Context-Aware Machine Translation

Rina Kawamura

10/07/2021

Presentation Structure

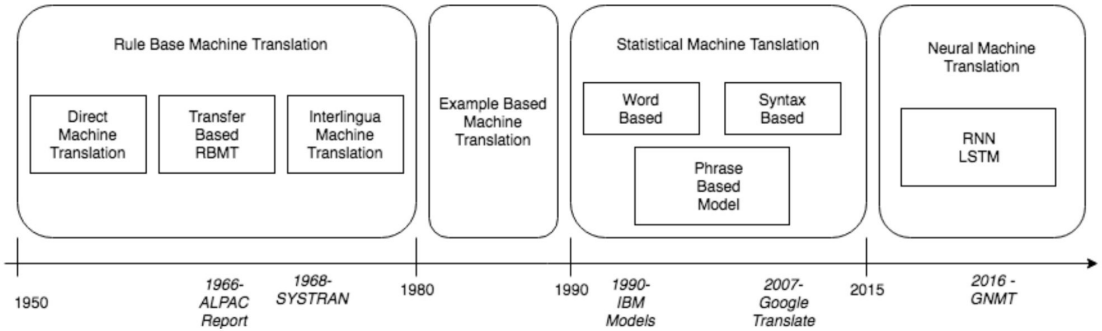
- Overview of Machine Translation
- Problems with sentence-level MT
- What is context-aware MT?
- Paper 1: An application of context-aware MT
- Paper 2: Evaluation and training of context-aware MT models
- Summary + Code repositories

Machine Translation

Objective: Translate one natural language into another automatically

History: Research started roughly around 1950

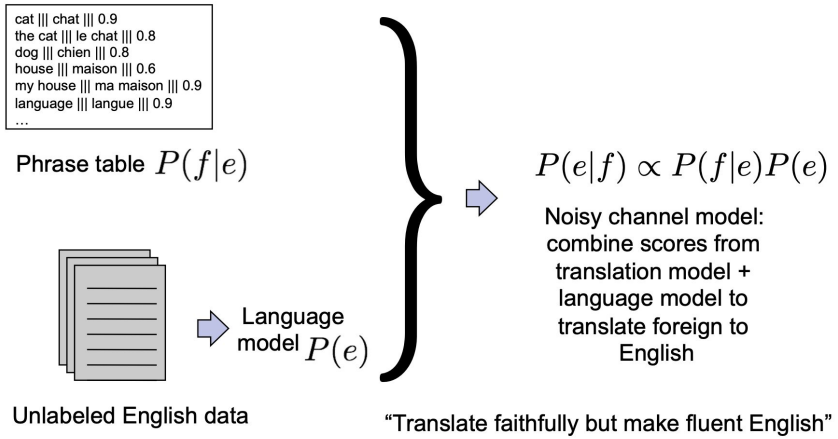
- Rule-based MT: Build dictionaries and transformation rules
- Example-based MT: Use a parallel corpora and perform translation by analogy
- Statistical MT: Based on probabilistic models constructed from parallel corpora
- Neural MT: Based on deep neural networks; Simplifies SMTs into single sequence model



https://cocoxu.github.io/courses/5525_slides_spring20/lec10-mt1.pdf

Statistical Machine Translation

- By far the most widely studied MT method before neural MT
- Translations generated based on statistical models formed from parallel corpora
- Phrase-based (as opposed to word based)
- Find best translation that maximizes P(e|f) = P(f|e)P(e)



https://cocoxu.github.io/courses/5525_slides_spring20/lec10-mt1.pdf

Statistical MT: Challenges & Shortcomings

Challenges

- Sentence alignment of parallel corpora data
- Word alignment
- Training the language model
- Finding heuristic to minimize search space to get best translation

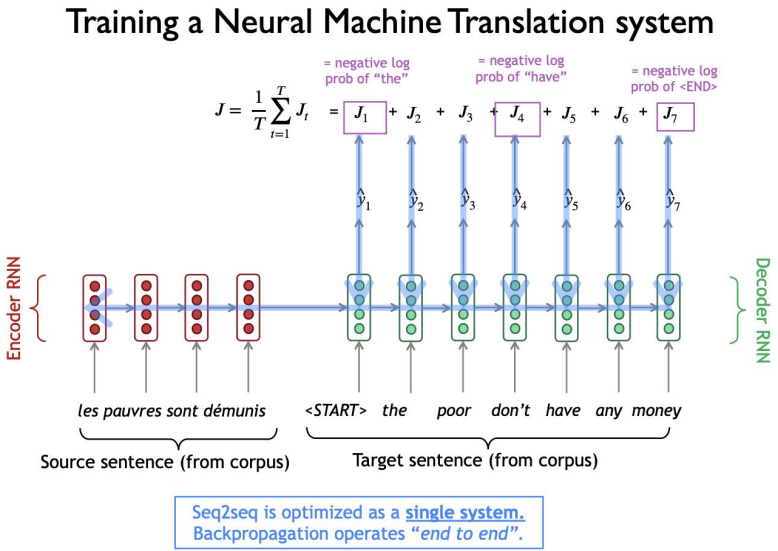
Shortcomings

- Not a lot of parallel corpus data available in some languages
- Doesn’t work as well for languages with different word order
- Specific errors hard to predict and fix
- Complex: Many components to model

Neural Machine Translation

- Use a single neural network to model machine translation
  - Encoder for natural language understanding of source language
  - Decoder for natural language generation in target language

- Sequence-to-sequence (Seq2Seq) model
  - Use two **RNNs** as encoder and decoder
  - Use **LSTMs** to learn on data with long range temporal dependencies
  - **Attention**: Focus on particular part of source at each time step of decoding
  - Auto-regressive encoding and decoding



https://courses.cs.washington.edu/courses/csep517/20wi/slides/csep517wi20-NeuralMT.pdf

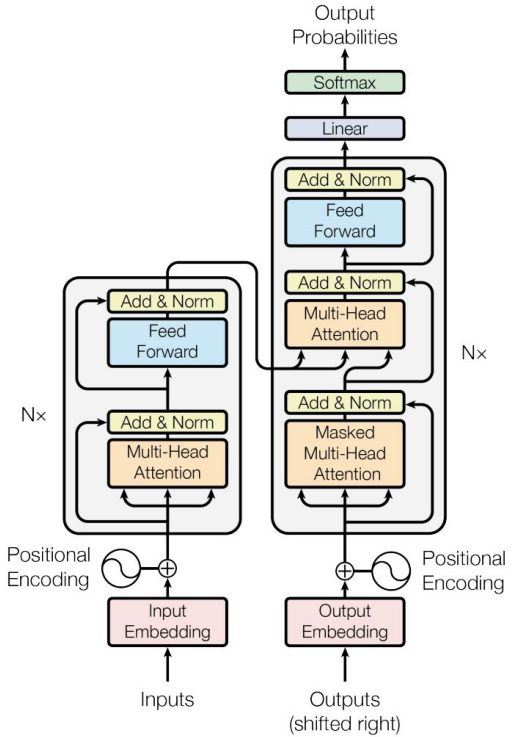
NMT Transformer Architecture

Replaced RNNs and became dominant architecture for NMT

- Serves as basis for many context-aware NMT models
- No recurrence structure so trains faster

