-SOTA-
Document-level Neural Machine Translation

by Lorenzo Lupo

April 2020

Plan

1. Modern Neural Machine Translation
   - Overview
2. Evaluation
   - Automatic metrics
   - Test Suites
   - Remarks and conclusions

Plan

1. Modern Neural Machine Translation
   - Overview
2. Evaluation
   - Automatic metrics
   - Test Suites
   - Remarks and conclusions

Overview

- MT objective
- from SMT to NMT (attention?)
- sota models
  - transformer
  - transformer variations like Compressive Transformer, Reformer, etc.
- has MT reached human parity? [Lubli et al., 2018]). No, we need DLNMT.
- discourse phenomena, what are they?
- DLNMT objective

Overview

Note: context here is mostly used to indicate the sentences of a document that are not the one currently being translated (both source or target side)
MT output is usually evaluated by **average translation quality** metrics such as BLUE [Papineni et al., 2002] and METEOR [Banerjee and Lavie, 2005]. They are calculate at sentence level by on the base of the number of overlapping n-grams between the translation and the reference. The document-level score is simply an average of the sentence-level scores.

Plan

1. Modern Neural Machine Translation
    - Overview

2. Evaluation
    - Automatic metrics
    - Test Suites
    - Remarks and conclusions

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 **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 **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 **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., 2019].

Evaluation

The evaluation 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 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 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 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 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 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.

Automatic metrics

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

Automatic metrics

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

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

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).

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.*

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*.

- *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.

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!

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:

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)

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.

Automatic metrics

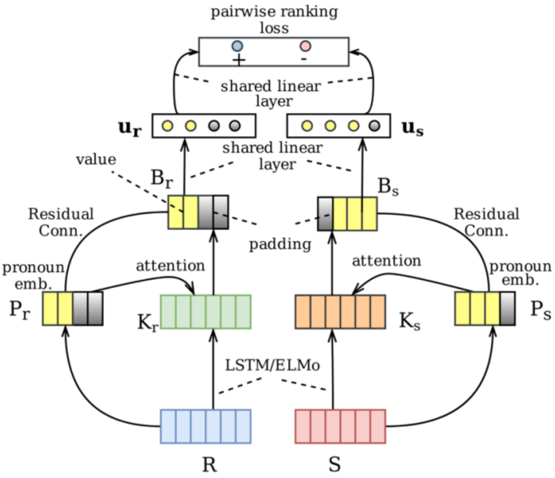

Figure: Pairwise ranking system by [Jwalapuram et al., 2019].

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.

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 $\rightarrow$ French (OpenSubtitles2016).

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.

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]

Contrastive Test Suites

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

- *Language* English $\longrightarrow$ 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

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.
- 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., 2019]

Contrastive Test Suites

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

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

Contrastive Test Suites

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

- *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 [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).

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 [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*.

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.

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 14), 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 14), 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.

Remarks and conclusions

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.

Thank you for your attention!

References I

Banerjee, S. and Lavie, A. (2005).
METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments.
In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.
00000.

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.
00055.

References II

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

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.

References III

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.

Lubli, S., Sennrich, R., and Volk, M. (2018).
Has Machine Translation Achieved Human Parity? A Case for Document-level Evaluation.
arXiv:1808.07048 [cs].
00035 arXiv: 1808.07048.

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

References IV

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.

References V

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002).
Bleu: a Method for Automatic Evaluation of Machine Translation.
In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.
10863.

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

References VI

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.

Voita, E., Sennrich, R., and Titov, I. (2019).
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.
00047.

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.

Markov Decision Processes

Reinforcement Learning

General class of algorithms that allow an agent to learn how to behave in a stochastic and possibly unknown environment by trial-and-error.

Markov Decision Process (MDP)

stochastic dynamical system specified by $< \mathbb{S}, \mathbb{A}, \mathcal{P}, \mathcal{R}, \gamma >$

1. $(\mathbb{S}, \mathcal{S})$ is a measurable state space
2. $(\mathbb{A}, \mathcal{A})$ is a measurable action space
3. $\mathcal{P} : \mathbb{S} \times \mathbb{A} \times \mathcal{S} \to \mathbb{R}$ is a Markov transition kernel
4. $\mathcal{R} : \mathbb{S} \times \mathbb{A} \to \mathbb{R}$ is a reward function
5. $0 < \gamma < 1$ is the discount factor.

