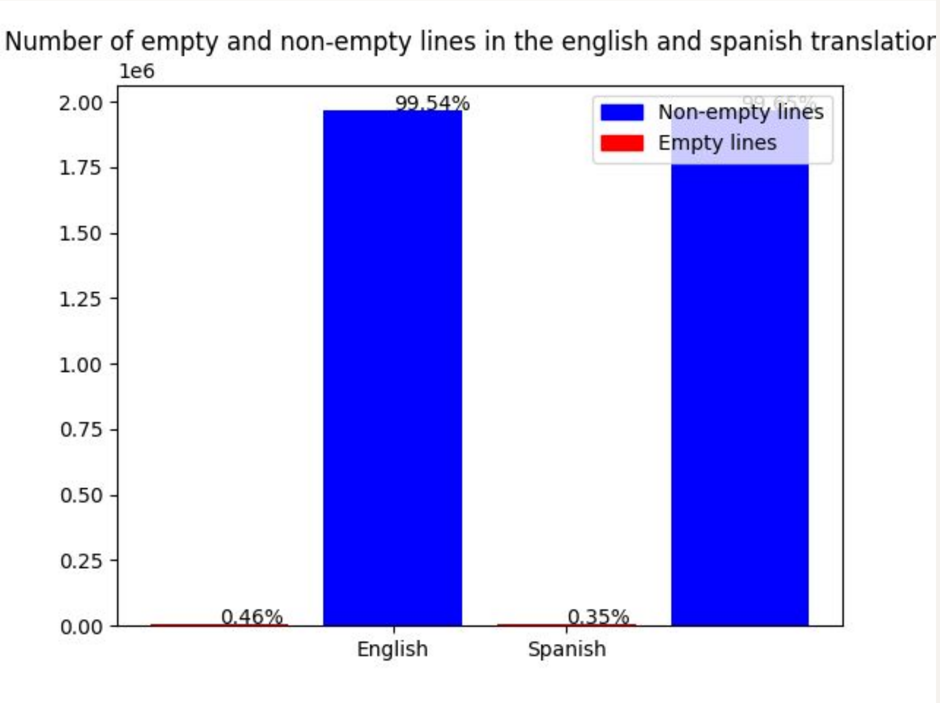
EuroparlExtract

EuroparlExtract is a corpus processing toolkit for the compilation of directional sub-corpora, both bilingual parallel and monoloingual comparable ones, from the original release of the European Parliament Proceedings Parallel Corpus (Europarl) by Philipp Koehn (2005). It therefore aims to maximise the utility of the Europarl corpus, which in its original release does not provide tools to compile directional subcorpora (i.e. corpora where the source and target language are made explicit for each language pair) and is therefore of limited use for linguistic and translation research.

The motivation behind the project is to make the wealth of linguistic data contained in the Europarl corpus easily

