Recent Advances In Document-level Neural Machine Translation

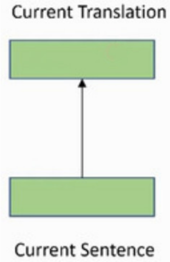
Lorenzo Lupo

Supervisors: Laurent Besacier, Marco Dinarelli

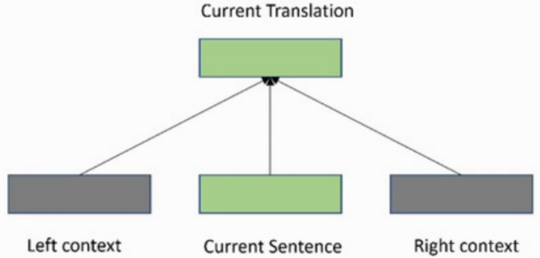
July 9, 2020

What is Document-level Machine Translation

Sentence-level MT

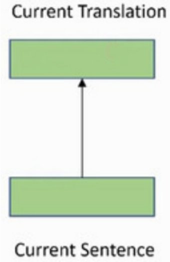


Document-level MT

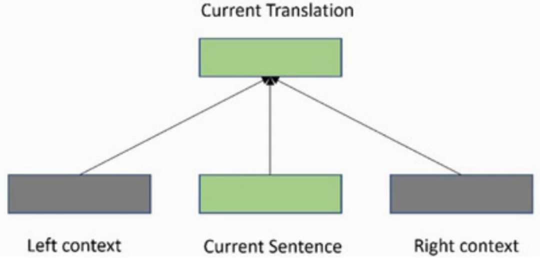


Document-level MT ↔ Context-aware MT

Context-agnostic MT



Context-aware MT



Sentence-level MT is inconsistent

B: Là, ils comprenaient l'importance de la cohésion lexicale.

Sentence-level MT is inconsistent

B: Là, ils comprenaient l'importance de la cohésion lexicale.

SENTENCE-LEVEL TRANSLATION

B: There they understood the importance of lexical cohesion.

Sentence-level MT is inconsistent

A: Nous avons refait l'exercice avec les mêmes etudiants.
Que pensez-vous qu'il est alors arrivé ?

B: Là, ils comprenaient l'importance de la cohésion lexicale.

SENTENCE-LEVEL TRANSLATION

B: There they understood the importance of lexical cohesion.

Sentence-level MT is inconsistent

A: Nous avons refait l'exercice avec les mêmes etudiants.
Que pensez-vous qu'il est alors arrivé ?

B: Là, ils comprenaient l'importance de la cohésion lexicale.

SENTENCE-LEVEL TRANSLATION

B: There they understood the importance of lexical cohesion.

CONTEXT-AWARE TRANSLATION

B: Now, they understood the importance of lexical cohesion.

How bad is it?

[Voita et al., 2019b] undertake a human study on context agnostic translation :

- 2000 pairs of consecutive English sentences (S1 + S2) from OpenSubtitles2018
- translate to Russian with Transformer model [Vaswani et al., 2017]

How bad is it?

[Voita et al., 2019b] undertake a human study on context agnostic translation :

- 2000 pairs of consecutive English sentences (S1 + S2) from OpenSubtitles2018
- translate to Russian with Transformer model [Vaswani et al., 2017]

all	one/both bad	both good	
		bad pair	good pair
2000	211	140	1649
100%	11%	7%	82%

Which kind of inconsistencies?

type of phenomena	frequency
deixis	37%
ellipsis	29%
lexical cohesion	14%
ambiguity	9%
anaphora	6%
other	5%

Figure: Types of phenomena causing discrepancies in context-agnostic translation of consecutive sentences when placed in the context of each other.

Objectives

Objectives

- **Design translation models** that solve discrepancies by taking context into account;

Objectives

- **Design translation models** that solve discrepancies by taking context into account;
- **Evaluate such models** in a proper way;

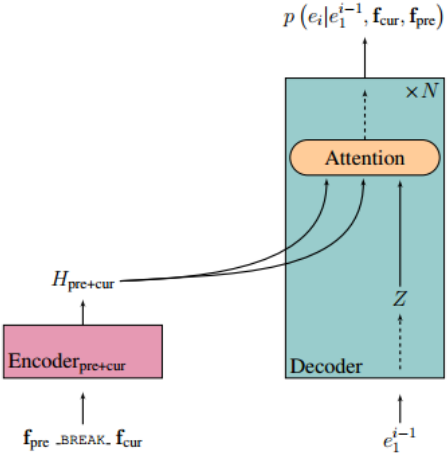
Plan

Plan

Plan

Concatenation Approaches

Concatenation approaches to DLNMT consist in feeding a standard encoder-decoder architecture with a concatenation of sentences.



Concatenation Approaches

For instance:

Concatenation Approaches

For instance:

- [Tiedemann and Scherrer, 2017] firstly introduced this approach proposing an RNN-based model that incorporate the preceding sentence by prepending it to the current one, separated by a <CONCAT> token. They propose two methods:

Concatenation Approaches

For instance:

- [Tiedemann and Scherrer, 2017] firstly introduced this approach proposing an RNN-based model that incorporate the preceding sentence by prepending it to the current one, separated by a <CONCAT> token. They propose two methods:
  - **2-TO-2**: the previous and the current sentences are translated together. The translation of the current sentence is then obtained by only retaining the tokens following the concatenation token.

Concatenation Approaches

For instance:

- [Tiedemann and Scherrer, 2017] firstly introduced this approach proposing an RNN-based model that incorporate the preceding sentence by prepending it to the current one, separated by a <CONCAT> token. They propose two methods:
  - **2-TO-2**: the previous and the current sentences are translated together. The translation of the current sentence is then obtained by only retaining the tokens following the concatenation token.
  - **2-TO-1**: only the current sentence is translated.

Concatenation Approaches

For instance:

- [Tiedemann and Scherrer, 2017] firstly introduced this approach proposing an RNN-based model that incorporate the preceding sentence by prepending it to the current one, separated by a <CONCAT> token. They propose two methods:
  - **2-TO-2**: the previous and the current sentences are translated together. The translation of the current sentence is then obtained by only retaining the tokens following the concatenation token.
  - **2-TO-1**: only the current sentence is translated.
- [Agrawal et al., 2018, Scherrer et al., 2019] investigated the concatenation approach with the Transformer as base model, extending the number of context sentences both on the source (s:-3,+1) and the target (t:-2) side.

Plan

Separate Encoding Approaches

Separate encoding approaches to DLNMT consist in encoder-decoder models that encode the current and context sentences separately. This can be undertaken by:

Separate Encoding Approaches

Separate encoding approaches to DLNMT consist in encoder-decoder models that encode the current and context sentences separately. This can be undertaken by:

- **Multiple encoders** working in parallel for the current and previous sentence. E.g. [Wang et al., 2017].

Separate Encoding Approaches

Separate encoding approaches to DLNMT consist in encoder-decoder models that encode the current and context sentences separately. This can be undertaken by:

- **Multiple encoders** working in parallel for the current and previous sentence. E.g. [Wang et al., 2017].
- **Multiple encoders with shared weights**. In this case, the parallel-working encoders not only have the same architecture, but also the same weights. E.g. [Voita et al., 2018].

Separate Encoding Approaches

Separate encoding approaches to DLNMT consist in encoder-decoder models that encode the current and context sentences separately. This can be undertaken by:

- **Multiple encoders** working in parallel for the current and previous sentence. E.g. [Wang et al., 2017].
- **Multiple encoders with shared weights**. In this case, the parallel-working encoders not only have the same architecture, but also the same weights. E.g. [Voita et al., 2018].
- **Two-pass approaches**, in which the encoder makes a first sentence-level encoding pass of the source, and a second in which it encodes contextual information too. See Slide 20.

Separate Encoding Approaches

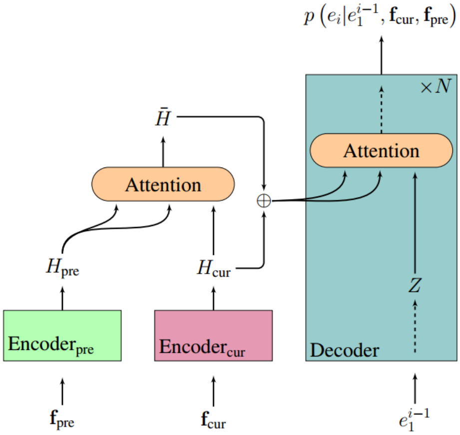
Separate encoding approaches to DLNMT consist in encoder-decoder models that encode the current and context sentences separately. This can be undertaken by:

- **Multiple encoders** working in parallel for the current and previous sentence. E.g. [Wang et al., 2017].
- **Multiple encoders with shared weights**. In this case, the parallel-working encoders not only have the same architecture, but also the same weights. E.g. [Voita et al., 2018].
- **Two-pass approaches**, in which the encoder makes a first sentence-level encoding pass of the source, and a second in which it encodes contextual information too. See Slide 20.
  - Remark: a powerful feature of two-pass approaches is their ability to exploit **future** target-side context.

Separate Encoding Approaches

Once the encoding of the current and the context sentences has been carried out, they can be integrated in different ways:

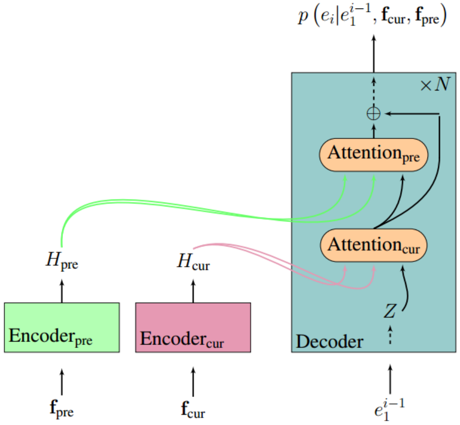
- **Outside** the decoder.
  - $(+)$ symbol represents a gate, a sum or a concatenation.



