Neural Machine Translation

With a focus on Unsupervised NMT

Niccolò Campolungo

campolungo@di.uniroma1.it
SapienzaNLP
Sapienza University of Rome

Table of Contents

SAPIENZA NLP

Table of Contents

SAPIENZA NLP

Definition

- **Machine translation** is the task of automatically converting source text in one language to text in another language
- One of the oldest problems in NLP
- Initially tackled with rule-based systems, then statistical models, now deep neural networks

SAPIENZA NLP

Table of Contents

SAPIENZA NLP

Rule-based systems

Based on sets of rules and alignment techniques.
These rules:

- Were developed by expert linguists;
- Took into account lexical, syntactic and semantic levels;
- Were very hard to handle due to the **high number of exceptions** that occurred.

SAPIENZA NLP

Example-based systems

Translation by examples (and bilingual vocabulary).
Example:

- We know that "`I went to the cinema`" → "`sono andato al cinema`";
- Our vocabulary knows that "`theatre`" → "`teatro`";
- Can we translate "`I went to the theatre`"? Yes!

Drawbacks:

- Similar to rule-based (still need linguists!);
- Limited by vocabulary coverage;
- Again, lots and lots of exceptions...

Statistical systems (v2)

Do not *write* rules, let statistical models **learn** them via supervision!

- ✓ No linguists needed (... mostly :));
- ✓ No intermediate tools (syntactic parsing, parallel vocabularies, etc);
- ✓ Only requirement: parallel data!
- ✗ Need to explicitly decide how to **generate** text;
- ✗ Need lots of data!
- ✗ Seriously... **Millions** of parallel sentences in the targeted language pair.

SAPIENZA NLP

Statistical systems (2)

Formally:

- Given a sentence $\mathcal{X}$ in source language S and a target language T
- Find a sentence $\mathcal{Y}$ in T such that $P(\mathcal{Y}|\mathcal{X})$ is maximized.

Where P is the probability emitted by the statistical model.
Might be word-based, syntax-based, phrase-based...

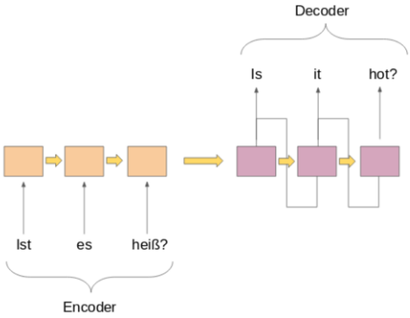
—

A good step towards better translations, but still far from human-level translations for more complex sentences!

Table of Contents

SAPIENZA NLP

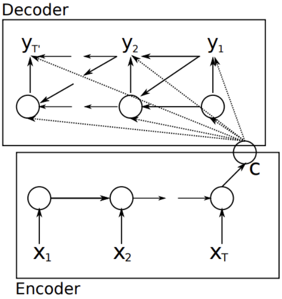
Neural systems: General framework



The two modules are usually jointly trained to maximize the conditional probability $P(\mathcal{Y}|\mathcal{X})$.

Recurrent Neural Networks [Cho et al., 2014]

- First seq2seq architecture!
- Encoder encodes the whole sentence as the last hidden vector of the RNN.

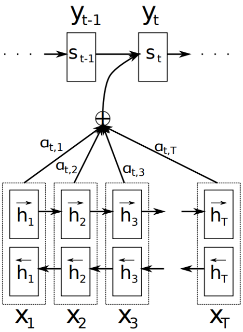


Drawbacks

1. Sentence is represented just by last vector of the RNN;
2. Network representation loses expressiveness due to the long-term dependency problem and information compression.

Attentive Seq2Seq [Bahdanau et al., 2014]

- First attentive seq2seq architecture, paved the way for many more;
- Performs attention over the encoder hidden states for each decoding step.



Benefits

1. A context vector is computed for each decoding step, performing attention over all the hidden states produced by the encoder;
2. Weighting is learned and performed by the network!