Monte-Carlo Policy Gradient: Pseudocode

Input: Stochastic policy π_θ, Initial parameters θ_0, learning rate $\{\alpha_k\}$
Output: Approximation of the optimal policy $\pi_{\theta*} \approx \pi_*$

repeat
    Sample M trajectories $h^{(m)} = \{(s_t^{(m)}, a_t^{(m)}, r_{t+1}^{(m)})\}_{t=0}^{T^{(m)}}$ under policy π_{θ_k}
    Approximate policy gradient

$$\nabla_\theta J(\theta_k) \approx \frac{1}{M} \sum_{m=0}^{M} \sum_{u=0}^{T^{(m)}-1} \nabla_\theta \log \pi_{\theta_k} \left(s_u^{(m)}, a_u^{(m)} \right) \sum_{v \geqslant u}^{T^{(m)}-1} \gamma^{v-u} r_{v+1}^{(m)}$$

    Update parameters using gradient ascent $\theta_{k+1} = \theta_k + \alpha_k \nabla_\theta J(\theta_k)$
    $k \leftarrow k + 1$
until converged

Episodic PGPE Algorithm: Pseudocode

Input: Controller F_θ, hyper-distribution p_ξ, initial guess ξ_0, learning rate $\{\alpha_k\}$
Output: Approximation of the optimal policy $F_{\xi*} \approx \pi_*$

repeat
    for $m = 1, \ldots, M$ **do**
        Sample controller parameters $\theta^{(m)} \sim p_{\xi_k}$
        Sample trajectory $h^{(m)} = \{(s_t^{(m)}, a_t^{(m)}, r_{t+1}^{(m)})\}_{t=0}^{T^{(m)}}$ under policy $F_{\theta^{(m)}}$
    end for
    Approximate policy gradient

$$\nabla_\xi J(\xi_k) \approx \frac{1}{M} \sum_{m=1}^{M} \nabla_\xi \log p_\xi \left(\theta^{(m)}\right) \left[G\left(h^{(m)}\right) - b\right]$$

    Update hyperparameters using gradient ascent $\xi_{k+1} = \xi_k + \alpha_k \nabla_\xi J(\xi_k)$
    $k \leftarrow k + 1$
until converged

Truncated Multiple Importance Sampling Estimator

Importance Sampling

Given a bounded function $f : \mathcal{Z} \to \mathbb{R}$, and a set of i.i.d. outcomes $z_1, \ldots, z_N$ sampled from Q, the importance sampling estimator of $\mu := \mathop{\mathbb{E}}_{z \sim P}[f(z)]$ is:

$$\widehat{\mu}_{\text{IS}} = \frac{1}{N} \sum_{i=1}^{N} f(z_i) w_{P/Q}(z_i), \tag{1}$$

which is an unbiased estimator, i.e., $\mathop{\mathbb{E}}_{z_i \overset{\text{iid}}{\sim} Q} [\widehat{\mu}_{IS}] = \mu$.

Truncated Estimator With Balance Heuristic

$$\check{\mu}_{\text{BH}} = \frac{1}{N} \sum_{k=1}^{K} \sum_{i=1}^{N_k} \min \left\{ M, \frac{p(z_{ik})}{\sum_{j=1}^{K} \frac{N_j}{N} q_j(z_{ik})} \right\} f(z_{ik}). \tag{2}$$

OPTIMIST2

Theorem

regretdiscretized Let $\mathcal{X}$ be a d-dimensional compact arm set with $\mathcal{X} \subseteq [-D, D]^d$. For any $\kappa \geqslant 2$, under Assumptions 1 and 2, OPTIMIST2 with confidence schedule $\delta_t = \dfrac{6\delta}{\pi^2 t^2 \left(1 + \lceil t^{1/\kappa} \rceil^d\right)}$ and discretization schedule $\tau_t = \lceil t^{\frac{1}{\kappa}} \rceil$ guarantees, with probability at least $1 - \delta$:

$$\begin{aligned} Regret(T) \leqslant \Delta_0 + C_1 T^{\left(1 - \frac{1}{\kappa}\right) d} + C_2 T^{\frac{1}{1+\epsilon}} \\ \cdot \left[v_\epsilon \left((2 + {}^{d}/_{\kappa}) \log T + d \log 2 + \log \frac{\pi^2}{3\delta} \right) \right]^{\frac{\epsilon}{1+\epsilon}}, \end{aligned}$$

where $C_1 = \dfrac{\kappa}{\kappa - 1} LD$, $C_2 = (1 + \epsilon) \left(2\sqrt{2} + \dfrac{5}{3} \right) \|f\|_\infty$, and Δ_0 is the instantaneous regret of the initial arm $\boldsymbol{x}_0$.