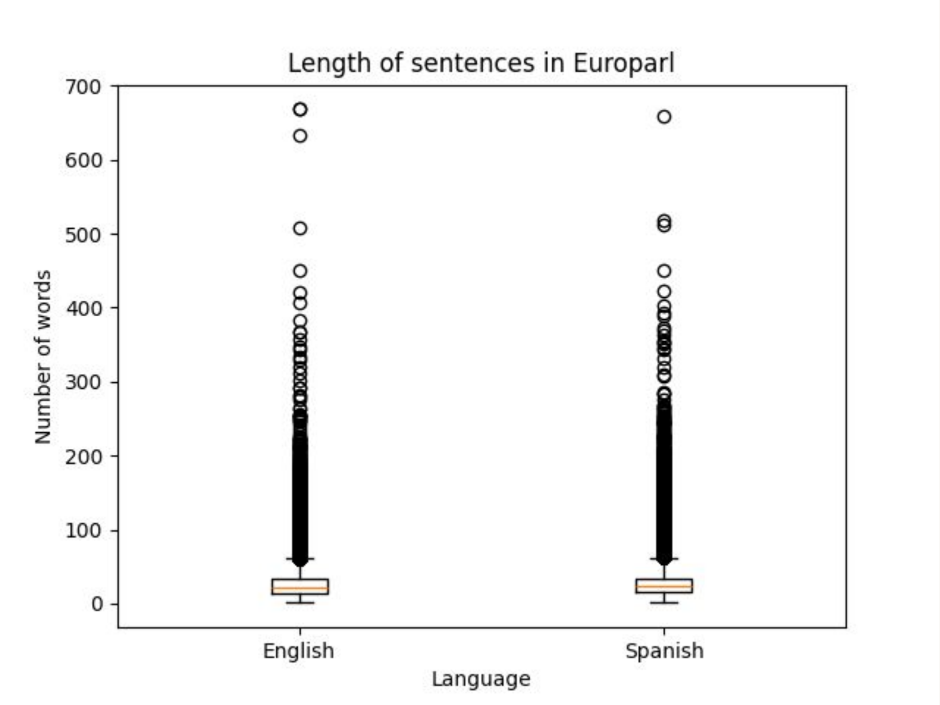
Preprocessing

- Identify speakers
- Manage empty lines



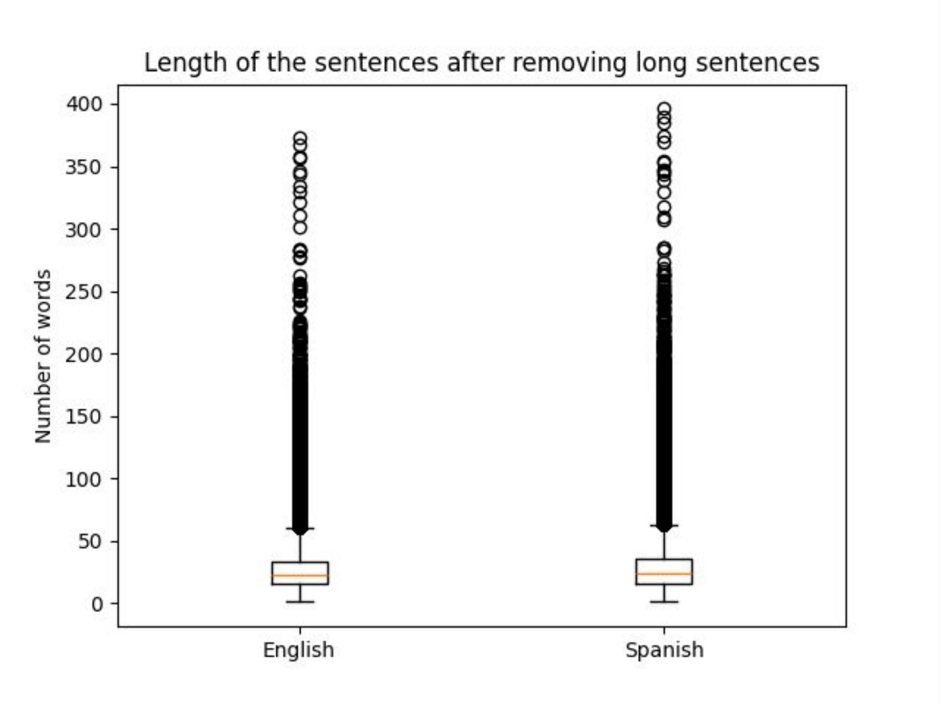
Preprocessing

- Identify speakers
- Manage empty lines
- Feature engineering



Preprocessing

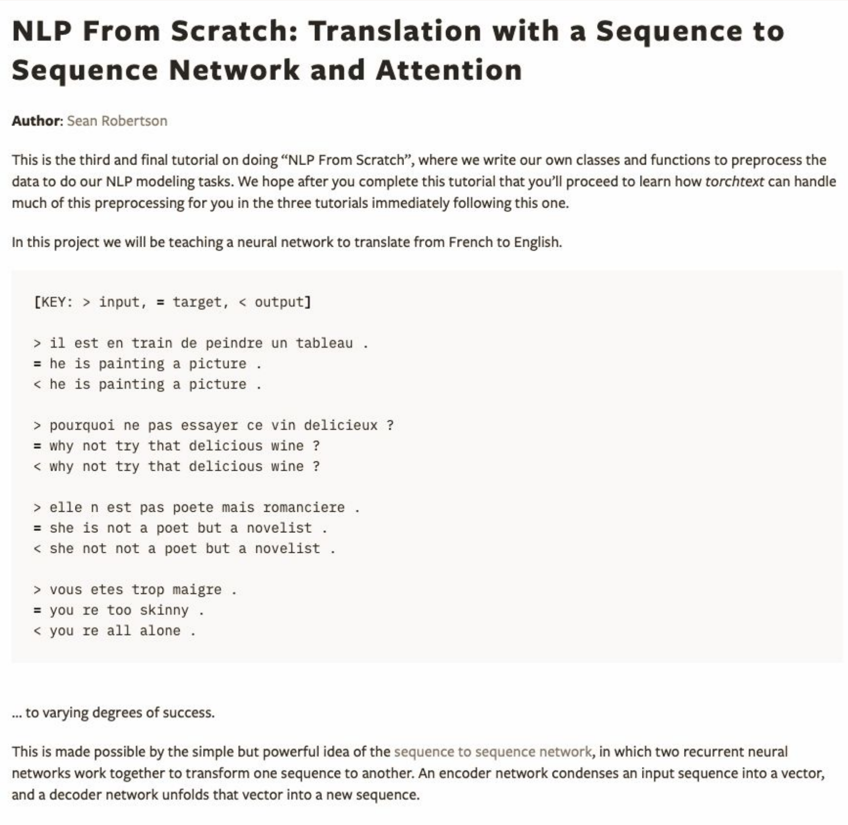
- Identify speakers
- Manage empty lines
- Feature engineering



Sequence to Sequence

Using ‘default’ hyperparameters

because of long training time



NLP From Scratch: Translation with a Sequence to Sequence Network and Attention

Author: Sean Robertson

This is the third and final tutorial on doing “NLP From Scratch”, where we write our own classes and functions to preprocess the data to do our NLP modeling tasks. We hope after you complete this tutorial that you’ll