Separate Encoding Approaches

Once the encoding of the current and the context sentences has been carried out, they can be integrated in different ways:

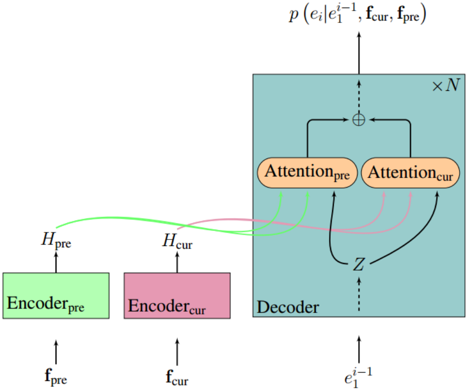
- **Outside** the decoder.
  - $(+)$ symbol represents a gate, a sum or a concatenation.
- **Inside** the decoder, **sequentially**.



Separate Encoding Approaches

Once the encoding of the current and the context sentences has been carried out, they can be integrated in different ways:

- **Outside** the decoder.
  - $(+)$ symbol represents a gate, a sum or a concatenation.
- **Inside** the decoder, **sequentially**.
- **Inside** the decoder, **in parallel**.



Architecture

The encoder-decoder architectures depicted above can be both RNN-based (until 2017) or Transfomer-based (after 2017), as for any approach to DLNMT. However, often some modifications are applied. For example:

Separate Encoding Approaches

Architecture

The encoder-decoder architectures depicted above can be both RNN-based (until 2017) or Transfomer-based (after 2017), as for any approach to DLNMT. However, often some modifications are applied. For example:

- In the case of RNN-based architectures, integration inside the decoder can be undertaken without attention by simply concatenating context representations to the cell state of the deocdrr's RNN [Wang et al., 2017].

Separate Encoding Approaches

Architecture

The encoder-decoder architectures depicted above can be both RNN-based (until 2017) or Transfomer-based (after 2017), as for any approach to DLNMT. However, often some modifications are applied. For example:

- In the case of RNN-based architectures, integration inside the decoder can be undertaken without attention by simply concatenating context representations to the cell state of the deocdrr's RNN [Wang et al., 2017].
- Beside contextual representation of words, the context encoder can also generate higher level representations such as sentence or document embeddings. This representations can also be attended by the decoder [Miculicich et al., 2018, Maruf et al., 2019a] or added to the word-representations [Tan et al., 2019].

Separate Encoding Approaches

Architecture

The encoder-decoder architectures depicted above can be both RNN-based (until 2017) or Transfomer-based (after 2017), as for any approach to DLNMT. However, often some modifications are applied. For example:

- In the case of RNN-based architectures, integration inside the decoder can be undertaken without attention by simply concatenating context representations to the cell state of the deocdrr's RNN [Wang et al., 2017].
- Beside contextual representation of words, the context encoder can also generate higher level representations such as sentence or document embeddings. This representations can also be attended by the decoder [Miculicich et al., 2018, Maruf et al., 2019a] or added to the word-representations [Tan et al., 2019].
- Parallel integration inside the decoder can also happen within a single multi-head attention that takes as values and queries the concatenations of the current and context sentence representations [Voita et al., 2019b]

Separate Encoding Approaches

Including target-side context

Despite some have considered including past target-side context harmful because of the *error propagation* problem [Zhang et al., 2018], most recent works have showed it to be of utmost importance for making the most out of context. Past works have successfully included target-side context information in different ways:

Separate Encoding Approaches

Including target-side context
Despite some have considered including past target-side context harmful because of the *error propagation* problem [Zhang et al., 2018], most recent works have showed it to be of utmost importance for making the most out of context. Past works have successfully included target-side context information in different ways:

- Translating past sentences (usually 1) along with the current one, and then discarding them, as in concatenation approaches [Bawden et al., 2018].

Separate Encoding Approaches

Including target-side context

Despite some have considered including past target-side context harmful because of the *error propagation* problem [Zhang et al., 2018], most recent works have showed it to be of utmost importance for making the most out of context. Past works have successfully included target-side context information in different ways:

- Translating past sentences (usually 1) along with the current one, and then discarding them, as in concatenation approaches [Bawden et al., 2018].
- By making the decoder attend the target-side hidden representations or embeddings of previously decoded sentences [Miculicich et al., 2018, Voita et al., 2019b, Maruf et al., 2019a, Zheng et al., 2020].

Separate Encoding Approaches

Reference	Context	Two-Pass Approach	Outside Integr.	Inside Integr.	Lang. Pair
[Wang et al., 2017]	s:-3		aut...	...aut	Zh→En
[Voita et al., 2018]	s:-1		yes		En→Ru
[Zhang et al., 2018]	s:-2		yes	sequential	Zh→En
[Miculicich et al., 2018]	s:-3; t:-3		yes		Zh/Es→En
[Maruf et al., 2019a]	s:all; t:all	optional	yes		En→De
[Zheng et al., 2020]	s:all; t:all	yes	yes		Zh/En→En/De
[Jean et al., 2017]	s:-1			parallel	En→De/Fr
[Bawden et al., 2018]	s:-1; t:-1			parallel	En→Fr
[Fu et al., 2019]	s:all	yes		parallel	En/Zh→De/En
[Voita et al., 2019b]	s:-3; t:-3	yes		parallel*	En→Ru
[Tan et al., 2019]	s:all	yes		parallel	Zh/De→En
[Wang et al., 2019]	s:2			sequential	Fr→En

Positional Embedding Schema

For many approaches to DNMT, the standard positional encoding proposed by [Vaswani et al., 2017] is insufficient because the DNMT system needs to tell context sentences from the current one. For this reason, many strategies have been proposed in the literature, such as:

Positional Embedding Schema

For many approaches to DNMT, the standard positional encoding proposed by [Vaswani et al., 2017] is insufficient because the DNMT system needs to tell context sentences from the current one. For this reason, many strategies have been proposed in the literature, such as:

1. Adding a **sentence distance embedding** to context sentences, that tell the model how far away they are from the current sentence [Voita et al., 2019b].

Positional Embedding Schema

For many approaches to DNMT, the standard positional encoding proposed by [Vaswani et al., 2017] is insufficient because the DNMT system needs to tell context sentences from the current one. For this reason, many strategies have been proposed in the literature, such as:

1. Adding a **sentence distance embedding** to context sentences, that tell the model how far away they are from the current sentence [Voita et al., 2019b].
2. Assign **positional embeddings progressively** to the current sentence, then to the previous one, and so on, so that far away sentences have high values of positional embedding [Li et al., 2019].

Positional Embedding Schema

For many approaches to DNMT, the standard positional encoding proposed by [Vaswani et al., 2017] is insufficient because the DNMT system needs to tell context sentences from the current one. For this reason, many strategies have been proposed in the literature, such as:

1. Adding a **sentence distance embedding** to context sentences, that tell the model how far away they are from the current sentence [Voita et al., 2019b].
2. Assign **positional embeddings progressively** to the current sentence, then to the previous one, and so on, so that far away sentences have high values of positional embedding [Li et al., 2019].
3. Adding a **segment embedding**, similar to classical positional encoding but for the position of the sentence/segment within the document [Zheng et al., 2020].

Plan

Cache Approaches

Cache approaches to DLNMT consist in encoder-decoder models that are equipped with one or more caches that store context information. The information stored can belong to both **source side or target side, past and future**.

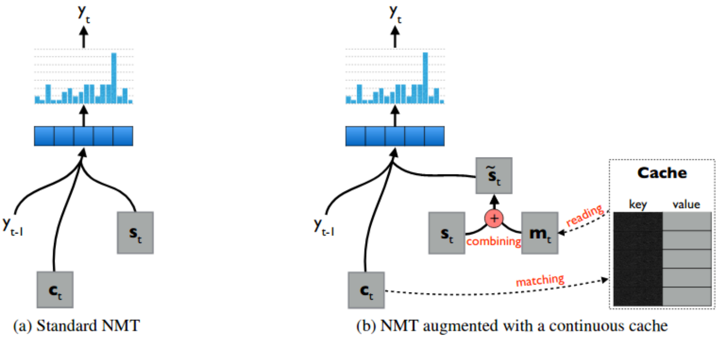
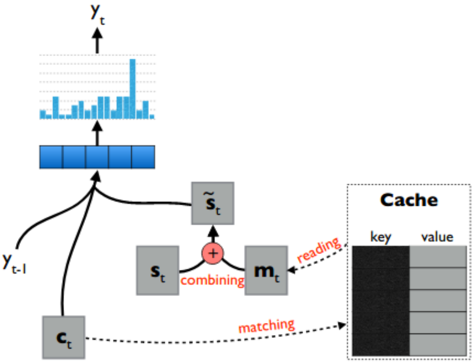
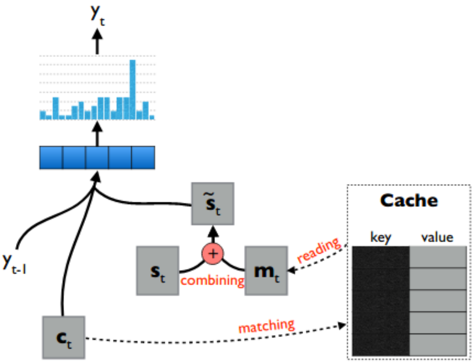


Figure: Continuous cache by [Tu et al., 2017]

Every cache slot is a **key-value-indicator** triplet (the key and the indicator are often the same thing). With these variables, we can **read** or **write** caches.