Mechanism

- Encoder-Decoder Architecture
- Transformer consists of stacked attention layers
  - Multi-head self attention
  - Feed forward network
- Self-attention: Each word computes attention over every other word
- Multiple heads allow model to attend to different positions
- Decoder uses attention matrix to focus on the relevant positions in source sentence
- Non-autoregressive transformer



https://arxiv.org/pdf/1706.03762.pdf

Evaluation of MT

Human Evaluation

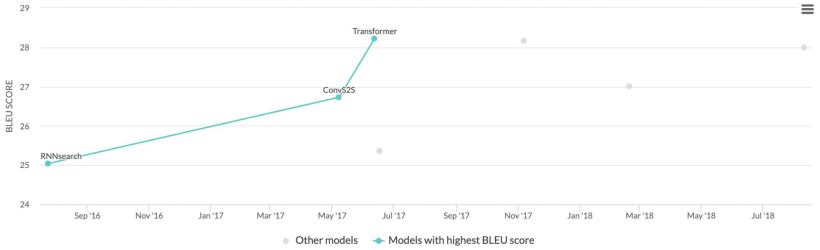
- Adequacy
- Grammaticality
- Fluency

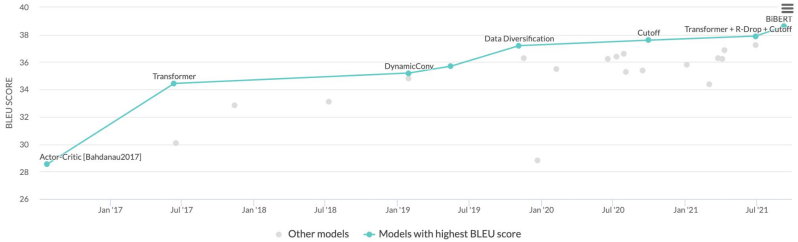
Automatic Methods

- BLEU Score
  - Score to compare candidate translation to one or more human reference translations
  - Calculate weighted geometric mean of multiple n-grams scores
  - Brevity penalty: Compare sentence lengths
  - Correlates well with human judgement
- Datasets: WMT (News commentaries, parliament proceedings, etc.), IWSLT (TED Talks)

Benchmarks

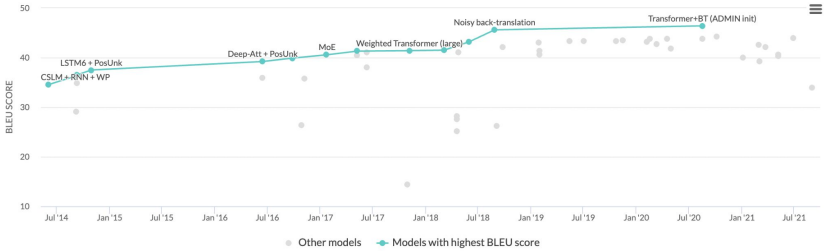
- IWSLT2015 English-German Dataset
  - Transformer: BLEU Score 28.23





- IWSLT2014 German-English Dataset
  - BiBERT: BLEU Score 38.61

- WMT2014 English-French Dataset
  - Transformer + BT: BLEU Score 46.4



Sentence-Level NMT

Fundamental assumption: Conditional independence between sentences

$$p(y^{(k)}|x^{(k)}) = \prod_{t=1}^{n} p(y_t^{(k)}|y_{<t}^{(k)}, x^{(k)}),$$

where $x^{(k)}$ and $y^{(k)}$ are the k-th source and target training sentences, and $y_t^{(k)}$ is the t-th token in $y^{(k)}$

Reported to have achieved human parity in some domains and language pairs

- Sentences evaluated in isolation

[Submitted on 15 Mar 2018 (v1), last revised 29 Jun 2018 (this version, v2)]

Achieving Human Parity on Automatic Chinese to English News Translation

Hany Hassan, Anthony Aue, Chang Chen, Vishal Chowdhary, Jonathan Clark, Christian Federmann, Xuedong Huang, Marcin Junczys-Dowmunt, William Lewis, Mu Li, Shujie Liu, Tie-Yan Liu, Renqian Luo, Arul Menezes, Tao Qin, Frank Seide, Xu Tan, Fei Tian, Lijun Wu, Shuangzhi Wu, Yingce Xia, Dongdong Zhang, Zhirui Zhang, Ming Zhou

Machine translation has made rapid advances in recent years. Millions of people are using it today in online translation systems and mobile applications in order to communicate across language barriers. The question naturally arises whether such systems can approach or achieve parity with human translations. In this paper, we first address the problem of how to define and accurately measure human parity in translation. We then describe Microsoft's machine translation system and measure the quality of its translations on the widely used WMT 2017 news translation task from Chinese to English. We find that our latest neural machine translation system has reached a new state-of-the-art, and that the translation quality is at human parity when compared to professional human translations. We also find that it significantly exceeds the quality of crowd-sourced non-professional translations.

Issues with sentence-level MT

Claims of human parity achievements do not hold for document-level evaluation.

Issues that arise:

- Anaphoric pronoun resolution
  - Antecedent word can be in previous sentences
- Lexical cohesion
  - Preservation of consistency in word choices throughout document
  - Eg. Model may translate the same Japanese word “時計” (tokei) into “clock” and “watch”
- Tense inconsistency
- Ellipsis
  - Problem if target sentences does not allow these constructions
- Ambiguity

EN However, the European Central Bank (ECB) took an interest in it. *It* describes bitcoin as “the most successful virtual currency”.

DE Dennoch hat die Europäische Zentralbank (EZB) Interesse hierfür gezeigt. *Sie* beschreibt Bitcoin als “die virtuelle Währung mit dem grössten Erfolg”.

https://aclanthology.org/W18-6307.pdf

My grandfather’s legs have failed because of the fluid. He had another visit today. Then his nature worsened. They can not speak for a few moments.

https://arxiv.org/pdf/1912.08494.pdf

Source	田中さん、よい時計をお持ちですね。 ありがとう、この時計は祖父の形見なんです。
Incohesive translations	You have a good clock, Mr. Tanaka. Thank you, this watch is a memento of my grandfather.
Cohesive translations	You have a good watch, Mr. Tanaka. Thank you, this watch is a memento of my grandfather.