proceed to learn how *torchtext* can handle much of this preprocessing for you in the three tutorials immediately following this one.

In this project we will be teaching a neural network to translate from French to English.

```
[KEY: > input, = target, < output]

> il est en train de peindre un tableau .
= he is painting a picture .
< he is painting a picture .

> pourquoi ne pas essayer ce vin delicieux ?
= why not try that delicious wine ?
< why not try that delicious wine ?

> elle n est pas poete mais romanciere .
= she is not a poet but a novelist .
< she not not a poet but a novelist .

> vous etes trop maigre .
= you re too skinny .
< you re all alone .
```

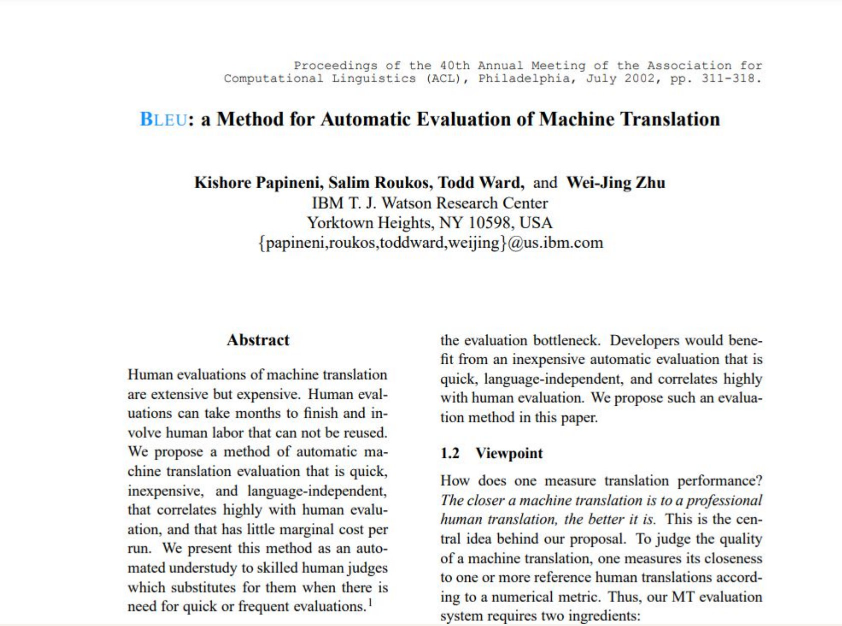
... to varying degrees of success.

