

-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

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

- ▶ 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 [Miller et al., 2018]. E.g. pronominal anaphora.
- ▶ Evaluation of **discourse phenomena** can be undertaken with:
 - ▶ automatic metrics
 - ▶ contrastive 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]
Test Suites	Pronouns	Discourse parser	[25, 39]
		Discourse parser	[99]
		En→Fr	[23]
		En→Fr (anaphora)	[7]
	Cohesion	En→De (anaphora)	[78]
		En→Fr	[7]
	Coherence	En→Ru	[115]
		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].

- ▶ The evaluation of discourse phenomena in document-level MT, *desiderata*, and particularly the test suites, should:
 - ▶ Provide inter-sentential context¹;
 - ▶ 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.

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

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

- *Compatible languages*: conceived for English to French but it has also been extended to other language pairs.
- *Functioning*:
 - Align source, reference and candidate translation with GIZA++ plus some heuristics;
 - 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*.

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

- *Rationale1*: **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 languages. The metric **only needs target-side inputs** \implies thus it can be trained and evaluated without the need of a parallel corpus for each source-target pair.
- *System input*: a pair $R = (C_r, r)$ and $S = (C_s, s)$ of translations to be compared, where:
 - 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

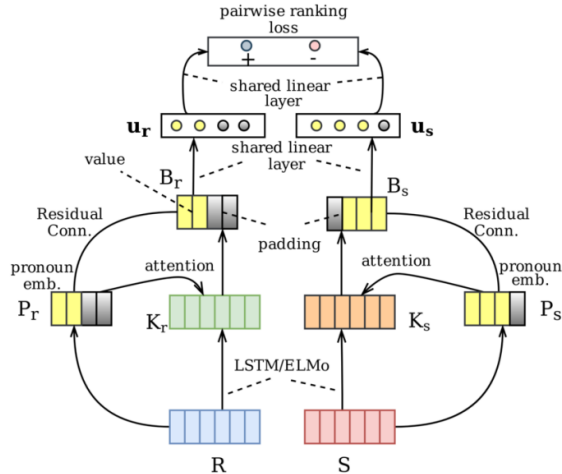


Figure: Pairwise ranking system by [Jwalapuram et al., 2019].

Lexical Cohesion extension [Wong and Kit, 2012]

- 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**.
- **WordNet** [Fellbaum, 1998] is used to cluster synonyms and superordinates into semantic groups;
 - Words belonging to the same semantic group or close semantic groups (near-synonyms) are defined and counted as **Lexical Cohesion Devices** (LCD).
- 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*.

Test suites


- ▶ [Bawden et al., 2018]: exemplary contrastive test suite, also good model reaching SOTA. Coherence very bad. Need for good models in coherence?
- ▶ [Miller et al., 2018]. Proposal: A Large-Scale Test Set for the Evaluation of Context-Aware Pronoun Translation in Neural Machine Translation.
 - ▶ Rationale: problems with previous contrastive test suites is that they are either too small to provide statistical significance [Bawden et al., 2018] or not adapted to properly test DLNMT systems because lemmatized or not always with context.
 - ▶ similar method will be adopted by [Jwalapuram et al., 2019]
 - ▶ Focus: inter-sentential anaphora, hard case, , i.e., it er, sie, es.

Remarks and conclusions


- automatic metrics
 - are less expensive than human annotation and thus more easily applicable to all languages
 - are noisy because they often rely on other imperfect NLP systems. E.g. alignment and coreference systems.
 - some automatic metrics might not be enough correlated with human judgment and miss the evaluation of some pronominal functions:
 - is the case for APT, for example [Guillou and Hardmeier, 2018]
 - there is nothing on coherence although it's the most relevant for post-editors together with cohesion
- test suites
 - systems trained on in-domain data perform better?
- what could we do?
 - strongly test new automatic metrics against human judgment
 - semi-automatic metrics: use a high precision automatic metric and a human to evaluate negative cases
 - keep designing test suites for very restricted scope

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

-  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.
-  Wu, Z. and Palmer, M. (1994).
Verbs semantics and lexical selection.
In Proceedings of the 32nd annual meeting on Association for Computational Linguistics, ACL '94, pages 133–138, Las Cruces, New Mexico. Association for Computational Linguistics. 03892.

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 $\langle \mathbb{S}, \mathbb{A}, \mathcal{P}, \mathcal{R}, \gamma \rangle$

1. $(\mathbb{S}, \mathcal{S})$ is a measurable state space
2. $(\mathbb{A}, \mathcal{A})$ is a measurable action space
3. $\mathcal{P} : \mathbb{S} \times \mathbb{A} \times \mathcal{S} \rightarrow \mathbb{R}$ is a Markov transition kernel
4. $\mathcal{R} : \mathbb{S} \times \mathbb{A} \rightarrow \mathbb{R}$ 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 $\pi_{\theta^*} \approx \pi_*$

1: **repeat**

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

3: Approximate policy gradient

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

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

5: $k \leftarrow k + 1$

6: **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 $F_{\xi^*} \approx \pi_*$

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

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

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

Truncated Multiple Importance Sampling Estimator

Importance Sampling

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

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

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

Truncated Estimator With Balance Heuristic

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

Theorem

regretdiscretized Let \mathcal{X} be a d -dimensional compact arm set with $\mathcal{X} \subseteq [-D, D]^d$. For any $\kappa \geq 2$, under Assumptions 1 and 2, OPTIMIST2 with confidence schedule

$$\delta_t = \frac{6\delta}{\pi^2 t^2 \left(1 + \lceil t^{1/\kappa} \rceil^d\right)} \text{ and discretization schedule } \tau_t = \lceil t^{\frac{1}{\kappa}} \rceil \text{ guarantees, with}$$

probability at least $1 - \delta$:

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

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