Table 1: Example of lexically incohesive and cohesive translation

https://arxiv.org/pdf/2010.05193.pdf

Need for context-aware MT to take discourse into account

Context-aware Machine Translation

- Improve translation quality by making use of additional context
- Relaxes the independence assumption of sentence-level MT

$$p(y^{(k)} \mid x^{(k)}) = \prod_{t=1}^{n} p(y_t^{(k)} \mid y_{<t}^{(k)}, X, Y^{(<k)})$$

where $X := \{x^{(1)}, \ldots, x^{(K)}\}$ are the document's source sentences and $Y(<k) := \{y^{(1)}, \ldots, y^{(k-1)}\}$ the previously generated target sentences.

Approaches

- Using source-side context
- Using both source-side and target-side context
- Using monolingual data

Differing findings on whether source or target context contributes to translation more

Context-aware Architectures: Source-side

Local context

- Differing architectures for where to combine the source sentence and context sentences/representations
- Concatenate previous sentence(s) to current source sentence
  - Simple approach competitive with more complex counterparts
- Use additional encoder to extract contextual information from previous sentences or from pre-trained embeddings using sentence-based Transformer encoder
  - Context vector representation used as auxiliary input to decoder

Global context

- Summarize global context from all previous sentences in document
- Combine all context sentence embeddings with source sentence
- Selective attention: Use sparse attention to select and attend to only the relevant sentences from global context

Context-aware Architectures: Source-side + Target-side

- Include both previous source and previous target sentences
- Methods
  - Extended translation units
  - Memory networks (https://aclanthology.org/P18-1118.pdf)
    - Interdependency between source sentence, target sentence, and all other source sentences in document => Encodings of source sentences in document using RNN
    - Interdependency between target sentence and all other target sentences in document => Last decoder states of current translations
    - Decoder attends to relevant parts of external memories to generate output
  - Cache (https://arxiv.org/pdf/1711.09367.pdf)
    - Remember translation history
    - Store key-value pairs where key = attention context vector, val = corresponding decoder state from previous translations
    - Decoder state cached once full target sentence generated
  - Hierarchical Attention Network
    - Can use selective attention for both source and target-side contexts

Context-aware Architectures: Monolingual Data

DocRepair: Approach to context-aware MT using only monolingual data (https://aclanthology.org/D19-1081.pdf)

- Separate monolingual seq2seq model used to correct sentence-level translations
- Training
  - Original monolingual data => Consistent group
  - Sample round-trip translations using sentence-level NMT models => Inconsistent group
  - Train model to correct inconsistent translations into original sentences
- Usage
  - Use sentence-level NMT model to translate source sentence
  - Run it through reparation model to fix inconsistencies

Combining sentence-level NMT model with another model abstracts away the logic for incorporating context => Can theoretically be used for any NMT model

Evaluation of Context-aware Models

- BLEU score cannot measure context usage
- Contrastive datasets
  - Focus on some specific discourse phenomena (eg. pronoun resolution)
  - Consists of pairs of source and target sentences with contrastive translation variants

source:	*It could get tangled in your hair.*
reference:	***Sie*** *könnte sich in deinem Haar verfangen.*
contrastive:	***Er*** *könnte sich in deinem Haar verfangen.*
contrastive:	***Es*** *könnte sich in deinem Haar verfangen.*
antecedent en:	a bat
antecedent de:	eine Fledermaus (f.)
antecedent distance :	1

https://aclanthology.org/W18-6307.pdf

Model Comparisons

- No state-of-art model or benchmark yet
- No widely used dataset across document-level NMT

Model	TED	News	Europarl
*Transformer-Doc (Zhang et al., 2018)	24.00	23.08	29.32
*HAN-Doc (Miculicich et al., 2018)	24.58	25.03	28.60
SAN-Doc (Maruf et al., 2019)	24.42	24.84	29.75
Capsule-Doc (Yang et al., 2019)	25.19	22.37	29.82
Global-Context (Zheng et al., 2020)	25.10	24.91	30.40
Flat-Transformer +BERT (Ma et al., 2020)	26.61	24.52	31.99

BLEU Scores for 3 different datasets for various context-aware NMT models

BLEU scores and contrastive dataset evaluation for discourse phenomena

Model	BLEU	Deixis(%)	Lexical cohesion(%)	Ellipsis inflection(%)	Ellipsis VP(%)
Transformer	32.40	50.0	45.9	53.0	28.4
CADec (Voita et al., 2019a)	32.38	81.6	58.1	72.2	80.0
DocRepair (Voita et al., 2019b)	34.60	91.8	80.6	86.4	75.2
Layer-Wise Weighting (Xu et al., 2020)	32.75	85.4	64.3	77.1	81.3

https://github.com/ZNLP/SOTA-MT

Surveys of Context-aware Machine Translation

https://arxiv.org/pdf/2010.09482.pdf

[Submitted on 19 Oct 2020]

Diving Deep into Context-Aware Neural Machine Translation

Jingjing Huo, Christian Herold, Yingbo Gao, Leonard Dahlmann, Shahram Khadivi, Hermann Ney

Context-aware neural machine translation (NMT) is a promising direction to improve the translation quality by making use of the additional context, e.g., document-level translation, or having meta-information. Although there exist various architectures and analyses, the effectiveness of different context-aware NMT models is not well explored yet. This paper analyzes the performance of document-level NMT models on four diverse domains with a varied amount of parallel document-level bilingual data. We conduct a comprehensive set of experiments to investigate the impact of document-level NMT. We find that there is no single best approach to document-level NMT, but rather that different architectures come out on top on different tasks. Looking at task-specific problems, such as pronoun resolution or headline translation, we find improvements in the context-aware systems, even in cases where the corpus-level metrics like BLEU show no significant improvement. We also show that document-level back-translation significantly helps to compensate for the lack of document-level bi-texts.

https://arxiv.org/abs/1912.08494

[Submitted on 18 Dec 2019 (v1), last revised 13 Jan 2021 (this version, v3)]

A Survey on Document-level Neural Machine Translation: Methods and Evaluation

Sameen Maruf, Fahimeh Saleh, Gholamreza Haffari

Machine translation (MT) is an important task in natural language processing (NLP) as it automates the translation process and reduces the reliance on human translators. With the resurgence of neural networks, the translation quality surpasses that of the translations obtained using statistical techniques for most language-pairs. Up until a few years ago, almost all of the neural translation models translated sentences independently, without incorporating the wider document-context and inter-dependencies among the sentences. The aim of this survey paper is to highlight the major works that have been undertaken in the space of document-level machine translation after the neural revolution, so that researchers can recognise the current state and future directions of this field. We provide an organisation of the literature based on novelties in modelling and architectures as well as training and decoding strategies. In addition, we cover evaluation strategies that have been introduced to account for the improvements in document MT, including automatic metrics and discourse-targeted test sets. We conclude by presenting possible avenues for future exploration in this research field.