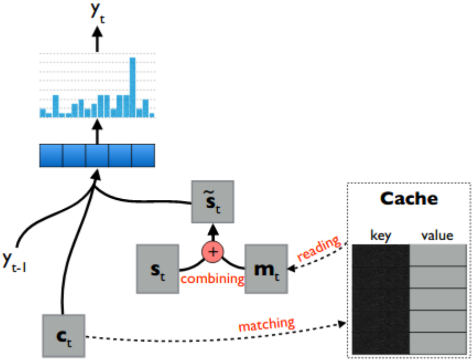
Cache reading involves:



Cache Approaches

Cache reading involves:

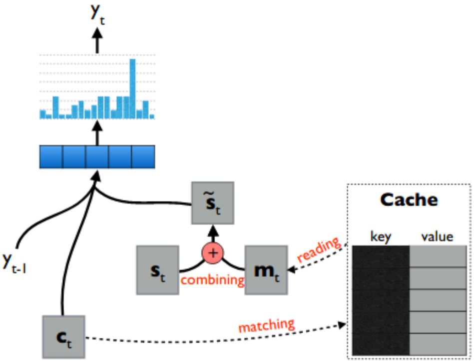
- Soft key matching



Cache Approaches

Cache reading involves:

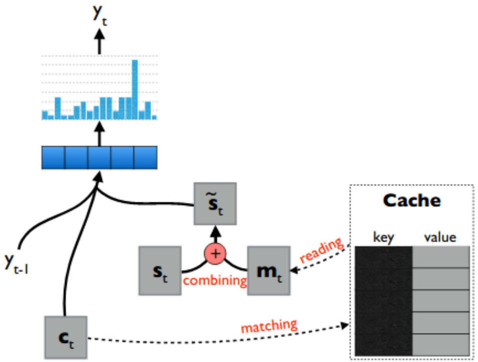
- Soft key matching
- Value reading



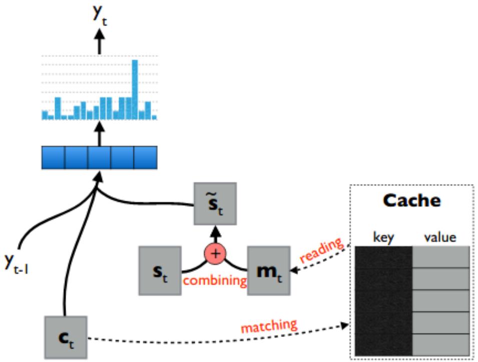
Cache Approaches

Cache reading involves:

- Soft key matching
- Value reading
- Combining



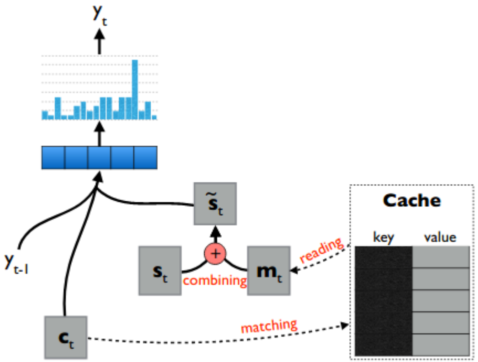
Cache writing can be undertaken after having translated one or more sentences. For every triplet:



Cache Approaches

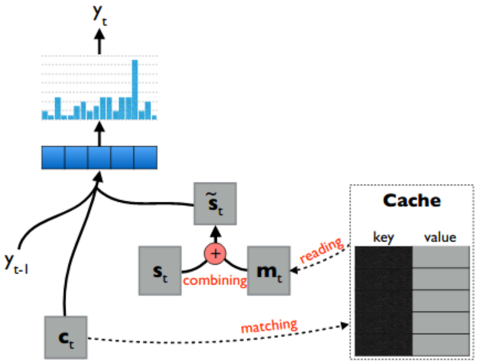
Cache writing can be undertaken after having translated one or more sentences. For every triplet:

- If the **indicator** already exists in the cache, we just update it's keys and values.



Cache writing can be undertaken after having translated one or more sentences. For every triplet:

- If the **indicator** already exists in the cache, we just update it's keys and values.
- Else, we write the triplet in an empty slot, after having emptied the oldest slot if the cache is full.



Cache Approaches

Reference	Caches	Size	Key (Indic.)	Value	Lang. Pair
[Tu et al., 2017]	single	$\leqslant$ 500	$c_t(\ y_{k<t})$	$s_{k<t}$	Zh→En
[Kuang et al., 2018]	dynamic topic	100 200	c_t	$y_{k<t}$ topic emb.	Zh→En
[Maruf and Haffari, 2018]	source target	doc.size	h_t s_t	$sent.emb.$ $s_{k<t}$	Fr/De/Et→En

Plan

On Parallel Corpora for Training

DLNMT systems require training on **document-level parallel corpora**. These corpora are usually released during workshops on machine translation like IWSLT and WMT, and hosted on open source web inventories. The most common ones, are extracted from:

On Parallel Corpora for Training

DLNMT systems require training on **document-level parallel corpora**. These corpora are usually released during workshops on machine translation like IWSLT and WMT, and hosted on open source web inventories. The most common ones, are extracted from:

- Movie subtitles (OpenSubtitles)

On Parallel Corpora for Training

DLNMT systems require training on **document-level parallel corpora**. These corpora are usually released during workshops on machine translation like IWSLT and WMT, and hosted on open source web inventories. The most common ones, are extracted from:

- Movie subtitles (OpenSubtitles)
- TED talks (WIT3)

On Parallel Corpora for Training

DLNMT systems require training on **document-level parallel corpora**. These corpora are usually released during workshops on machine translation like IWSLT and WMT, and hosted on open source web inventories. The most common ones, are extracted from:

- Movie subtitles (OpenSubtitles)
- TED talks (WIT3)
- News articles (LDC)

On Parallel Corpora for Training

DLNMT systems require training on **document-level parallel corpora**. These corpora are usually released during workshops on machine translation like IWSLT and WMT, and hosted on open source web inventories. The most common ones, are extracted from:

- Movie subtitles (OpenSubtitles)
- TED talks (WIT3)
- News articles (LDC)
- Parliamentary interventions (Europarl)

On Parallel Corpora for Training

On Parallel Corpora for Training

- Unfortunately, document-level parallel corpora are often insufficient to train DLNMT systems from scratch, although it is often possible to make them converge to a local optimum.

On Parallel Corpora for Training

- Unfortunately, document-level parallel corpora are often insufficient to train DLNMT systems from scratch, although it is often possible to make them converge to a local optimum.
- [Kim et al., 2019] pointed out that when constraining training on such small datasets, model comparison becomes misleading because gains in performance are mainly related to better regularization.

On Parallel Corpora for Training

- Unfortunately, document-level parallel corpora are often insufficient to train DLNMT systems from scratch, although it is often possible to make them converge to a local optimum.
- [Kim et al., 2019] pointed out that when constraining training on such small datasets, model comparison becomes misleading because gains in performance are mainly related to better regularization.
- A popular solution to this problem is the **two-step training strategy**: [Tu et al., 2017, Zhang et al., 2018, Miculicich et al., 2018]

On Parallel Corpora for Training

- Unfortunately, document-level parallel corpora are often insufficient to train DLNMT systems from scratch, although it is often possible to make them converge to a local optimum.
- [Kim et al., 2019] pointed out that when constraining training on such small datasets, model comparison becomes misleading because gains in performance are mainly related to better regularization.
- A popular solution to this problem is the **two-step training strategy**: [Tu et al., 2017, Zhang et al., 2018, Miculicich et al., 2018]
  1. Distinguish two integrated components in your model with params $\Theta = [\theta_S; \theta_D]$:

On Parallel Corpora for Training

- Unfortunately, document-level parallel corpora are often insufficient to train DLNMT systems from scratch, although it is often possible to make them converge to a local optimum.
- [Kim et al., 2019] pointed out that when constraining training on such small datasets, model comparison becomes misleading because gains in performance are mainly related to better regularization.
- A popular solution to this problem is the **two-step training strategy**: [Tu et al., 2017, Zhang et al., 2018, Miculicich et al., 2018]
  1. Distinguish two integrated components in your model with params $\Theta = [\theta_S; \theta_D]$:
     - A self-standing sentence-level NMT system with parameters θ_S.

On Parallel Corpora for Training

- Unfortunately, document-level parallel corpora are often insufficient to train DLNMT systems from scratch, although it is often possible to make them converge to a local optimum.
- [Kim et al., 2019] pointed out that when constraining training on such small datasets, model comparison becomes misleading because gains in performance are mainly related to better regularization.
- A popular solution to this problem is the **two-step training strategy**: [Tu et al., 2017, Zhang et al., 2018, Miculicich et al., 2018]
  1. Distinguish two integrated components in your model with params $\Theta = [\theta_S; \theta_D]$:
     - A self-standing sentence-level NMT system with parameters θ_S.
     - Some context-handling modules with parameters θ_D.

On Parallel Corpora for Training

- Unfortunately, document-level parallel corpora are often insufficient to train DLNMT systems from scratch, although it is often possible to make them converge to a local optimum.
- [Kim et al., 2019] pointed out that when constraining training on such small datasets, model comparison becomes misleading because gains in performance are mainly related to better regularization.
- A popular solution to this problem is the **two-step training strategy**: [Tu et al., 2017, Zhang et al., 2018, Miculicich et al., 2018]
  1. Distinguish two integrated components in your model with params $\Theta = [\theta_S; \theta_D]$:
     - A self-standing sentence-level NMT system with parameters θ_S.
     - Some context-handling modules with parameters θ_D.
  2. Train θ_S independently on a sentence-level parallel corpus C_S.