This is made possible by the simple but powerful idea of the sequence to sequence network, in which two recurrent neural networks work together to transform one sequence to another. An encoder network condenses an input sequence into a vector, and a decoder network unfolds that vector into a new sequence.

Evaluation metrics

BLEU score

- Quick to understand
- Similar to how humans evaluate translated sentences
- Widely used



Proceedings of the 40th Annual Meeting of the Association for
Computational Linguistics (ACL), Philadelphia, July 2002, pp. 311-318.

BLEU: a Method for Automatic Evaluation of Machine Translation

Kishore Papineni, Salim Roukos, Todd Ward, and **Wei-Jing Zhu**
IBM T. J. Watson Research Center
Yorktown Heights, NY 10598, USA
{papineni,roukos,toddward,weijing}@us.ibm.com

Abstract

Human evaluations of machine translation are extensive but expensive. Human evaluations can take months to finish and involve human labor that can not be reused. We propose a method of automatic machine translation evaluation that is quick, inexpensive, and language-independent, that correlates highly with human evaluation, and that has little marginal cost per run. We present this method as an automated understudy to skilled human judges which substitutes for them when there is need for quick or frequent evaluations.[1]

the evaluation bottleneck. Developers would benefit from an inexpensive automatic evaluation that is quick, language-independent, and correlates highly with human evaluation. We propose such an evaluation method in this paper.

1.2 Viewpoint

How does one measure translation performance? *The closer a machine translation is to a professional human translation, the better it is.* This is the central idea behind our proposal. To judge the quality of a machine translation, one measures its closeness to one or more reference human translations according to a numerical metric. Thus, our MT evaluation system requires two ingredients:

$$Bleu\ (N) = Brevity\ Penalty \cdot Geometric\ Average\ Precision\ Scores\ (N)$$

3.

Results

Partial results

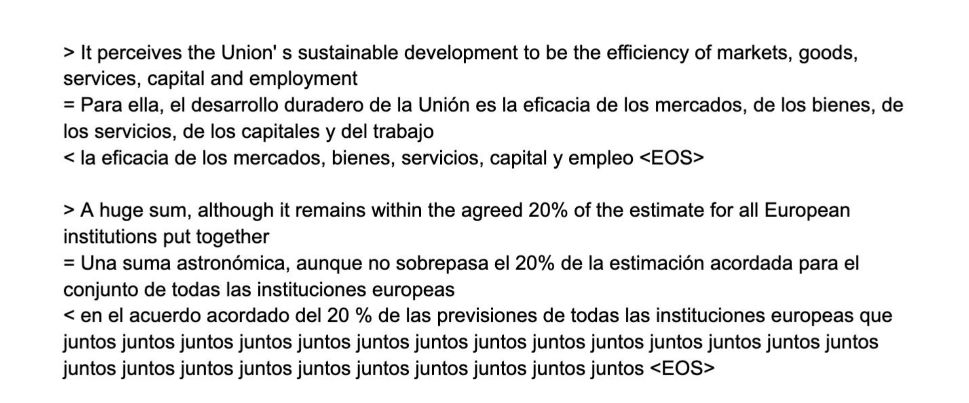
We can evaluate the results in terms of

- Adequacy: Translation w(t) should adequately reflect the linguistic content of w(s)
- Fluency: Translation w(t) should be fluent text in the target language
- BLEU Score

Partial results Bahdanau

We can evaluate the results in terms of

- Adequacy: Translation w(t) should adequately reflect the linguistic content of w(s)
- Fluency: Translation w(t) should be fluent text in the target language
- BLEU Score