Discussion Question:

How do you think humans use context when translating a document into another language? Do you think the context-aware models we have seen today represent this process?

Paper 1: Context-Aware Neural Machine Translation for Korean Honorific Expressions (2021)

Published in *Electronics* **2021**

Yongkeun Hwang [1], Yanghoon Kim [1] and Kyomin Jung [1,2,*]

Motive for choosing paper: An application of context aware MT

- Problem: Handling the translation of honorifics
  - Honorifics frequently used in some languages: Korean, Japanese, Hindi, etc.
  - Preserve meaning and culture
  - Challenging: Different languages have different honorific systems

- Proposed Solution
  - Use both source and target side contexts
  - Encoder captures source language contextual information => Determines target honorifics
  - Context-aware post-editing system captures target sentence context to refine translation

How are honorifics captured in context?

We can use surrounding source sentences to deduce the appropriate honorific of the target sentence.

Example: The English word “wait” can be translated into honorific style (기다려요, gi-da-lyeo-yo) and non-honorific style (기다려, gi-da-lyeo)

Sentence	English	Korean
context_1	Come on, **dad**. Don’t you even take something?	**아빠**, 뭐라도 안 드세요?
context_0	Okay, give me some coffee.	좋아, 그럼 커피 좀 줘.
source/target	Wait a minute, please.	잠시만 기다려요.

Blue words are translated into honorific style and red words are translated into non-honorific style.

More examples

Sentence	English	Korean
context_1	You're back.	자네들 또 왔구만.
context_0	Yes, **sir**, we are.	예, **어르신**.
source/target	We're addicted to **your** citrus.	**어르신**의 감귤류에 중독 됐어요.

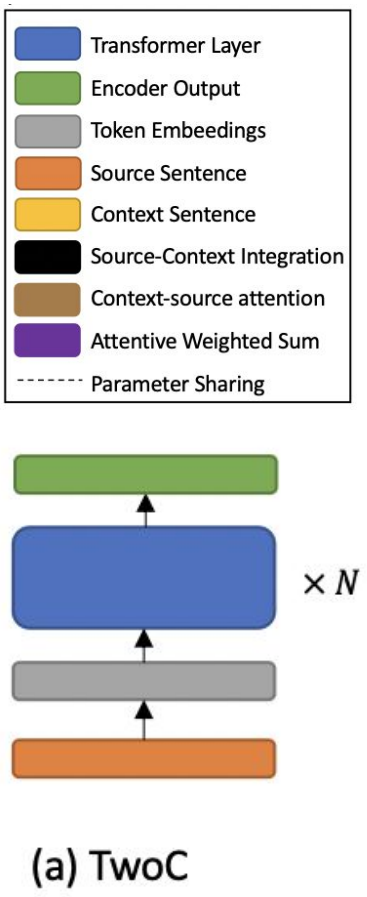
(a)

Sentence	English	Korean
context_1	You need to relax, okay?	진정해 주실래요?
context_0	You are not a **suspect**.	당신은 **용의자**가 아닙니다.
source/target	We should find Jessica right now.	저희는 빨리 제시카를 찾아야만 합니다.

Model Frameworks

5 transformer-based NMT models used in experiment

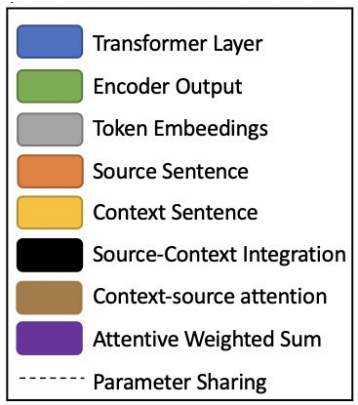
- **Transformer without contexts (TwoC)**
  - Baseline model
  - Transformer architecture that takes source sentence only as input

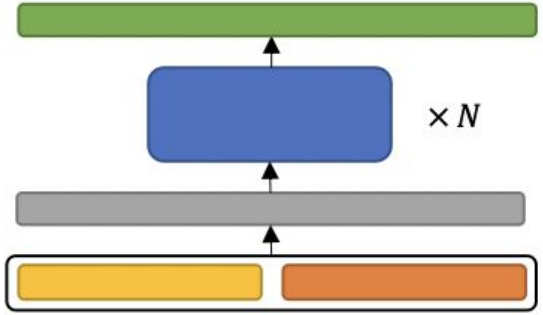


Model Frameworks

5 transformer-based NMT models used in experiment

- Transformer without contexts (TwoC)
- **Transformer with contexts (TwC)**
  - Concatenate all contextual sentences with source sentence
  - Use as single input to transformer model (same as TwoC)

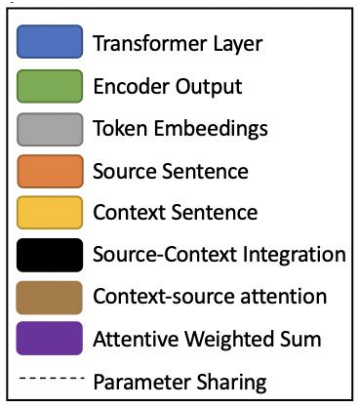


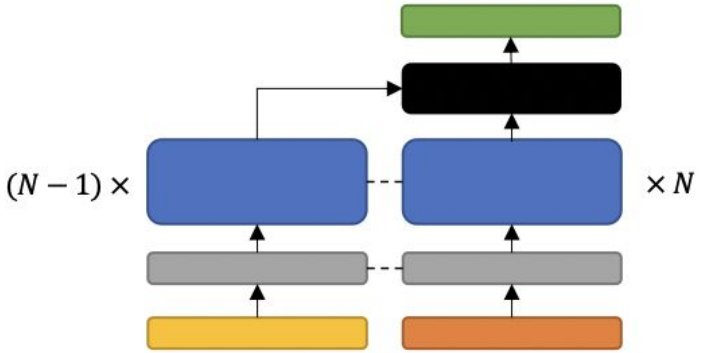


Model Frameworks

5 transformer-based NMT models used in experiment

- Transformer without contexts (TwoC)
- Transformer with contexts (TwC)
- **Discourse Aware Transformer (DAT)**
  - Additional encoder takes in concatenation of context sentences
  - Context and source encoder share same structure with same weights
  - Encoded source and contextual embedding incorporated with a source-to-context attention mechanism and a gated summation