On Parallel Corpora for Training

- Unfortunately, document-level parallel corpora are often insufficient to train DLNMT systems from scratch, although it is often possible to make them converge to a local optimum.
- [Kim et al., 2019] pointed out that when constraining training on such small datasets, model comparison becomes misleading because gains in performance are mainly related to better regularization.
- A popular solution to this problem is the **two-step training strategy**: [Tu et al., 2017, Zhang et al., 2018, Miculicich et al., 2018]
  1. Distinguish two integrated components in your model with params $\Theta = [\theta_S; \theta_D]$:
     - A self-standing sentence-level NMT system with parameters θ_S.
     - Some context-handling modules with parameters θ_D.
  2. Train θ_S independently on a sentence-level parallel corpus C_S.
  3. Train θ_D on a document-level parallel corpus C_D while fine-tuning θ_S, or freezing them [Zhang et al., 2018].

Exploiting Monolingual Corpora

Another solution to the lack of vast document-level parallel corpora is leveraging on huge *monolingual* document-level corpora like BookCorpus [Zhu et al., 2015] and PG-19 [Rae et al., 2019]. In the literature, we can find various approaches to leverage monolingual corpora:

Exploiting Monolingual Corpora

Another solution to the lack of vast document-level parallel corpora is leveraging on huge *monolingual* document-level corpora like BookCorpus [Zhu et al., 2015] and PG-19 [Rae et al., 2019]. In the literature, we can find various approaches to leverage monolingual corpora:

- **Back-translate** target-side corpus to augment dl corpus [Sugiyama and Yoshinaga, 2019].

Exploiting Monolingual Corpora

Another solution to the lack of vast document-level parallel corpora is leveraging on huge *monolingual* document-level corpora like BookCorpus [Zhu et al., 2015] and PG-19 [Rae et al., 2019]. In the literature, we can find various approaches to leverage monolingual corpora:

- **Back-translate** target-side corpus to augment dl corpus [Sugiyama and Yoshinaga, 2019].
- Train **context-aware language models** on target/source-side corpus, then:

Exploiting Monolingual Corpora

Another solution to the lack of vast document-level parallel corpora is leveraging on huge *monolingual* document-level corpora like BookCorpus [Zhu et al., 2015] and PG-19 [Rae et al., 2019]. In the literature, we can find various approaches to leverage monolingual corpora:

- **Back-translate** target-side corpus to augment dl corpus [Sugiyama and Yoshinaga, 2019].
- Train **context-aware language models** on target/source-side corpus, then:
  - Generate translations by fusioning the decoder and the LM's scores to candidate words [Martnez Garcia et al., 2019].

Exploiting Monolingual Corpora

Another solution to the lack of vast document-level parallel corpora is leveraging on huge *monolingual* document-level corpora like BookCorpus [Zhu et al., 2015] and PG-19 [Rae et al., 2019]. In the literature, we can find various approaches to leverage monolingual corpora:

- **Back-translate** target-side corpus to augment dl corpus [Sugiyama and Yoshinaga, 2019].
- Train **context-aware language models** on target/source-side corpus, then:
  - Generate translations by fusioning the decoder and the LM's scores to candidate words [Martnez Garcia et al., 2019].
  - Initialize the econder (or decoder) of a DLNMT model [Li et al., 2019].

Exploiting Monolingual Corpora

Another solution to the lack of vast document-level parallel corpora is leveraging on huge *monolingual* document-level corpora like BookCorpus [Zhu et al., 2015] and PG-19 [Rae et al., 2019]. In the literature, we can find various approaches to leverage monolingual corpora:

- **Back-translate** target-side corpus to augment dl corpus [Sugiyama and Yoshinaga, 2019].
- Train **context-aware language models** on target/source-side corpus, then:
  - Generate translations by fusioning the decoder and the LM's scores to candidate words [Martnez Garcia et al., 2019].
  - Initialize the econder (or decoder) of a DLNMT model [Li et al., 2019].
- Train **Automatic Post Editing** systems on target-side corpus (See next slide).

Automatic Post Editing (APE)
[Voita et al., 2019a] devised an APE system called DocRepair, that turns a sentence-level translation into a context-aware translation. DocRepair can work on top of whatever sentence-level MT system.

Cache Approaches

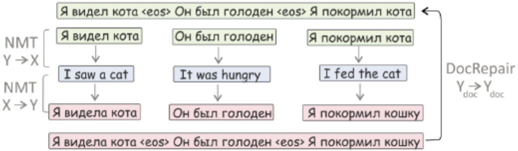


Figure 1: Training procedure of DocRepair. First, round-trip translations of individual sentences are produced to form an inconsistent text fragment (in the example, both genders of the speaker and the cat became inconsistent). Then, a repair model is trained to produce an original text from the inconsistent one.

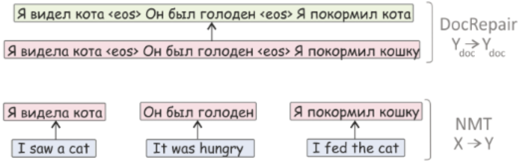


Figure 2: The process of producing document-level translations at test time is two-step: (1) sentences are translated independently using a sentence-level model, (2) DocRepair model corrects translation of the resulting text fragment.

Plan

Others

Approaches Including Additional Discourse Information as Input
These approaches consist in concatenation approaches or separate encoding approaches that also integrate discourse-related information as additional input features. Examples of extra features are:

Others

Approaches Including Additional Discourse Information as Input
These approaches consist in concatenation approaches or separate encoding approaches that also integrate discourse-related information as additional input features. Examples of extra features are:

- Lexical chains of semantically similar words to promote word sense disambiguation [Rios Gonzales et al., 2017].

Others

Approaches Including Additional Discourse Information as Input
These approaches consist in concatenation approaches or separate encoding approaches that also integrate discourse-related information as additional input features. Examples of extra features are:

- Lexical chains of semantically similar words to promote word sense disambiguation [Rios Gonzales et al., 2017].
- Coreference chains to promote coreference resolution [Stojanovski and Fraser, 2018, Ohtani et al., 2019].

Learning Approaches
[Jean and Cho, 2019] looked at the problem from a learning perspective and designed a regularisation term to encourage a DLNMT model to exploit the additional context in a useful way . This regularisation term is applied at the token, sentence and corpus levels and is based on pair-wise ranking loss, that is, it helps to assign a higher log-probability to a translation paired with the correct context than to the translation without context.

Plan

Possible Future Research Directions

Remarks and conclusions

Possible Future Research Directions

- Build a large DL corpus for training systems, or find automatic approaches to generate synthetic data other than back-translation.

Remarks and conclusions

Possible Future Research Directions

- Build a large DL corpus for training systems, or find automatic approaches to generate synthetic data other than back-translation.
  - E.g. imputing context sentences [Jean et al., 2019].

Remarks and conclusions

Possible Future Research Directions

- Build a large DL corpus for training systems, or find automatic approaches to generate synthetic data other than back-translation.
  - E.g. imputing context sentences [Jean et al., 2019].
- Design models with good results on lexical cohesion [Voita et al., 2019b].

Remarks and conclusions

Possible Future Research Directions

- Build a large DL corpus for training systems, or find automatic approaches to generate synthetic data other than back-translation.
  - E.g. imputing context sentences [Jean et al., 2019].
- Design models with good results on lexical cohesion [Voita et al., 2019b].
- Design models exploiting full context in a memory-efficient way:

Possible Future Research Directions

- Build a large DL corpus for training systems, or find automatic approaches to generate synthetic data other than back-translation.
    - E.g. imputing context sentences [Jean et al., 2019].
- Design models with good results on lexical cohesion [Voita et al., 2019b].
- Design models exploiting full context in a memory-efficient way:
    - Dynamic context integration.

Remarks and conclusions

Possible Future Research Directions

- Build a large DL corpus for training systems, or find automatic approaches to generate synthetic data other than back-translation.
  - E.g. imputing context sentences [Jean et al., 2019].
- Design models with good results on lexical cohesion [Voita et al., 2019b].
- Design models exploiting full context in a memory-efficient way:
  - Dynamic context integration.
  - Caches integrated to Transformer-based models.

Remarks and conclusions

Possible Future Research Directions

- Build a large DL corpus for training systems, or find automatic approaches to generate synthetic data other than back-translation.
  - E.g. imputing context sentences [Jean et al., 2019].
- Design models with good results on lexical cohesion [Voita et al., 2019b].
- Design models exploiting full context in a memory-efficient way:
  - Dynamic context integration.
  - Caches integrated to Transformer-based models.
- Design automatic post-processing models that are lightweight and can be trained on little data [Kim et al., 2019].

Possible Future Research Directions

- Build a large DL corpus for training systems, or find automatic approaches to generate synthetic data other than back-translation.
  - E.g. imputing context sentences [Jean et al., 2019].
- Design models with good results on lexical cohesion [Voita et al., 2019b].
- Design models exploiting full context in a memory-efficient way:
  - Dynamic context integration.
  - Caches integrated to Transformer-based models.
- Design automatic post-processing models that are lightweight and can be trained on little data [Kim et al., 2019].
- Study pre-trained language models for DLNMT decoder.

Remarks and conclusions

Possible Future Research Directions

- Build a large DL corpus for training systems, or find automatic approaches to generate synthetic data other than back-translation.
  - E.g. imputing context sentences [Jean et al., 2019].
- Design models with good results on lexical cohesion [Voita et al., 2019b].
- Design models exploiting full context in a memory-efficient way:
  - Dynamic context integration.
  - Caches integrated to Transformer-based models.
- Design automatic post-processing models that are lightweight and can be trained on little data [Kim et al., 2019].
- Study pre-trained language models for DLNMT decoder.
- Study other learning methods that foster document-level modeling [Jean and Cho, 2019].