Transformer [Vaswani et al., 2017]

- First fully-attentive seq2seq architecture;
- Encoder: series of self-attention blocks followed by feedforward networks;
- Decoder: same as encoder, but with a cross-attention module over encoder output, and basic self-attention is now causally masked (can't look at next tokens);

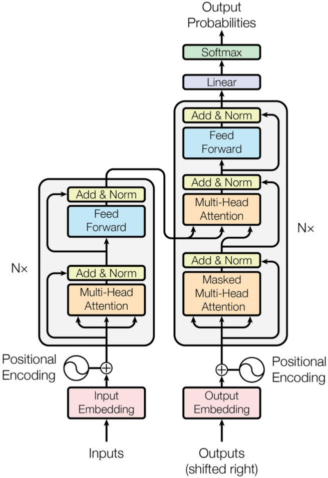


Figure 1: The Transformer - model architecture.

Transformer (2)

- **Many** stacked layers form the full architecture;
- Uses SentencePiece (subwords) instead of plain tokens;
- Decoder performs cross-attention over encoder output, hence at subword level!
- Pushes SotA in a **wide variety** of Natural Language Generation tasks.

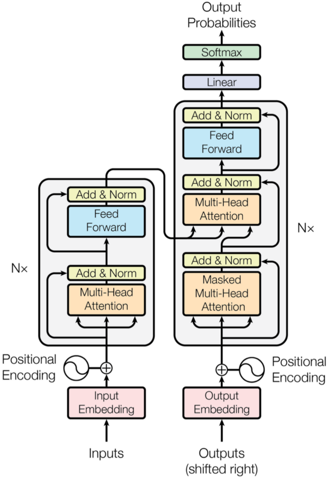


Figure 1: The Transformer - model architecture.

Table of Contents

SAPIENZA NLP

Unsupervised setting

Standard setting: parallel dataset $\mathcal{D} = \{(x, y) \mid \forall\ x \in \mathcal{D}_x, y \in \mathcal{D}_y\}$, where y is a translation in language T of x (written in language S).

Let's now consider some other dataset $\hat{\mathcal{D}}_x$ of sentences in language S. We can assume that there exists a dataset $\hat{\mathcal{D}}_y$ of sentences in language T which are translations of each $x \in \hat{\mathcal{D}}_x$.

—

Problem: we **don't** have access to $\hat{\mathcal{D}}_y$!

SAPIENZA NLP

Table of Contents

SAPIENZA NLP

Definition

Back-translation

A process where a translated text is re-translated back into the source language, without direct access to the original text.

Definition

Back-translation

A process where a translated text is re-translated back into the source language, without direct access to the original text.

The presence of discrepancies between the back-translated text and the original one is an *indication* of **translation errors** in the target language.

SAPIENZA NLP

Definition

Back-translation

A process where a translated text is re-translated back into the source language, without direct access to the original text.

The presence of discrepancies between the back-translated text and the original one is an *indication* of **translation errors** in the target language.

How do we apply this to NMT?

Definition

Back-translation

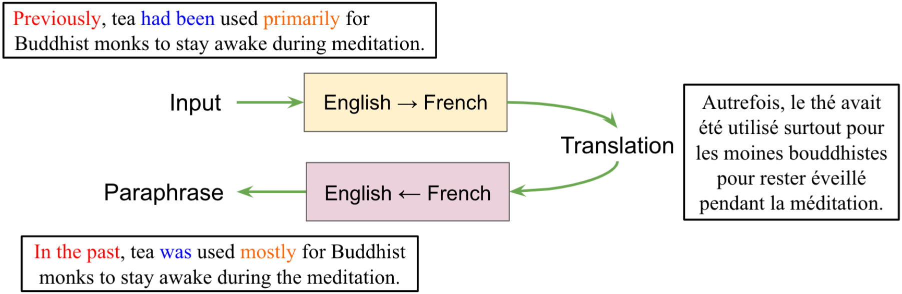
A process where a translated text is re-translated back into the source language, without direct access to the original text.

The presence of discrepancies between the back-translated text and the original one is an *indication* of **translation errors** in the target language.

How do we apply this to NMT? Minimize the **reconstruction error** from the back-translated sentence and the original one.

SAPIENZA NLP

Back-translation Example



SAPIENZA NLP

Table of Contents

First appearance [Sennrich et al., 2016]

- First introduced as a data augmentation technique to perform **semi-supervised** learning;
- Back-translated data and gold parallel data were treated exactly the same;
- Attentive RNN-based architecture;
- Achieved SotA performances on standard MT datasets, beating supervised systems (25 BLEU EN-FR).