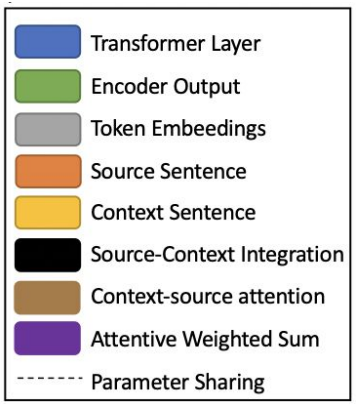


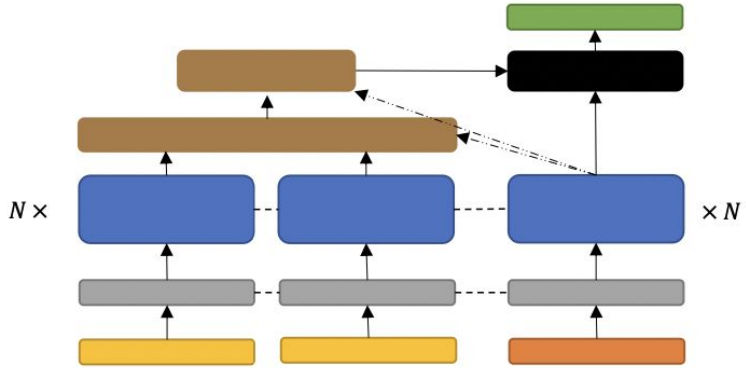


Model Frameworks

5 transformer-based NMT models used in experiment

- Transformer without contexts (TwoC)
- Transformer with contexts (TwC)
- Discourse Aware Transformer (DAT)
- **Hierarchical Attention Networks (HAN)**
  - Each contextual sentence encoded and summarized using word level attention
  - Sentence-level representations concatenated and summarized using sentence-level attention
  - Gated summation of source and contextual embeddings

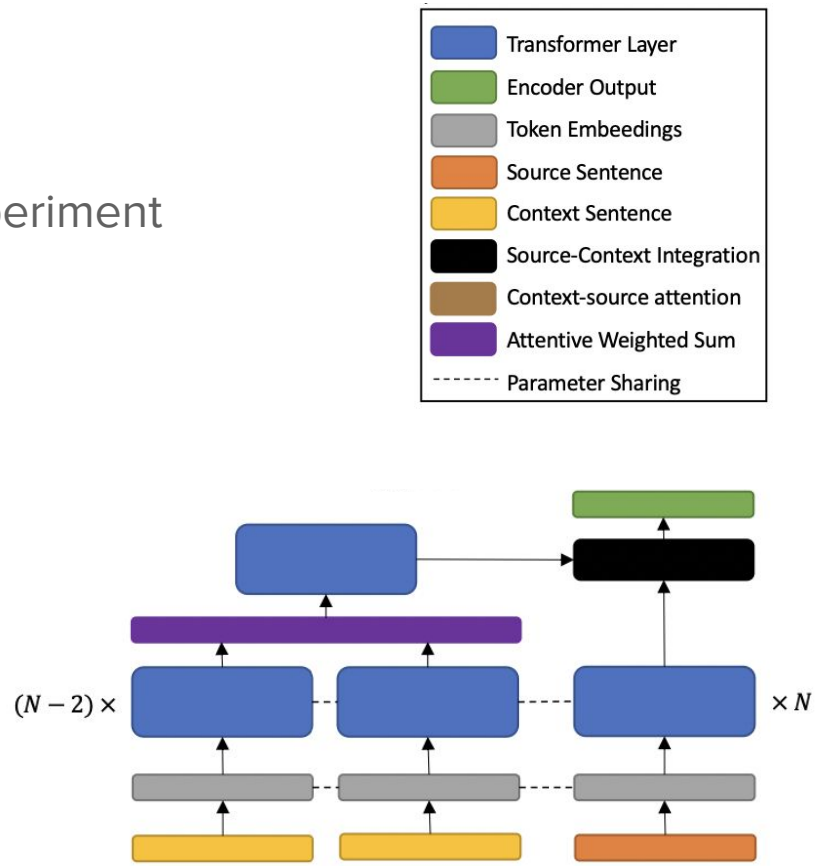




Model Frameworks

5 transformer-based NMT models used in experiment

- Transformer without contexts (TwoC)
- Transformer with contexts (TwC)
- Discourse Aware Transformer (DAT)
- Hierarchical Attention Networks (HAN)
- **Hierarchical Context Encoder (HCE)**
  - Similar to HAN
  - Sentence-level encoding summarized using self-attentive weighted sum module
  - Vectors inputted into another encoder layer to encode into single embedding
  - Combined with source sentence as in HAN

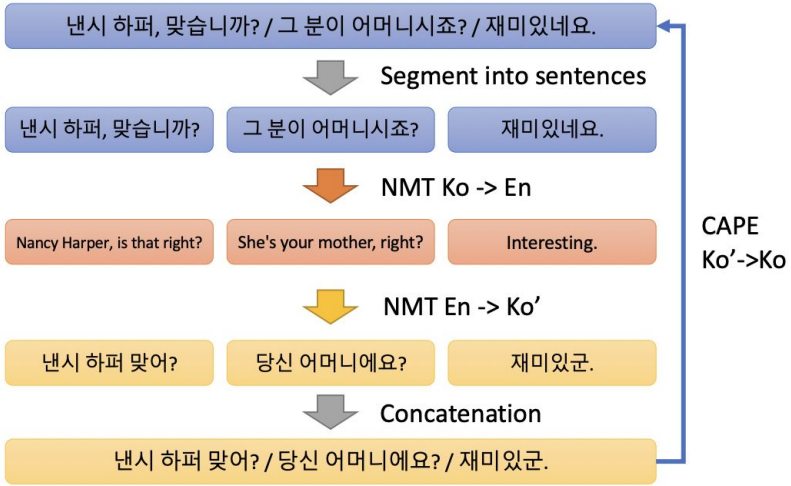


Context-Aware Post Editing (CAPE)

- Variant of automatic post-editing systems
- Independent from MT model
  - Can hypothetically be used to correct translations from any black-box MT system
- First approach to context-aware MT using monolingual data only
  - Voita et al.'s paper describes this approach in more detail
- Some types of inconsistencies can only be identified in the target sentences
  - Eg. Inter-sentence disagreement of honorifics

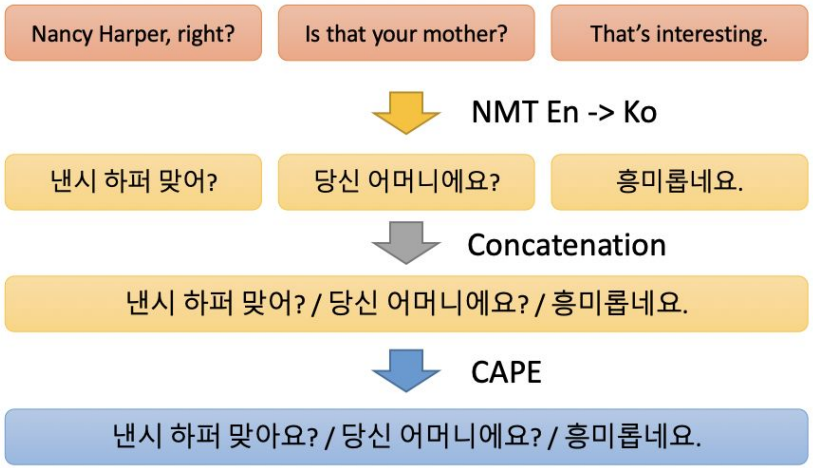
CAPE Training

- Round-trip Translation
  - (Ex. For evaluation of English => Korean NMT)
  - Korean sentences translated to English using Korean => English NMT
  - Back-translated into English using target English => Korean NMT
  - Inconsistencies between original Korean sentences and round-trip translations
- CAPE, a seq-to-seq model, is trained to minimize the inconsistencies