Plan

Evaluation

Evaluation

- Classical metrics such as BLUE and METEOR are inadequate in evaluating document-level MT because they evaluate **average translation quality** at **sentence-level**. Thus:

Evaluation

- Classical metrics such as BLUE and METEOR are inadequate in evaluating document-level MT because they evaluate **average translation quality** at **sentence-level**. Thus:
  - they are unable to capture document-wide phenomena like coherence and cohesion [Wong and Kit, 2012].

Evaluation

- Classical metrics such as BLUE and METEOR are inadequate in evaluating document-level MT because they evaluate **average translation quality** at **sentence-level**. Thus:
  - they are unable to capture document-wide phenomena like coherence and cohesion [Wong and Kit, 2012].
  - they are not able to measure improvements over discourse phenomena that affect few words but heavily influence fluency and correctness of the translation [Mller et al., 2018]. E.g. pronomial anaphora.

Evaluation

- Classical metrics such as BLUE and METEOR are inadequate in evaluating document-level MT because they evaluate **average translation quality** at **sentence-level**. Thus:
  - they are unable to capture document-wide phenomena like coherence and cohesion [Wong and Kit, 2012].
  - they are not able to measure improvements over discourse phenomena that affect few words but heavily influence fluency and correctness of the translation [Mller et al., 2018]. E.g. pronomial anaphora.
- Evaluation of **translation of discourse phenomena** can be undertaken with:

Evaluation

- Classical metrics such as BLUE and METEOR are inadequate in evaluating document-level MT because they evaluate **average translation quality** at **sentence-level**. Thus:
  - they are unable to capture document-wide phenomena like coherence and cohesion [Wong and Kit, 2012].
  - they are not able to measure improvements over discourse phenomena that affect few words but heavily influence fluency and correctness of the translation [Mller et al., 2018]. E.g. pronomial anaphora.
- Evaluation of **translation of discourse phenomena** can be undertaken with:
  - automatic metrics.

Evaluation

- Classical metrics such as BLUE and METEOR are inadequate in evaluating document-level MT because they evaluate **average translation quality** at **sentence-level**. Thus:
  - they are unable to capture document-wide phenomena like coherence and cohesion [Wong and Kit, 2012].
  - they are not able to measure improvements over discourse phenomena that affect few words but heavily influence fluency and correctness of the translation [Mller et al., 2018]. E.g. pronomial anaphora.
- Evaluation of **translation of discourse phenomena** can be undertaken with:
  - automatic metrics.
  - test suites.

Evaluation

Evaluation Type	Discourse Phenomena	Dependency	Reference
Automatic Metric	Pronouns	Alignments, Pronoun lists	[29]
		Alignments, Pronoun lists	[77]
		English in target (anaphoric)	[43]
	Lexical Cohesion	Lexical cohesion devices	[120]
		Topic model, Lexical chain	[21]
	Discourse Connectives	Alignments, Dictionary	[26]
		Discourse parser	[25, 39]
		Discourse parser	[99]
Test Suites	Pronouns	En→Fr	[23]
		En→Fr (anaphora)	[7]
		En→De (anaphora)	[78]
	Cohesion	En→Fr	[7]
		En→Ru	[115]
	Coherence	En→Fr	[7]
		En↔De, Cs↔De, En→Cs	[117]
		En→Cs	[90]
	Conjunction	En/Fr→De	[85]
	Deixis, Ellipsis	En→Ru	[115]
	Grammatical Phenomena	En→De	[93]
		De→En	[2]
	Word Sense Disambiguation	De→En/Fr	[89, 88]
		En↔De/Fi/Lt/Ru, En→Cs	[86]

Figure: Overview of works on discourse phenomena evaluation in MT [Maruf et al., 2019b].

Evaluation

The evaluation of translation of discourse-phenomena in document-level MT should:

[1] in the remainder of this presentation, we refer to inter-sentential context simply as context.

Evaluation

The evaluation of translation of discourse-phenomena in document-level MT should:

- Provide inter-sentential context[1].

[1] in the remainder of this presentation, we refer to inter-sentential context simply as context.

Evaluation

The evaluation of translation of discourse-phenomena in document-level MT should:

- Provide inter-sentential context[1].
- Focus on context-dependent cases.

[1] in the remainder of this presentation, we refer to inter-sentential context simply as context.

Evaluation

The evaluation of translation of discourse-phenomena in document-level MT should:

- Provide inter-sentential context[1].
- Focus on context-dependent cases.
    - E.g., pronominal anaphora cases in which the antecedent is in a previous sentence (context-dependent), instead of being in the same sentence (context-independent).

[1] in the remainder of this presentation, we refer to inter-sentential context simply as context.

Evaluation

The evaluation of translation of discourse-phenomena in document-level MT should:

- Provide inter-sentential context[1].
- Focus on context-dependent cases.
  - E.g., pronominal anaphora cases in which the antecedent is in a previous sentence (context-dependent), instead of being in the same sentence (context-independent).
- Focus on hard cases.

[1] in the remainder of this presentation, we refer to inter-sentential context simply as context.

Evaluation

The evaluation of translation of discourse-phenomena in document-level MT should:

- Provide inter-sentential context[1].
- Focus on context-dependent cases.
  - E.g., pronominal anaphora cases in which the antecedent is in a previous sentence (context-dependent), instead of being in the same sentence (context-independent).
- Focus on hard cases.
  - E.g., when translating English to French, **he** is easy whereas **it** is hard to translate because ambiguous.

[1] in the remainder of this presentation, we refer to inter-sentential context simply as context.

Plan

Automatic metrics

Accuracy of Pronoun Translation [Miculicich Werlen and Popescu-Belis, 2017]:

Accuracy of Pronoun Translation [Miculicich Werlen and Popescu-Belis, 2017]:

1. Align source, reference and candidate translation with GIZA++ plus some heuristics.

Accuracy of Pronoun Translation [Miculicich Werlen and Popescu-Belis, 2017]:

1. Align source, reference and candidate translation with GIZA++ plus some heuristics.
2. Compare candidate and reference pronouns taking into account **equivalent** pronouns and identical pronouns with **different forms** (target language-specific).

Automatic metrics

Accuracy of Pronoun Translation [Miculicich Werlen and Popescu-Belis, 2017]:

1. Align source, reference and candidate translation with GIZA++ plus some heuristics.
2. Compare candidate and reference pronouns taking into account **equivalent** pronouns and identical pronouns with **different forms** (target language-specific).
   - E.g. *it is difficult* → *il/ce/c' est difficile.*

Accuracy of Pronoun Translation [Miculicich Werlen and Popescu-Belis, 2017]:

1. Align source, reference and candidate translation with GIZA++ plus some heuristics.
2. Compare candidate and reference pronouns taking into account **equivalent** pronouns and identical pronouns with **different forms** (target language-specific).
   - E.g. *it is difficult* → *il/ce/c' est difficile*.

- *Compatible languages*: conceived for English to French but it has also been extended to other language pairs.

Automatic metrics

Pronoun Pair-wise Ranking [Jwalapuram et al., 2019]

Automatic metrics

Pronoun Pair-wise Ranking [Jwalapuram et al., 2019]

- *Rationale*: **ranking-based evaluation** measures can achieve higher correlations with human judgments, as rankings are simpler to obtain from humans and to train models on.

Pronoun Pair-wise Ranking [Jwalapuram et al., 2019]

- *Rationale*: **ranking-based evaluation** measures can achieve higher correlations with human judgments, as rankings are simpler to obtain from humans and to train models on.
- *Compatible languages*: all target languages on which it is possible to built a training set (parsers needed). all source languages!

Pronoun Pair-wise Ranking [Jwalapuram et al., 2019]

- *Rationale*: **ranking-based evaluation** measures can achieve higher correlations with human judgments, as rankings are simpler to obtain from humans and to train models on.
- *Compatible languages*: all target languages on which it is possible to built a training set (parsers needed). all source languages!
- *System input*: a pair $R = (C_r, r)$ and $S = (C_s, s)$ of translations to be compared:

Pronoun Pair-wise Ranking [Jwalapuram et al., 2019]

- *Rationale*: **ranking-based evaluation** measures can achieve higher correlations with human judgments, as rankings are simpler to obtain from humans and to train models on.
- *Compatible languages*: all target languages on which it is possible to built a training set (parsers needed). all source languages!
- *System input*: a pair $R = (C_r, r)$ and $S = (C_s, s)$ of translations to be compared:
  - C_r, C_s are the two translations. Each C can comprise one or multiple sentences (context)

Automatic metrics

Pronoun Pair-wise Ranking [Jwalapuram et al., 2019]

- *Rationale*: **ranking-based evaluation** measures can achieve higher correlations with human judgments, as rankings are simpler to obtain from humans and to train models on.
- *Compatible languages*: all target languages on which it is possible to built a training set (parsers needed). all source languages!
- *System input*: a pair $R = (C_r, r)$ and $S = (C_s, s)$ of translations to be compared:
  - C_r, C_s are the two translations. Each C can comprise one or multiple sentences (context)
  - r, s are the positions of the pronouns to be compared in the translation R and S, respectively.