More recently...

Same principle applied with advanced decoding algorithms (and larger monolingual corpora) achieves impressive results (45 BLEU EN-FR!) [Edunov et al., 2018]

SAPIENZA NLP

Joint fine-tuning [Lample and Conneau, 2019]

XLM – NeurIPS2019

- Transformer-based architecture;
- Encoders were pre-trained on large multilingual corpora (no parallel data) through Masked Language Modeling;
- Back-translation as a fine-tuning technique on pre-trained encoders;
- Jointly fine-tuned on denoising auto-encoding and **online** back-translation.

SAPIENZA NLP

Monolingual pre-training [Song et al., 2019]

MASS – ICML2019

- Pre-train **whole** Transformer on monolingual data through MASS (random input segment masked and predicted);
- Fine-tune through online back-translation **alone**;
- Results: SotA in UNMT! Not as good as supervised systems (37 vs 41 BLEU EN-FR)

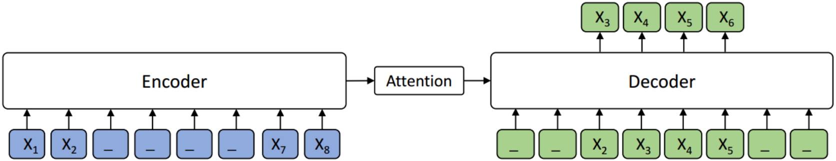


Table of Contents

SAPIENZA NLP

Data-dependent Gaussian Prior objective [Li et al., 2020]

D2GPo – ICLR2020

Negative diversity invariance

Maximum Likelihood Estimation fails to assign proper scores to different incorrect model outputs, which means that all incorrect outputs are treated equally during training.

How to deal with this?

Data-dependent Gaussian Prior objective [Li et al., 2020]

D2GPo – ICLR2020

Negative diversity invariance

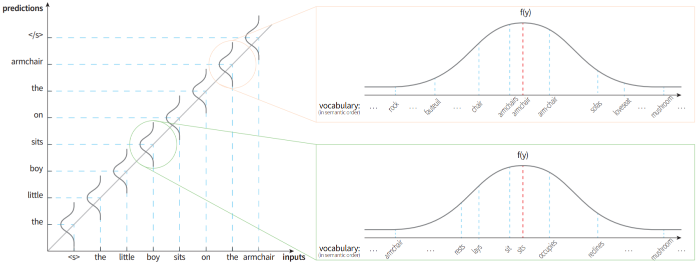
Maximum Likelihood Estimation fails to assign proper scores to different incorrect model outputs, which means that all incorrect outputs are treated equally during training.

How to deal with this?

- Intuition: if correct token is `armchair`, penalize `deckchair` much less than `mushroom`!
- Add a penalization term to the training loss, based on some distance function of each true word to its corresponding prediction.

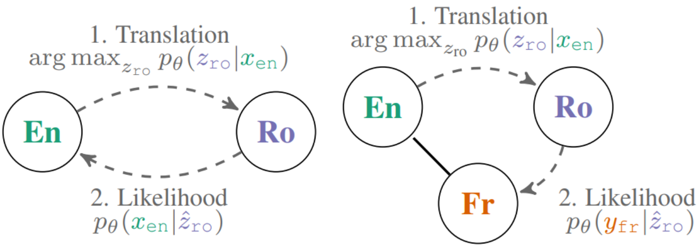
SAPIENZA NLP

- ✓ Very intuitive!
- ✓ Consistently improves SotA on a wide range of tasks;
- ✗ Strictly vocabulary-specific! (even at subword level)



Introducing cross-translation

Instead of simply back-translating, exploit parallel data to perform **cross-translation**.



(a) Back-translation (b) Cross-translation

Multilingual Unsupervised NMT (2)

- ✓ Elegant extension of back-translation;
- ✓ Easily extended to multiple languages;
- ✓ Exploit high-resource languages to train low-resource ones!
- ✗ Requires parallel data.

SAPIENZA NLP

Multilingual Unsupervised NMT (2)

- ✓ Elegant extension of back-translation;
- ✓ Easily extended to multiple languages;
- ✓ Exploit high-resource languages to train low-resource ones!
- ✗ Requires parallel data.