CAPE Testing

- Target NMT translates each sentence first
- CAPE takes group of translated sentences and fixes them to produce final translation



Experiment Model

- Use the NMT model with contextual encoder to capture surrounding sentence context
  - Model should implicitly be able to control output honorific level using context
- Use CAPE to improve inconsistent sentence-level translation of honorifics
  - Use TwoC Korean => English and English => Korean models trained on separate corpus for round-trip translations
- Apply CAPE to NMT models with contextual encoders

Data

- Constructed context-aware parallel corpora
- Used bilingual English-Korean corpus of subtitles of movies/TV shows
  - Contains many honorific expressions
- Obtained contextual sentences using timestamp based heuristics
  - Sentences appear within a short period of time (3000 ms) => contains contextual information
- Additional Korean sentences from subtitles to train CAPE

- Developed rule-based honorific labeling scheme
  - Label as honorific/non-honorific
  - Extract morphologies and POS tags and look at the sentence endings
- Used to label test set to evaluate experimental method

Results

Metrics

- Tokenized BLEU scores (Tokenize translations before scoring)
- Honorific accuracy: Ratio of translations with same type of honorific as reference translation

Results: Contextual Encoders

- All contextual encoder models outperform non-contextual model (TwoC) in terms of BLEU score
- Difference in performance of contextual models could be the difference in how the contextual information is encoded.

Models	BLEU		Accuracy	Accuracy
	En-Ko	Ko-En	All Test Set	Polite Targets
TwoC	9.16/12.45	23.81	64.34	39.27
TwC	9.6/13.2	24.35	66.85	**44.08**
DAT [6]	9.36/12.98	23.96	65.12	38.7
HAN [28]	9.50/13.08	24.54	66.3	42.26
HCE [34]	**10.23/14.75**	**26.63**	**67.94**	42.42

Results: Contextual Encoder w/ Special Token

- In past research, honorific translation was done by annotating the training data with a special token to indicate the honorific style of the target sentence.
  - Eg. Adding <F> or <I> to indicate a formal or informal target sentence
- Tried using this approach with a contextual encoder model (HCE)
- Performance improves using the contextual encoder model

Models	BLEU	Accuracy All Test Set	Accuracy Polite Targets
TwoC + Special Token	9.36/12.68	99.46	98.91
HCE + Special Token	**10.83/14.79**	**99.49**	**99.04**

Results: Number of Contextual Sentences

- Experiment to examine effect of number of contextual sentences used
- Too much information hurts performance
- 2 contextual sentences shows best performance

Table 4. English-Korean BLEU scores (normal/tokenized) and accuracy (%) by the number of contextual sentences on HCE

# Contextual Sents.	BLEU	Accuracy All Test Set	Accuracy Polite Targets
1	9.23/12.88	65.42	40.31
2	**10.23/14.75**	**67.94**	**42.42**
3	9.83/13.49	66.56	41.93
4	9.31/12.92	64.8	39.27
5	8.98/12.09	63.3	36.48

Results: CAPE

- CAPE improves performance for both contextual and non-contextual models
- Improvement in honorific accuracy suggests CAPE can repair honorific inconsistency in output
- Improvement in HCE shows CAPE (trained with TwoC models) can be applied to other MT models (including context-aware ones)

Models	BLEU	Accuracy All Test Set	Accuracy Polite Targets
TwoC	9.16/12.45	64.34	39.27
+CAPE	**10.03/14.38**	**67.5**	**43.81**
HCE	10.23/14.65	67.94	42.42
+CAPE	**10.55/15.03**	**69.16**	**46.51**

Results: Examples of Translations by Models

- First example: HCE correctly outputs honorific style with TwoC fails to do so

- Second example: TwC correctly outputs honorific style while HCE could not
  - Hypothesize that TwC's (simple and direct use of context) leads to better performance when contextual sentences are simple and short (eg. Daddy!)

En (Context_1)	Life must go on as it always has. (언제나처럼 인생은 계속되어야죠.)
En (Context_0)	Come on, let's eat. (어서 먹자.)
En (Source)	How's mom?
Ko (TwoC)	엄마는 어때?
Ko (HCE)	엄마는 어떠세요?
Ko (Reference)	엄마는 어때요?

En (Context_1)	Daddy! (아빠!)
En (Context_0)	Hey, I'm here. (어, 나 여기 있어.)
En (Source)	I'm sorry. I should have listened to **you**.
Ko (TwoC)	미안해. **네** 말을 들었어야 했는데.
Ko (HCE)	미안해. **네** 말을 들었어야 했어.
Ko (TwC)	죄송해요. **아빠** 말을 들었어야 했는데요.
Ko (Reference)	미안해요. **아빠** 말을 들을 걸 그랬어요.

Conclusion

- Context-aware NMT models can translate sentences with the proper honorific style by capturing speaker information
- CAPE can correct inconsistent use of honorifics and improve translation
- Context-aware NMT can improve prior special token methods

Future work

- Apply method to other languages which have complex and widely used honorific systems

Discussion Question:

In this paper, we saw how context-aware NMT can be used to improve honorific translation. What is another potential application of context-aware MT?