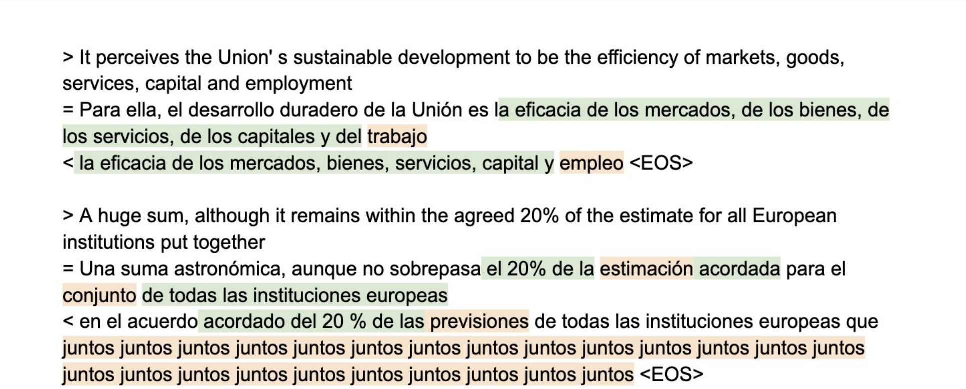
```
> It perceives the Union' s sustainable development to be the efficiency of markets, goods,
services, capital and employment
= Para ella, el desarrollo duradero de la Unión es la eficacia de los mercados, de los bienes, de
los servicios, de los capitales y del trabajo
< la eficacia de los mercados, bienes, servicios, capital y empleo <EOS>

> A huge sum, although it remains within the agreed 20% of the estimate for all European
institutions put together
= Una suma astronómica, aunque no sobrepasa el 20% de la estimación acordada para el
conjunto de todas las instituciones europeas
< en el acuerdo acordado del 20 % de las previsiones de todas las instituciones europeas que
juntos juntos juntos juntos juntos juntos juntos juntos juntos juntos juntos juntos juntos juntos
juntos juntos juntos juntos juntos juntos juntos juntos juntos juntos <EOS>
```

Partial results Bahdanau

We can evaluate the results in terms of

- Adequacy: Translation w(t) should adequately reflect the linguistic content of w(s)
- Fluency: Translation w(t) should be fluent text in the target language



> It perceives the Union' s sustainable development to be the efficiency of markets, goods, services, capital and employment
= Para ella, el desarrollo duradero de la Unión es la eficacia de los mercados, de los bienes, de los servicios, de los capitales y del trabajo
< la eficacia de los mercados, bienes, servicios, capital y empleo <EOS>

> A huge sum, although it remains within the agreed 20% of the estimate for all European institutions put together
= Una suma astronómica, aunque no sobrepasa el 20% de la estimación acordada para el conjunto de todas las instituciones europeas
< en el acuerdo acordado del 20 % de las previsiones de todas las instituciones europeas que juntos <EOS>

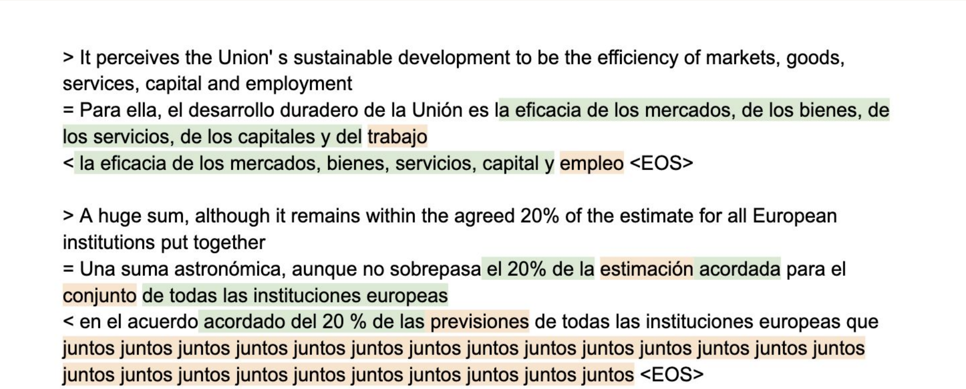
In orange, synonyms, in green coincidences between true translation and model prediction