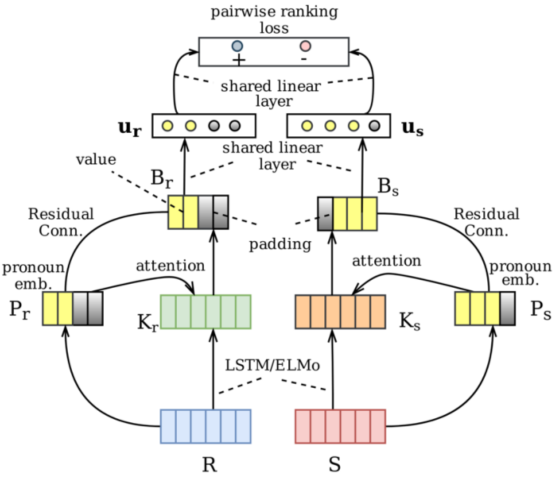


Figure: Pairwise ranking system by [Jwalapuram et al., 2019].

Automatic metrics

Lexical Cohesion Devices [Wong and Kit, 2012]

Automatic metrics

Lexical Cohesion Devices [Wong and Kit, 2012]

1. A **stemming algorithm** [Porter, 1980] is used to identify word stems for each content word .

Automatic metrics

Lexical Cohesion Devices [Wong and Kit, 2012]

1. A **stemming algorithm** [Porter, 1980] is used to identify word stems for each content word .
    - Words with the same stem are defined and counted as **Repetitions**.

Automatic metrics

Lexical Cohesion Devices [Wong and Kit, 2012]

1. A **stemming algorithm** [Porter, 1980] is used to identify word stems for each content word .
   - Words with the same stem are defined and counted as **Repetitions**.
2. **WordNet** [Fellbaum, 1998] is used to cluster synonims and superordinates into semantic groups.

Automatic metrics

Lexical Cohesion Devices [Wong and Kit, 2012]

1. A **stemming algorithm** [Porter, 1980] is used to identify word stems for each content word .
   - Words with the same stem are defined and counted as **Repetitions**.
2. **WordNet** [Fellbaum, 1998] is used to cluster synonims and superordinates into semantic groups.
   - Words belonging to the same semantic group or close semantic groups (near-synonims) are defined and counted as **Lexical Cohesion Devices** (LCD).

Automatic metrics

Lexical Cohesion Devices [Wong and Kit, 2012]

1. A **stemming algorithm** [Porter, 1980] is used to identify word stems for each content word .
   - Words with the same stem are defined and counted as **Repetitions**.
2. **WordNet** [Fellbaum, 1998] is used to cluster synonims and superordinates into semantic groups.
   - Words belonging to the same semantic group or close semantic groups (near-synonims) are defined and counted as **Lexical Cohesion Devices** (LCD).
3. A **hybrid metric** can then be defined as weighted average of:

Automatic metrics

Lexical Cohesion Devices [Wong and Kit, 2012]

1. A **stemming algorithm** [Porter, 1980] is used to identify word stems for each content word .
   - Words with the same stem are defined and counted as **Repetitions**.
2. **WordNet** [Fellbaum, 1998] is used to cluster synonims and superordinates into semantic groups.
   - Words belonging to the same semantic group or close semantic groups (near-synonims) are defined and counted as **Lexical Cohesion Devices** (LCD).
3. A **hybrid metric** can then be defined as weighted average of:
   - a **classic sentence-level metric**, e.g. BLEU, METEOR, TER.

Automatic metrics

Lexical Cohesion Devices [Wong and Kit, 2012]

1. A **stemming algorithm** [Porter, 1980] is used to identify word stems for each content word .
    - Words with the same stem are defined and counted as **Repetitions**.
2. **WordNet** [Fellbaum, 1998] is used to cluster synonims and superordinates into semantic groups.
    - Words belonging to the same semantic group or close semantic groups (near-synonims) are defined and counted as **Lexical Cohesion Devices** (LCD).
3. A **hybrid metric** can then be defined as weighted average of:
    - a **classic sentence-level metric**, e.g. BLEU, METEOR, TER.
    - **a lexical cohesion metric**, e.g. *Repetitions/content words* or *LCD/content words*.

Automatic metrics

Lexical Cohesion Devices [Wong and Kit, 2012]

1. A **stemming algorithm** [Porter, 1980] is used to identify word stems for each content word .
   - Words with the same stem are defined and counted as **Repetitions**.
2. **WordNet** [Fellbaum, 1998] is used to cluster synonims and superordinates into semantic groups.
   - Words belonging to the same semantic group or close semantic groups (near-synonims) are defined and counted as **Lexical Cohesion Devices** (LCD).
3. A **hybrid metric** can then be defined as weighted average of:
   - a **classic sentence-level metric**, e.g. BLEU, METEOR, TER.
   - **a lexical cohesion metric**, e.g. *Repetitions*/*content words* or *LCD*/*content words*.

- *Compatible languages*: all languages with stemmers and WordNets available.

Plan

Test Suites

In the literature we can distinguish three kinds of test suites:

Test Suites

In the literature we can distinguish three kinds of test suites:

- **Manual test suites** consist in the manual evaluation of a number of test cases (e.g. discourse phenomena) based on a given machine translation task.

Test Suites

In the literature we can distinguish three kinds of test suites:

- **Manual test suites** consist in the manual evaluation of a number of test cases (e.g. discourse phenomena) based on a given machine translation task.
  - For instance, WMT2019 not only provided ratings for each system output but also some manual evaluation and analysis of the outputs, like the English → Czech test suite one on coherence by [Rysov et al., 2019].

Test Suites

In the literature we can distinguish three kinds of test suites:

- **Manual test suites** consist in the manual evaluation of a number of test cases (e.g. discourse phenomena) based on a given machine translation task.
  - For instance, WMT2019 not only provided ratings for each system output but also some manual evaluation and analysis of the outputs, like the English $\rightarrow$ Czech test suite one on coherence by [Rysov et al., 2019].
- **Specialized test sets** are like normal MT test sets but consist of sentence pairs that are more densely populated with specific discourse phenomena. Translations are evaluated on such tests sets by means of average quality metrics like BLEU.

Test Suites

In the literature we can distinguish three kinds of test suites:

- **Manual test suites** consist in the manual evaluation of a number of test cases (e.g. discourse phenomena) based on a given machine translation task.
  - For instance, WMT2019 not only provided ratings for each system output but also some manual evaluation and analysis of the outputs, like the English → Czech test suite one on coherence by [Rysov et al., 2019].
- **Specialized test sets** are like normal MT test sets but consist of sentence pairs that are more densely populated with specific discourse phenomena. Translations are evaluated on such tests sets by means of average quality metrics like BLEU.
  - E.g. [Voita et al., 2018] build a specialized English → Russian test set by retrieving from OpenSubtitles2016 all the sentences containing pronouns that are coreferent to an expression in the previous sentence.

Test Suites

In the literature we can distinguish three kinds of test suites:

- **Manual test suites** consist in the manual evaluation of a number of test cases (e.g. discourse phenomena) based on a given machine translation task.
  - For instance, WMT2019 not only provided ratings for each system output but also some manual evaluation and analysis of the outputs, like the English → Czech test suite one on coherence by [Rysov et al., 2019].
- **Specialized test sets** are like normal MT test sets but consist of sentence pairs that are more densely populated with specific discourse phenomena. Translations are evaluated on such tests sets by means of average quality metrics like BLEU.
  - E.g. [Voita et al., 2018] build a specialized English → Russian test set by retrieving from OpenSubtitles2016 all the sentences containing pronouns that are coreferent to an expression in the previous sentence.
- **Contrastive test suites** consists in blocks of few candidate translations of a given source in which one translation is correct and the others are not. MT systems are assessed on their ability to rank correct translations higher than the incorrect ones.

Contrastive Test Suites

Pronomial Anaphora, Lexical Coherence and Cohesion [Bawden et al., 2018]

Contrastive Test Suites

Pronomial Anaphora, Lexical Coherence and Cohesion [Bawden et al., 2018]

- *Language* English → French (OpenSubtitles2016).

Contrastive Test Suites

Pronomial Anaphora, Lexical Coherence and Cohesion [Bawden et al., 2018]

- *Language* English $\longrightarrow$ French (OpenSubtitles2016).
- One test suite on **pronomial anaphora** comprised of 50 blocks.

Contrastive Test Suites

Source:

context:	Oh, I hate **flies**. Look, there's another one!
current sent.:	Don't worry, I'll kill **it** for you.

Target:

1	context:	Ô je déteste les **mouches**. Regarde, il y en a une autre !
	correct:	T'inquiète, je **la** tuerai pour toi.
	incorrect:	T'inquiète, je **le** tuerai pour toi.
2	context:	Ô je déteste les **moucherons**. Regarde, il y en a un autre !
	correct:	T'inquiète, je **le** tuerai pour toi.
	incorrect:	T'inquiète, je **la** tuerai pour toi.
3	context:	Ô je déteste les **araignées**. Regarde, il y en a une autre !
	semi-correct:	T'inquiète, je **la** tuerai pour toi.
	incorrect:	T'inquiète, je **le** tuerai pour toi.
4	context:	Ô je déteste les **papillons**. Regarde, il y en a un autre !
	semi-correct:	T'inquiète, je **le** tuerai pour toi.
	incorrect:	T'inquiète, je **la** tuerai pour toi.

Figure: Example block of the pronomial anaphora test suite.

Contrastive Test Suites

Pronomial Anaphora, Lexical Coherence and Cohesion [Bawden et al., 2018]

- *Language* English → French (OpenSubtitles2016).
- One test suite on **pronomial anaphora** comprised of 50 blocks.
- One on **lexical coherence and cohesion**, comprised of 100 blocks.

Contrastive Test Suites

Source:	
context:	So what do you say to £50?
current sent.:	It's a little **steeper** than I was expecting.
Target:	
context:	Qu'est-ce que vous en pensez de 50£ ?
correct:	C'est un peu plus **cher** que ce que je pensais.
incorrect:	C'est un peu plus **raide** que ce que je pensais.