Other application of context-aware NMT

- Russian pronoun resolution
  - Context-Aware Neural Machine Translation Learns Anaphora Resolution
  - Voita et al., 2018

- Writing style & terminology improvements based on context
  - Context-aware Neural Machine Translation with Mini-batch Embedding
  - Morishita et al., 2021

Paper 2: Measuring and Increasing Context Usage in Context-Aware Machine Translation (2021)

Venue: ACL 2021

Patrick Fernandes[1,2,3] **Kayo Yin**[1] **Graham Neubig**[1] **André F. T. Martins**[2,3,4]

- Motive: How do we measure how much context is actually captured in the models?
  - Many papers present models that can *theoretically* capture context

- Paper Contribution
  - New metric, **conditional cross-mutual information**, to quantify context usage
    - Empirical analysis of metric
  - New training method, **context-aware word dropout**, to increase context usage

Current methods of evaluation

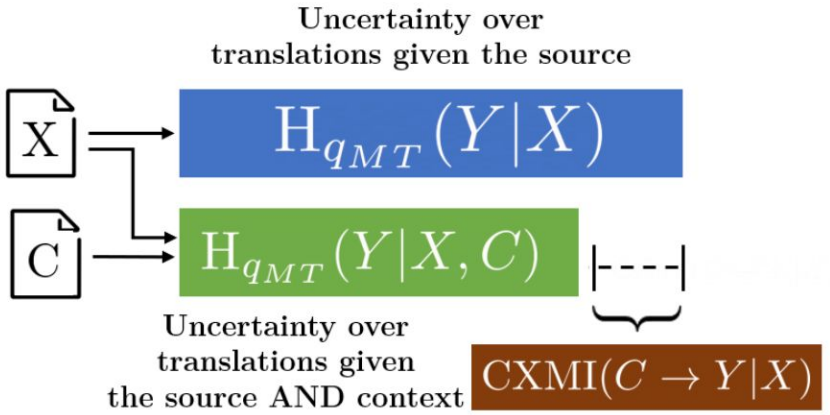
- Methods of Evaluation
  - BLEU Score
    - Ill equipped to measure gains from additional context
  - Contrastive evaluation to measure performance on inter-sentential discourse phenomena
    - Assessed on ability to distinguish correct translations from contrastive ones
    - This paper introduces a large-scale test set focused specifically on pronoun translation
    - However, confined to narrow set of phenomena
    - Can fail to capture other, unknown ways context could be captured
      - Most improvements to translation quality due to non-interpretable usages of context (eg. noise regularizes encoder/decoder)
      - No clear definition of “context”

- No method to measure impact of context in a broader setting

Proposed metric: Conditional cross-mutual information (CXMI)

Measures quantitatively how much context-aware models use the provided context **by comparing the model distributions over a dataset with and without context**.

- Applies to any probabilistic context- aware model
- Models should hypothetically be a single model able to translate with or without context => Eliminates extraneous factors



Conditional cross-mutual information (CXMI)

Measures quantitatively how much context-aware models use the provided context **by comparing the model distributions over a dataset with and without context**.

$$\mathbf{CXMI}(C \to Y|X) = \mathbf{H}_{q_{MT_A}}(Y|X) - \mathbf{H}_{q_{MT_C}}(Y|X,C)$$

H_{qMTA} = Entropy of a context-agnostic model H_{qMTC} = Entropy of a context-aware model

$$\mathbf{CXMI}(C \to Y|X) \approx -\frac{1}{N}\sum_{i=1}^{N} \log \frac{q_{MT_A}(y^{(i)}|x^{(i)})}{q_{MT_C}(y^{(i)}|x^{(i)}, C^{(i)})}$$

qMTA = Context-agnostic model == qMTC = Context-aware model

Data

- Pretraining: Paracrawl to get 82M and 104M sentence pairs
- Training: IWSLT 2017 dataset for EN => DE and EN => FR pairs (200K sentences)
- Validation: 2011-2014 datasets
- Testing: 2015 dataset

- Evaluation using two contrastive datasets
  - ContraPro (Müller et al., 2018) for EN => DE: Targets anaphoric pronoun resolution
  - Bawden et al. (2018) dataset for EN => FR: Anaphoric pronoun resolution + Lexical cohesion

Experiment

Models

- Transformer encoder-decoder architecture
- Include context by concatenating context (source/target) to source sentence
- Pre-train two models (one for source, one for target) to receive context of size 0 to 5 => Can translate with context of any size in interval
- Measure CXMI for source and target models (both with and without pretraining) for different context sizes

Results

- For non-pretrained model, target context captured slightly more than source
- For stronger, pretrained model, target context captured while source hardly captured at all
- Biggest jump in context usage is increasing from 0 to 1 context sentences
    - After that, diminishing increases or reduction in context usage

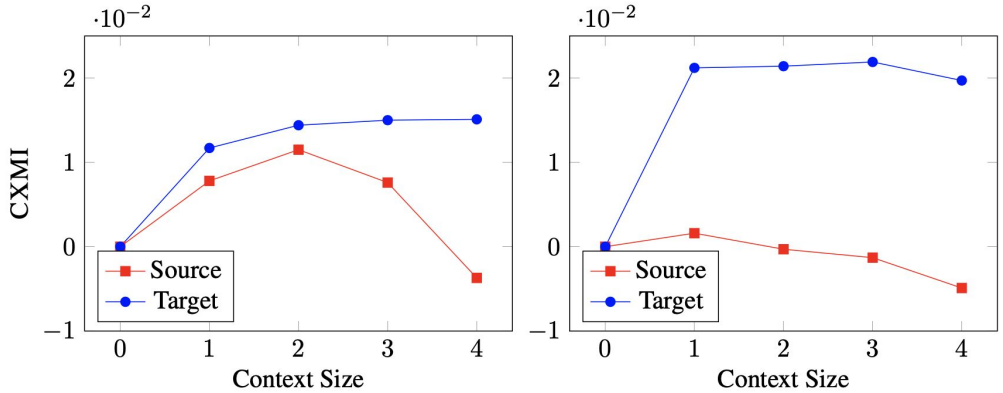


Figure 2: CXMI values for the EN → DE as a function of source and target context sizes for non-pretrained (left) and pretrained (right) models.

Does CXMI really capture context usage?

- Perform correlation analysis with performance in contrastive datasets
- Compare CXMI score and binary variable that takes value 1 if contextual model picks correct translation or non-contextual model picks wrong translation
- Conclude that CXMI captures previous measures of context usage to some extent

	r_{pb}		
Context Size	(1)	(2)	(3)
1	**0.365**	**0.315**	**0.206**
2	**0.366**	-	-
3	**0.367**	-	-
4	**0.366**	-	-

Table 2: Point-Biserial correlation coefficients on the contrastive datasets with pretrained models for different context sizes. Measured on *ContraPro* (1) and Bawden et al. (2018), both for pronoun resolution (2) and lexical cohesion (3). Bold values mean the correlation is statistically significant with $p < 0.01$.