	En – Fr	Fr – En	En – De	De – En	En – Ro	Ro – En
Models without auxiliary parallel data						
XLM (Lample & Conneau, 2019)	33.4	33.3	27.0	34.3	33.3	31.8
MASS (Song et al., 2019)	37.50	34.90	28.30	35.20	35.20	33.10
D2GPo (Li et al., 2020)	37.92	34.94	28.42	35.62	36.31	33.41
Artetxe et al. (2019)	36.2	33.5	26.9	34.4	-	-
Ren et al. (2019)	35.4	34.9	27.7	35.6	34.9	34.1
mBART (Liu et al., 2020)[5]	-	-	**29.8**	34.0	35.0	30.5
M-UNMT	36.25	33.50	25.47	32.32	34.87	32.10
Models with auxiliary parallel data						
mBART (Liu et al., 2020)	-	-	-	-	-	33.9
M-UNMT (Only Pre-Train)	29.22	33.84	18.33	29.04	25.25	32.64
M-UNMT (Fine-Tuned)	**38.34**	**36.05**	28.73	**35.98**	**37.4**	**35.75**

SAPIENZA NLP

Table of Contents

SAPIENZA NLP

Table of Contents

SAPIENZA NLP

Language Generation Problems [1]

NOTE

These limitations do not concern models performing machine translation **only**, but all models whose end task is to **generate** natural language!

[1]Taken from Li et al. [2020].

Language Generation Problems [1]

NOTE

These limitations do not concern models performing machine translation **only**, but all models whose end task is to **generate** natural language!

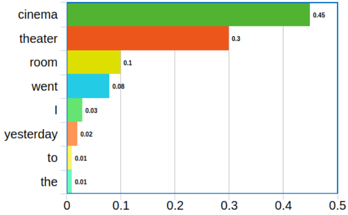
1. Exposure bias: the model is not exposed to the full range of errors during training (due to teacher forcing);
2. Loss mismatch: the model is (usually) trained on token-level loss(es), but evaluation is performed on different metrics;
3. Generation diversity: dull, generic, repetitive, short-sighted generations (i.e. not human-like);
4. Negative diversity invariance: failure to properly score mistakes during training.

[1] Taken from Li et al. [2020].

Decoding example

X: `ieri sono andato al cinema`
Y: `yesterday I went to the...`

How to choose the next token y_5?

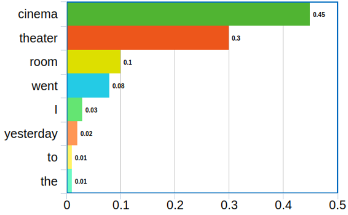


$P(y_i | y_{1:i-1}, X)$
Next-token scores for decoding token after `the`.

Decoding example: Greedy Decoding

X: `ieri sono andato al cinema`
Y: `yesterday I went to the...`

Pick the next most-probable token.
In this case, cinema.



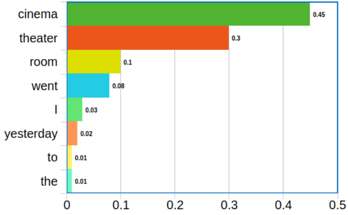
$P(y_i|y_{1:i-1}, X)$
Next-token scores for decoding token after `the`.

SAPIENZA NLP

Decoding example: Pure Sampling

X: `ieri sono andato al cinema`
Y: `yesterday I went to the...`

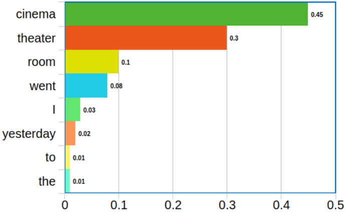
Choose a token from the vocabulary sampling according to the next-token scores. Any token in the vocabulary might be chosen, but pick is based on their probability.



$P(y_i|y_{1:i-1}, X)$
Next-token scores for decoding token after `the`.

X: `ieri sono andato al cinema`
Y: `yesterday I went to the...`

Choose a token among the top-k scoring ones, sampling according to the next-token scores.
Only top scoring tokens can be picked, avoiding garbage outputs like `went` or `yesterday`.



$P(y_i|y_{1:i-1}, X)$
Next-token scores for decoding token after `the`.