Partial results Bahdanau

We can evaluate the results in terms of

- Adequacy: Translation w(t) should adequately reflect the linguistic content of w(s)
- Fluency: Translation w(t) should be fluent text in the target language
- BLEU Score: between 0 and 1 out of 100 pairs

Bahdanau attention:
3.99967114868466e-06

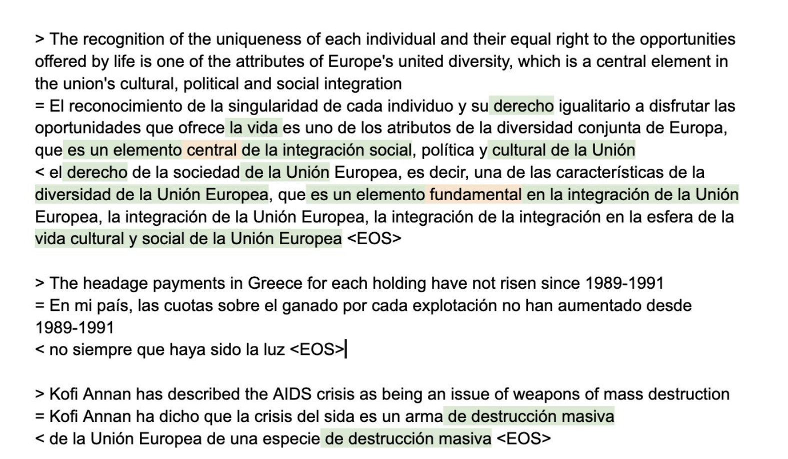


> It perceives the Union' s sustainable development to be the efficiency of markets, goods, services, capital and employment
= Para ella, el desarrollo duradero de la Unión es la eficacia de los mercados, de los bienes, de los servicios, de los capitales y del trabajo
< la eficacia de los mercados, bienes, servicios, capital y empleo <EOS>

> A huge sum, although it remains within the agreed 20% of the estimate for all European institutions put together
= Una suma astronómica, aunque no sobrepasa el 20% de la estimación acordada para el conjunto de todas las instituciones europeas
< en el acuerdo acordado del 20 % de las previsiones de todas las instituciones europeas que juntos <EOS>

In orange, synonyms, in green coincidences between true translation and model prediction

Full results Multihead & Bahdanau

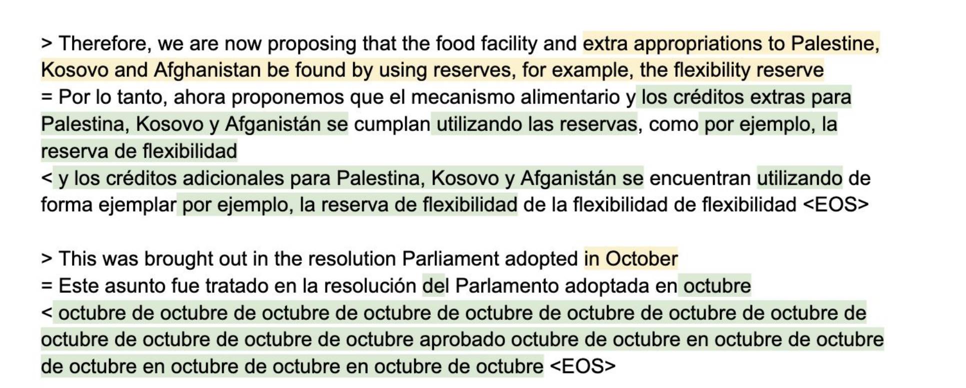


> The recognition of the uniqueness of each individual and their equal right to the opportunities offered by life is one of the attributes of Europe's united diversity, which is a central element in the union's cultural, political and social integration
= El reconocimiento de la singularidad de cada individuo y su derecho igualitario a disfrutar las oportunidades que ofrece la vida es uno de los atributos de la diversidad conjunta de Europa, que es un elemento central de la integración social, política y cultural de la Unión
< el derecho de la sociedad de la Unión Europea, es decir, una de las características de la diversidad de la Unión Europea, que es un elemento fundamental en la integración de la Unión Europea, la integración de la Unión Europea, la integración de la integración en la esfera de la vida cultural y social de la Unión Europea <EOS>