Source:	
context:	How are your feet holding up?
current sent.:	It's a little **steeper** than I was expecting.
Target:	
context:	Comment vont tes pieds ?
correct:	C'est un peu plus **raide** que ce que je pensais.
incorrect:	C'est un peu plus **cher** que ce que je pensais.

Figure: Example block of the lexical coherence and cohesion test suite.

Contrastive Test Suites

Deixis, Ellipsis, and Lexical Cohesion [Voita et al., 2019b]

Contrastive Test Suites

Deixis, Ellipsis, and Lexical Cohesion [Voita et al., 2019b]

- *Language*: English → Russian (OpenSubtitles2018).

Contrastive Test Suites

Deixis, Ellipsis, and Lexical Cohesion [Voita et al., 2019b]

- *Language*: English → Russian (OpenSubtitles2018).
- *Design method*: manual design preceded by a human analysis on the most common translation errors in the target language pair.

Contrastive Test Suites

Large Contrastive Test-suite for Pronoun Translation [Mller et al., 2018]

Contrastive Test Suites

Large Contrastive Test-suite for Pronoun Translation [Mller et al., 2018]

- *Rationale*: previous contrastive test suites are not suitable for DLNMT systems because either they **don't provid context**, or they are **too small** to provide statistical significance [Bawden et al., 2018].

Contrastive Test Suites

Large Contrastive Test-suite for Pronoun Translation [MIler et al., 2018]

- *Rationale*: previous contrastive test suites are not suitable for DLNMT systems because either they **don't provid context**, or they are **too small** to provide statistical significance [Bawden et al., 2018].
- *Language*: English → German (OpenSubtitles2016).

Contrastive Test Suites

Large Contrastive Test-suite for Pronoun Translation [Mller et al., 2018]

- *Rationale*: previous contrastive test suites are not suitable for DLNMT systems because either they **don't provid context**, or they are **too small** to provide statistical significance [Bawden et al., 2018].
- *Language*: English → German (OpenSubtitles2016).
- *Focus*: *it* → *er, sie, es* (hard cases of inter-sentential anaphora).

Contrastive Test Suites

Large Contrastive Test-suite for Pronoun Translation [Mller et al., 2018]

- *Rationale*: previous contrastive test suites are not suitable for DLNMT systems because either they **don't provid context**, or they are **too small** to provide statistical significance [Bawden et al., 2018].
- *Language*: English → German (OpenSubtitles2016).
- *Focus*: *it* → *er, sie, es* (hard cases of inter-sentential anaphora).
- *Method*:

Contrastive Test Suites

Large Contrastive Test-suite for Pronoun Translation [Mller et al., 2018]

- *Rationale*: previous contrastive test suites are not suitable for DLNMT systems because either they **don't provid context**, or they are **too small** to provide statistical significance [Bawden et al., 2018].
- *Language*: English → German (OpenSubtitles2016).
- *Focus*: *it* → *er, sie, es* (hard cases of inter-sentential anaphora).
- *Method*:
  - **Align and parse** source-reference pairs with coreference annotators (e.g. CoreNLP for En)

Contrastive Test Suites

Large Contrastive Test-suite for Pronoun Translation [Mller et al., 2018]

- *Rationale*: previous contrastive test suites are not suitable for DLNMT systems because either they **don't provid context**, or they are **too small** to provide statistical significance [Bawden et al., 2018].
- *Language*: English → German (OpenSubtitles2016).
- *Focus*: *it* → *er, sie, es* (hard cases of inter-sentential anaphora).
- *Method*:
  - **Align and parse** source-reference pairs with coreference annotators (e.g. CoreNLP for En)
  - filter aligned sentences containing aligned pronouns and antecedents.

Contrastive Test Suites

Large Contrastive Test-suite for Pronoun Translation [MIler et al., 2018]

- *Rationale*: previous contrastive test suites are not suitable for DLNMT systems because either they **don't provid context**, or they are **too small** to provide statistical significance [Bawden et al., 2018].
- *Language*: English → German (OpenSubtitles2016).
- *Focus*: *it* → *er, sie, es* (hard cases of inter-sentential anaphora).
- *Method*:
  - **Align and parse** source-reference pairs with coreference annotators (e.g. CoreNLP for En)
  - filter aligned sentences containing aligned pronouns and antecedents.
  - **Randomly sample** 4000 instances of each of the three translations of *it* under consideration: *er,sie,es*.

Contrastive Test Suites

Large Contrastive Test-suite for Pronoun Translation [Mller et al., 2018]

- *Rationale*: previous contrastive test suites are not suitable for DLNMT systems because either they **don't provid context**, or they are **too small** to provide statistical significance [Bawden et al., 2018].
- *Language*: English → German (OpenSubtitles2016).
- *Focus*: *it* → *er, sie, es* (hard cases of inter-sentential anaphora).
- *Method*:
    - **Align and parse** source-reference pairs with coreference annotators (e.g. CoreNLP for En)
    - filter aligned sentences containing aligned pronouns and antecedents.
    - **Randomly sample** 4000 instances of each of the three translations of *it* under consideration: *er,sie,es*.
    - **Generate two contrastive translations for each** of the 12000 reference translations, by swapping the correct German pronoun with the two incorrect ones.

Plan

Remarks and conclusions

Automatic Metrics

Remarks and conclusions

Automatic Metrics

+ They are **unexpensive** w.r.t. human annotation.

Remarks and conclusions

Automatic Metrics

+ They are **unexpensive** w.r.t. human annotation.
+ they can be easily extended to all languages.

Remarks and conclusions

Automatic Metrics

+ They are **unexpensive** w.r.t. human annotation.
+ they can be easily extended to all languages.
- They are **noisy** because they often rely on other imperfect NLP systems. E.g. alignment and coreference systems.

Remarks and conclusions

Automatic Metrics

+ They are **unexpensive** w.r.t. human annotation.
+ they can be easily extended to all languages.
- They are **noisy** because they often rely on other imperfect NLP systems. E.g. alignment and coreference systems.
- They might **not be enough correlated with human judgment**:

Remarks and conclusions

Automatic Metrics

+ They are **unexpensive** w.r.t. human annotation.
+ they can be easily extended to all languages.
– They are **noisy** because they often rely on other imperfect NLP systems. E.g. alignment and coreference systems.
– They might **not be enough correlated with human judgment**:
    - is the case of APT, for example, which has been shown by [Guillou and Hardmeier, 2018] not to be suitable to evaluate the translation of pronouns with certain functions.

Remarks and conclusions

Automatic Metrics

+ They are **unexpensive** w.r.t. human annotation.
+ they can be easily extended to all languages.

– They are **noisy** because they often rely on other imperfect NLP systems. E.g. alignment and coreference systems.

– They might **not be enough correlated with human judgment**:
  - is the case of APT, for example, which has been shown by [Guillou and Hardmeier, 2018] not to be suitable to evaluate the translation of pronouns with certain functions.

– No existing metrics for coherence although it's very relevant for users.

Remarks and conclusions

Test Suites

[2] During scoring, the model is also provided with reference translations as target context (easier). Instead, during translation, the model needs to predict the full sequence, thus being subject to beam search failures and error propagation.

Remarks and conclusions

Test Suites

+ They can evaluate discourse phenomena translations with **high precision** and, if well designed, **hig recall**.

[2] During scoring, the model is also provided with reference translations as target context (easier). Instead, during translation, the model needs to predict the full sequence, thus being subject to beam search failures and error propagation.

Remarks and conclusions

Test Suites

+ They can evaluate discourse phenomena translations with **high precision** and, if well designed, **hig recall**.
– Excepts for specialized test sets (slide 49), test suites have a **limited scope**: fixed language pair, fixed number of context sentences (past and future).

[2] During scoring, the model is also provided with reference translations as target context (easier). Instead, during translation, the model needs to predict the full sequence, thus being subject to beam search failures and error propagation.

Remarks and conclusions

Test Suites

+ They can evaluate discourse phenomena translations with **high precision** and, if well designed, **hig recall**.
– Excepts for specialized test sets (slide 49), test suites have a **limited scope**: fixed language pair, fixed number of context sentences (past and future).
– Contrastive evaluation has **limited guarantees**: only permits to conclude whether or not the reference translation is more probable than a contrastive variant. It is not guaranteed at all that the MT system will output such reference translation.[2]

[2] During scoring, the model is also provided with reference translations as target context (easier). Instead, during translation, the model needs to predict the full sequence, thus being subject to beam search failures and error propagation.

Possible Future Research Directions

Remarks and conclusions

Possible Future Research Directions

- New automatic metrics strongly tested against human judgment.

Remarks and conclusions

Possible Future Research Directions

- New automatic metrics strongly tested against human judgment.
  - Works on coherence and cohesion are particularly lacking.

Remarks and conclusions

Possible Future Research Directions

- New automatic metrics strongly tested against human judgment.
  - Works on coherence and cohesion are particularly lacking.
- Semi-automatic metrics: use a high precision automatic metric and a human to evaluate negative cases.

Remarks and conclusions

Possible Future Research Directions

- New automatic metrics strongly tested against human judgment.
  - Works on coherence and cohesion are particularly lacking.
- Semi-automatic metrics: use a high precision automatic metric and a human to evaluate negative cases.
- New test suites for restricted scope.

Remarks and conclusions

Possible Future Research Directions

- New automatic metrics strongly tested against human judgment.
  - Works on coherence and cohesion are particularly lacking.
- Semi-automatic metrics: use a high precision automatic metric and a human to evaluate negative cases.
- New test suites for restricted scope.
  - Considering other documents other than movie subtitles for building test sets would be interesting for various reasons:
    - No multiple speakers, no unavailable context (the video), more phenomena related to future context.