SAPIENZA NLP

Decoding example: Top-p Sampling

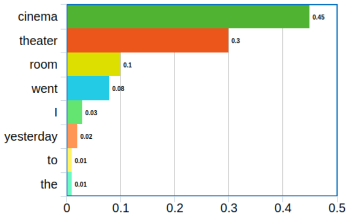
X: `ieri sono andato al cinema`

Y: `yesterday I went to the...`

Choose a token among the top-p scoring ones, sampling according to the next-token scores.

$$\text{Top-}p = min_{|\mathcal{T}|} \sum_{t \in \mathcal{T}} P(t) \geq p$$

More adaptive threshold to restrict token choice compared to top-k.



$P(y_i | y_{1:i-1}, X)$

Next-token scores for decoding token after `the`.

SAPIENZA NLP

Beam search

All previous algorithms have one flaw: they consider **only** next-token prediction.

Solution: keep yourself open to the possibility that one decoding path might become more likely than another **after a few decoding steps**.

- ✓ Every decoding step creates (and prunes) new branches of the *decoding tree*;
- ✓ More likely to avoid sub-optimal solutions;
- ✓ Both path choice and token sampling can be chosen according to different policies!
- ✗ Complexity grows exponentially with beam size;
- ✗ Biased towards shorter sequences (path probability is non-increasing);

NOTE Heuristics (length / coverage penalty) help mitigate problems.

SAPIENZA NLP

Table of Contents

SAPIENZA NLP

Evaluation example

Fundamental problem in automatic translation systems: **how can we evaluate translations?** When is a translation better than another one?

Example

- Source: `ieri sono andato al cinema`
- Reference: `I went to the cinema yesterday`
- Translation: `yesterday I went to the cinema`

What should be the score of this translation?

- BLEU: 79.53/100
- METEOR: 98.14/100
- ROUGE-L: 83.33/100

SAPIENZA NLP

Evaluation example (2)

Fundamental problem in automatic translation systems: **how can we evaluate translations?** When is a translation better than another one?

Example

- Source: `ieri sono andato al cinema`
- Reference: `I went to the cinema yesterday`
- Translation: `I went to yesterday the cinema`

What should be the score of this translation?

- BLEU: **0**/100 (no common 4-grams!)
- METEOR: 93.75/100
- ROUGE-L: 83.33/100

SAPIENZA NLP

How to evaluate translations

Ideally

- Semantically equivalent sentences should have high similarity scores
  - Order of words should not matter
  - **As long as** grammar is not disrupted
- Synonyms should be scored according to their contextual relevance
- Rephrasing / paraphrasing should not be considered a severe mistake

Actually

- Semantically equivalent sentences might have very different similarity scores;
- Changing word order, even when following grammar rules, changes similarity scores;
- Synonyms / rephrasing / paraphrasing might completely break common similarity metrics.

Table of Contents

SAPIENZA NLP

Conclusions

1. In the last few years, translation quality has **dramatically** improved
2. Exciting new unsupervised directions!
3. Unsupervised techniques are incredibly useful when applied jointly with supervised ones
4. We still don't know how to properly extract knowledge from translation models
5. The evaluation suite is poor and models are compared on inadequate metrics

SAPIENZA NLP

Thank you! Questions?

References I

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, 2014.

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. Understanding back-translation at scale. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 489–500, 2018.

Xavier Garcia, Pierre Foret, Thibault Sellam, and Ankur P Parikh. A multilingual view of unsupervised machine translation. *arXiv preprint arXiv:2002.02955*, 2020.

Guillaume Lample and Alexis Conneau. Cross-lingual language model pretraining. *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Zuchao Li, Rui Wang, Kehai Chen, Masso Utiyama, Eiichiro Sumita, Zhuosheng Zhang, and Hai Zhao. Data-dependent gaussian prior objective for language generation. In *International Conference on Learning Representations*, 2020. URL `https://openreview.net/forum?id=S1efxTVYDr`.

SAPIENZA NLP

References II

Rico Sennrich, Barry Haddow, and Alexandra Birch. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, 2016.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. MASS: masked sequence to sequence pre-training for language generation. In *Proc. of ICML*, pages 5926–5936, 2019. URL `http://proceedings.mlr.press/v97/song19d.html`.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

SAPIENZA NLP