> The headage payments in Greece for each holding have not risen since 1989-1991
= En mi país, las cuotas sobre el ganado por cada explotación no han aumentado desde 1989-1991
< no siempre que haya sido la luz <EOS>

> Kofi Annan has described the AIDS crisis as being an issue of weapons of mass destruction
= Kofi Annan ha dicho que la crisis del sida es un arma de destrucción masiva
< de la Unión Europea de una especie de destrucción masiva <EOS>

In orange, synonyms, in green coincidences between true translation and model prediction, in yellow original input correctly translated in model prediction



> Therefore, we are now proposing that the food facility and extra appropriations to Palestine, Kosovo and Afghanistan be found by using reserves, for example, the flexibility reserve
= Por lo tanto, ahora proponemos que el mecanismo alimentario y los créditos extras para Palestina, Kosovo y Afganistán se cumplan utilizando las reservas, como por ejemplo, la reserva de flexibilidad
< y los créditos adicionales para Palestina, Kosovo y Afganistán se encuentran utilizando de forma ejemplar por ejemplo, la reserva de flexibilidad de la flexibilidad de flexibilidad <EOS>

> This was brought out in the resolution Parliament adopted in October
= Este asunto fue tratado en la resolución del Parlamento adoptada en octubre
< octubre de octubre de octubre de octubre de octubre de octubre de octubre de octubre de octubre de octubre de octubre de octubre aprobado octubre de octubre en octubre de octubre de octubre en octubre de octubre en octubre de octubre <EOS>

In orange, synonyms, in green coincidences between true translation and model prediction, in yellow original input correctly translated in model prediction

BLEU Score:

Bahdanau attention:
3.99967114868466e-06

BLEU Score:

Multi-head attention:
1.7302134354230151e-06

4.

Discussion

Conclusion & Limitations

- It needs a lot of computational power to train
- Not reliable (the BLEU scores are low)
- Limited context understanding
- Longer sentences are more difficult to translate accurately

Potential Fixes

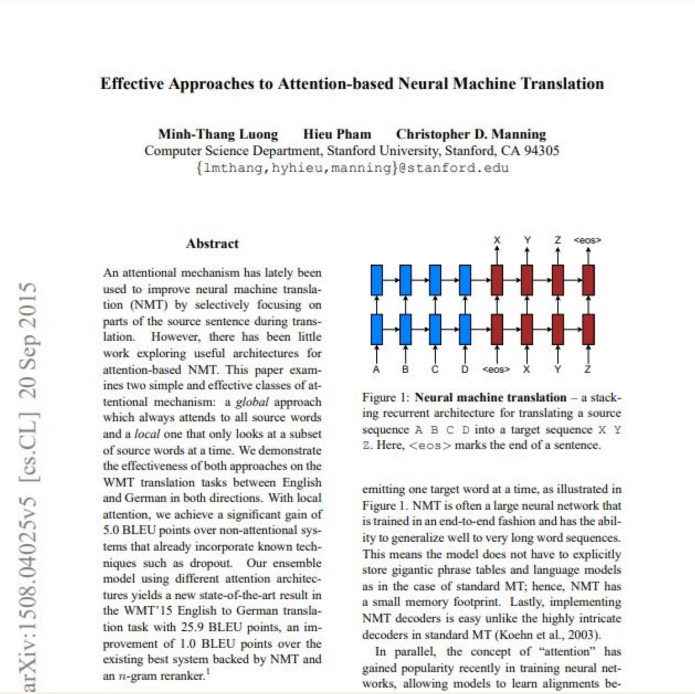
- Limit the sentence length
  - + The training will be faster, so we will be able to train more epochs
  - - We will have to limit the size of the dataset