Thank you for your attention!

References I

Agrawal, R. R., Turchi, M., and Negri, M. (2018).
Contextual Handling in Neural Machine Translation: Look Behind, Ahead and on Both Sides.
pages 11–20.
00007 Accepted: 2018-08-08T15:15:28Z.

Bawden, R., Sennrich, R., Birch, A., and Haddow, B. (2018).
Evaluating Discourse Phenomena in Neural Machine Translation.
In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1304–1313, New Orleans, Louisiana. Association for Computational Linguistics.
00056.

References II

Fellbaum, C. (1998).
A Semantic Network of English: The Mother of All WordNets.
Computers and the Humanities, 32(2):209–220.
00194.

Fu, H., Liu, C., and Sun, J. (2019).
Reference Network for Neural Machine Translation.
In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3002–3012, Florence, Italy. Association for Computational Linguistics.
00000.

References III

Guillou, L. and Hardmeier, C. (2018).
Automatic Reference-Based Evaluation of Pronoun Translation Misses the Point.
In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4797–4802, Brussels, Belgium. Association for Computational Linguistics.
00008.

Jean, S., Bapna, A., and Firat, O. (2019).
Fill in the Blanks: Imputing Missing Sentences for Larger-Context Neural Machine Translation.
arXiv:1910.14075 [cs].
00000 arXiv: 1910.14075.

References IV

Jean, S. and Cho, K. (2019).
Context-Aware Learning for Neural Machine Translation.
arXiv:1903.04715 [cs].
00003 arXiv: 1903.04715.

Jean, S., Lauly, S., Firat, O., and Cho, K. (2017).
Does Neural Machine Translation Benefit from Larger Context?
arXiv:1704.05135 [cs, stat].
00039 arXiv: 1704.05135.

Jwalapuram, P., Joty, S., Temnikova, I., and Nakov, P. (2019).
Evaluating Pronominal Anaphora in Machine Translation: An Evaluation Measure and a Test Suite.
arXiv:1909.00131 [cs].
00002 arXiv: 1909.00131.

References V

Kim, Y., Tran, D. T., and Ney, H. (2019).
When and Why is Document-level Context Useful in Neural Machine Translation?
arXiv:1910.00294 [cs].
00001 arXiv: 1910.00294.

Kuang, S., Xiong, D., Luo, W., and Zhou, G. (2018).
Modeling Coherence for Neural Machine Translation with Dynamic and Topic Caches.
In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 596–606, Santa Fe, New Mexico, USA. Association for Computational Linguistics.
00012.

References VI

- Li, L., Jiang, X., and Liu, Q. (2019).
  Pretrained Language Models for Document-Level Neural Machine Translation.
  arXiv:1911.03110 [cs].
  00001 arXiv: 1911.03110.
- Martnez Garcia, E., Creus, C., and Espaa-Bonet, C. (2019).
  Context-Aware Neural Machine Translation Decoding.
  In *Proceedings of the Fourth Workshop on Discourse in Machine Translation (DiscoMT 2019)*, pages 13–23, Hong Kong, China. Association for Computational Linguistics.
  00000.

References VII

Maruf, S. and Haffari, G. (2018).
Document Context Neural Machine Translation with Memory Networks.
In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1275–1284, Melbourne, Australia. Association for Computational Linguistics.
00000.

Maruf, S., Martins, A. F. T., and Haffari, G. (2019a).
Selective Attention for Context-aware Neural Machine Translation.
arXiv:1903.08788 [cs].
00012.

Maruf, S., Saleh, F., and Haffari, G. (2019b).
A Survey on Document-level Machine Translation: Methods and Evaluation.
arXiv:1912.08494 [cs].
00000 arXiv: 1912.08494.

Miculicich, L., Ram, D., Pappas, N., and Henderson, J. (2018).
Document-Level Neural Machine Translation with Hierarchical Attention Networks.
arXiv:1809.01576 [cs].
00029 arXiv: 1809.01576.

Miculicich Werlen, L. and Popescu-Belis, A. (2017).
Validation of an Automatic Metric for the Accuracy of Pronoun Translation (APT).
In *Proceedings of the Third Workshop on Discourse in Machine Translation*, pages 17–25, Copenhagen, Denmark. Association for Computational Linguistics.
00000.

Mller, M., Rios, A., Voita, E., and Sennrich, R. (2018).
A Large-Scale Test Set for the Evaluation of Context-Aware Pronoun Translation in Neural Machine Translation.
In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 61–72, Brussels, Belgium. Association for Computational Linguistics.
00010.

Ohtani, T., Kamigaito, H., Nagata, M., and Okumura, M. (2019).
Context-aware Neural Machine Translation with Coreference Information.
In *Proceedings of the Fourth Workshop on Discourse in Machine Translation (DiscoMT 2019)*, pages 45–50, Hong Kong, China. Association for Computational Linguistics.
00000.

References X

Porter, M. (1980).
An algorithm for suffix stripping.
Program, 40(3):211–218.
10830.

Rae, J. W., Potapenko, A., Jayakumar, S. M., and Lillicrap, T. P. (2019).
Compressive Transformers for Long-Range Sequence Modelling.
arXiv:1911.05507 [cs, stat].
00000 arXiv: 1911.05507.

Rios Gonzales, A., Mascarell, L., and Sennrich, R. (2017).
Improving Word Sense Disambiguation in Neural Machine Translation with Sense Embeddings.
In *Proceedings of the Second Conference on Machine Translation*, pages 11–19, Copenhagen, Denmark. Association for Computational Linguistics.
00030.

References XI

Rysov, K., Rysov, M., Musil, T., Polkov, L., and Bojar, O. (2019).
A Test Suite and Manual Evaluation of Document-Level NMT at WMT19.
In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 455–463, Florence, Italy. Association for Computational Linguistics.
00001.

Scherrer, Y., Tiedemann, J., and Loiciga, S. (2019).
Analysing concatenation approaches to document-level NMT in two different domains.
In *Proceedings of the Fourth Workshop on Discourse in Machine Translation (DiscoMT 2019)*, pages 51–61, Hong Kong, China. Association for Computational Linguistics.
00001.

References XII

Stojanovski, D. and Fraser, A. (2018).
Coreference and Coherence in Neural Machine Translation: A Study Using Oracle Experiments.
In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 49–60, Brussels, Belgium. Association for Computational Linguistics.
00003.

Sugiyama, A. and Yoshinaga, N. (2019).
Data augmentation using back-translation for context-aware neural machine translation.
In *Proceedings of the Fourth Workshop on Discourse in Machine Translation (DiscoMT 2019)*, pages 35–44, Hong Kong, China. Association for Computational Linguistics.
00000.

Tan, X., Zhang, L., Xiong, D., and Zhou, G. (2019).
Hierarchical Modeling of Global Context for Document-Level Neural Machine Translation.
In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1576–1585, Hong Kong, China. Association for Computational Linguistics.
00002.

Tiedemann, J. and Scherrer, Y. (2017).
Neural Machine Translation with Extended Context.
In *Proceedings of the Third Workshop on Discourse in Machine Translation*, pages 82–92, Copenhagen, Denmark. Association for Computational Linguistics.
00040.

Tu, Z., Liu, Y., Shi, S., and Zhang, T. (2017).
Learning to Remember Translation History with a Continuous Cache.
arXiv:1711.09367 [cs].
00041 arXiv: 1711.09367.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017).
Attention Is All You Need.
arXiv:1706.03762 [cs].
05728 arXiv: 1706.03762.

Voita, E., Sennrich, R., and Titov, I. (2019a).
Context-Aware Monolingual Repair for Neural Machine Translation.
arXiv:1909.01383 [cs].
00003 arXiv: 1909.01383.

Voita, E., Sennrich, R., and Titov, I. (2019b).
When a Good Translation is Wrong in Context: Context-Aware Machine Translation Improves on Deixis, Ellipsis, and Lexical Cohesion.
In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1198–1212, Florence, Italy. Association for Computational Linguistics.
00007.

Voita, E., Serdyukov, P., Sennrich, R., and Titov, I. (2018).
Context-Aware Neural Machine Translation Learns Anaphora Resolution.
In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1264–1274, Melbourne, Australia. Association for Computational Linguistics.
00049.

References XVI

Wang, L., Tu, Z., Way, A., and Liu, Q. (2017).
Exploiting Cross-Sentence Context for Neural Machine Translation.
In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2826–2831, Copenhagen, Denmark. Association for Computational Linguistics.
00048.

Wang, X., Weston, J., Auli, M., and Jernite, Y. (2019).
Improving Conditioning in Context-Aware Sequence to Sequence Models.
00002.

References XVII

Wong, B. T. M. and Kit, C. (2012).
Extending Machine Translation Evaluation Metrics with Lexical Cohesion to Document Level.
In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1060–1068, Jeju Island, Korea. Association for Computational Linguistics.
00044.

Zhang, J., Luan, H., Sun, M., Zhai, F., Xu, J., Zhang, M., and Liu, Y. (2018).
Improving the Transformer Translation Model with Document-Level Context.
In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 533–542, Brussels, Belgium. Association for Computational Linguistics.
00028.

Zheng, Z., Yue, X., Huang, S., Chen, J., and Birch, A. (2020).
Toward Making the Most of Context in Neural Machine Translation.
arXiv:2002.07982 [cs].
00000 arXiv: 2002.07982.

Zhu, Y., Kiros, R., Zemel, R., Salakhutdinov, R., Urtasun, R., Torralba, A., and Fidler, S. (2015).
Aligning Books and Movies: Towards Story-like Visual Explanations by Watching Movies and Reading Books.
arXiv:1506.06724 [cs].
00450 arXiv: 1506.06724.