- Optimize hyperparameters
  - + The models would converge faster
  - - The process is resource intensive

Future (potential) work

- Fixing the network

Future (potential) work



Effective Approaches to Attention-based Neural Machine Translation

Minh-Thang Luong **Hieu Pham** **Christopher D. Manning**
Computer Science Department, Stanford University, Stanford, CA 94305
{lmthang,hyhieu,manning}@stanford.edu

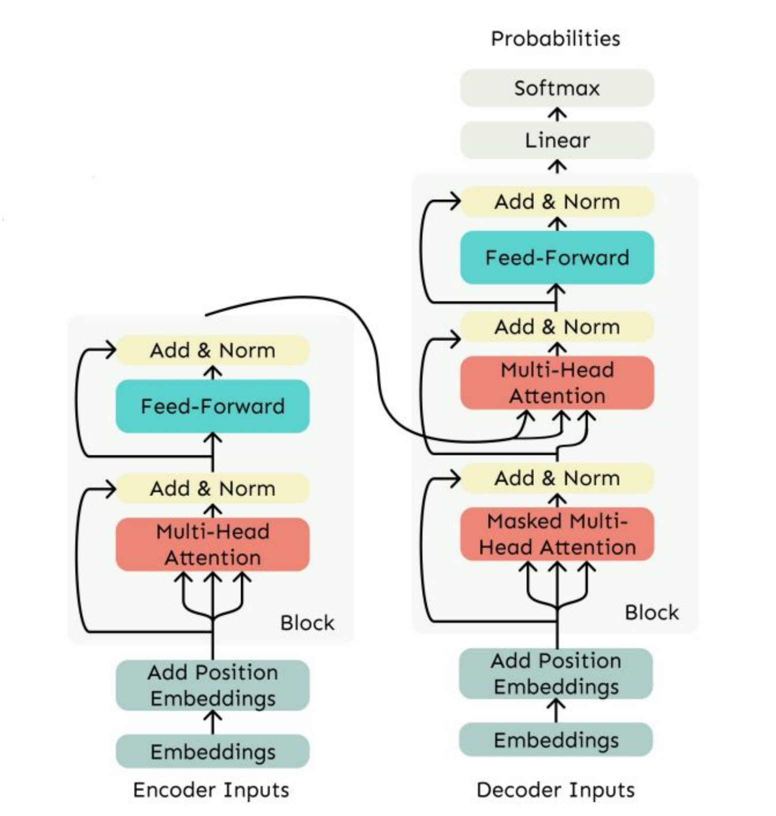
Abstract

An attentional mechanism has lately been used to improve neural machine translation (NMT) by selectively focusing on parts of the source sentence during translation. However, there has been little work exploring useful architectures for attention-based NMT. This paper examines two simple and effective classes of attentional mechanism: a *global* approach which always attends to all source words and a *local* one that only looks at a subset of source words at a time. We demonstrate the effectiveness of both approaches on the WMT translation tasks between English and German in both directions. With local attention, we achieve a significant gain of 5.0 BLEU points over non-attentional systems that already incorporate known techniques such as dropout. Our ensemble model using different attention architectures yields a new state-of-the-art result in the WMT'15 English to German translation task with 25.9 BLEU points, an improvement of 1.0 BLEU points over the existing best system backed by NMT and an n-gram reranker.[1]

Figure 1: **Neural machine translation** – a stacking recurrent architecture for translating a source sequence `A B C D` into a target sequence `X Y Z`. Here, `<eos>` marks the end of a sentence.

emitting one target word at a time, as illustrated in Figure 1. NMT is often a large neural network that is trained in an end-to-end fashion and has the ability to generalize well to very long word sequences. This means the model does not have to explicitly store gigantic phrase tables and language models as in the case of standard MT; hence, NMT has a small memory footprint. Lastly, implementing NMT decoders is easy unlike the highly intricate decoders in standard MT (Koehn et al., 2003).

In parallel, the concept of "attention" has gained popularity recently in training neural networks, allowing models to learn alignments be-



Thank you!