Proposed training method: Context-aware Word Dropout (COWORD Dropout)

- Training method to increase context usage of models
  - Based on word dropout, a popular regularization technique used in sentence-level MT
- Results from first experiment showed limited usage of context with standard MLE training

COWORD Dropout

- Randomly drop words from current source sentence and replace with MASK token with probability *p*
  - Do not drop words from the context: Provides inductive bias that context important
- Encourages model to use extra-sentential information to compensate for missing information

$$p_\theta(y^{(i)}|x^{(i)}) = \prod_{t=1}^{T} p_\theta(y_t^{(i)}|\tilde{x}^{(i)}, y_{<t}^{(i)}, C^{(i)})$$

$$r_t^{(i)} \sim \text{Bernoulli}(p)$$

$$\tilde{x}_t^{(i)} = \begin{cases} \langle \text{MASK} \rangle & \text{if } r_t^{(i)} = 1 \\ x_t^{(i)} & \text{otherwise.} \end{cases}$$

Experiment

Context Usage

- Same models as in first experiment
  - Target context only model
  - Target + source context model
- Dynamic context size
- Disable COWORD dropout during test: Used only during training
- Measure CXMI values as function of context size for different dropout values *p*

Translation Quality

- 3 fixed-size models: No context (baseline), Target context (one-to-two), Source + Target context (two-to-two)
- Explore COWORD's effects on other architectures
- Standard BLEU score & COMET score
- Contrastive Datasets

Results: Context Usage

- CXMI score increases with increase in *p*, i.e. more dropout
  - Increased context usage

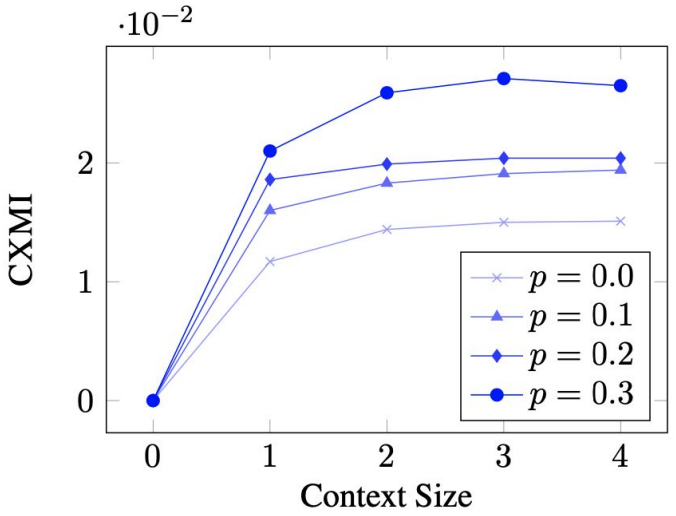


Figure 3: CXMI values as a function target context size for different values of CoWord dropout

Results: Context Usage

- Models trained with COWORD seem to recognize the words that use the most context according to CXMI score

Source Context	Source	Target Context	Target	ΔCXMI
More people watched games because it was faster.	It was more entertaining	Mehr Menschen sahen zu, die Spiele wurden schneller	und unterhaltsamer.	0.53
The ball comes off track.	You don't know where it's going to land	Der Ball ist außer Kontrolle	Sie wissen nicht, wo er landet.	0.33
I really think that this lie that we've been sold about disability is the greatest injustice	It makes life hard for us	Meiner Meinung nach ist diese Luge über Behinderung eine schreiende Ungerechtigkeit	Sie macht uns das Leben schwer.	0.25

Table 3: Examples where models with CoWord dropout use the target context more than models trained without it. Word highlighted blue in the context are used to disambiguate translations while highlighted green in the target use context according to native speakers. Words underlined in the target are the ones with the highest *per-word* CXMI i.e. the ones that use the most context according to the model

Results: Translation Quality w/ BLEU & COMET

- Models trained with COWORD generally outperform models trained without
- More improvement seen for models without pre-training
  - Better models harder to improve?
- Helpful for other architectures that don't use concatenation

		EN → DE				EN → FR			
				w/ pretraining				w/ pretraining	
	p	BLEU	COMET	BLEU	COMET	BLEU	COMET	BLEU	COMET
baseline	0.0	26.36	0.083	35.10	0.521	37.62	0.450	42.98	0.679
	0.1	**27.26**	0.159	**35.15**	**0.525**	38.16	0.472	**43.28**	**0.679**
	0.2	26.97	**0.163**	35.13	0.524	**38.34**	**0.474**	42.99	0.678
1-to-2	0.0	26.60	0.087	**35.22**	**0.528**	37.59	0.450	42.89	0.672
	0.1	**27.36**	0.174	34.92	0.527	38.25	0.472	42.88	0.677
	0.2	27.33	**0.193**	34.75	0.524	**38.27**	**0.485**	**42.90**	**0.678**
2-to-2	0.0	26.85	0.090	34.47	0.471	37.54	0.453	**42.97**	0.674
	0.1	**27.72**	0.169	34.51	0.522	**38.30**	0.467	42.95	**0.676**
	0.2	27.21	**0.177**	**34.65**	**0.525**	38.15	**0.468**	42.88	0.675

Table 4: Results on IWSLT2017 with different probabilities for CoWord dropout. Averaged across three runs for each method.

		EN → DE		EN → FR	
	p	BLEU	COMET	BLEU	COMET
baseline	0.0	26.36	0.083	37.62	0.450
	0.1	**27.26**	0.159	38.16	0.472
	0.2	26.97	**0.163**	**38.34**	**0.474**
multi	0.0	26.64	0.104	37.85	0.466
	0.1	**27.45**	0.190	37.98	0.460
	0.2	27.31	**0.190**	**38.30**	**0.484**

Table 5: Results on IWSLT2017 for a multi-encoder 1-to-2 model with different probabilities for CoWord dropout. Averaged across three runs for each method.

Results: Translation Quality w/ Contrastive Datasets

- Generally improved performance with COWORD dropout
- More improvement for models not pre-trained
- Greater increase in performance for EN => DE pronoun resolution
  - Smaller dataset for EN => FR contrastive dataset
- Improves performance for other architectures

		EN → DE		EN → FR			
			w/ pretraining			w/ pretraining	
	p	Pronouns	Pronouns	Pronouns	Cohesion	Pronouns	Cohesion
baseline	0.0	42.96	46.57	50.00	50.00	50.00	50.00
1-to-2	0.0	57.36	76.79	68.16	49.99	**86.83**	**56.83**
	0.1	58.70	76.28	72.33	51.49	86.49	56.66
	0.2	**60.72**	**77.52**	**72.99**	**52.16**	85.66	56.49
2-to-2	0.0	61.06	80.33	72.16	50.99	85.66	64.33
	0.1	**66.00**	**80.35**	**73.99**	**52.49**	87.16	**65.99**
	0.2	65.47	79.97	**73.99**	**52.49**	**88.49**	63.99

Table 6: Results on anaphoric pronoun resolution and lexical cohesion contrastive datasets with different probabilities for CoWord dropout. Averaged across three runs for each method.

		EN → DE	EN → FR	
	p	Pronouns	Pronouns	Cohesion
baseline	0.0	42.96	50.00	50.00
multi	0.0	42.85	49.74	49.99
	0.1	47.29	51.74	50.24
	0.2	**47.57**	**52.50**	**50.99**

Table 7: Results on anaphoric pronoun resolution and lexical cohesion contrastive datasets for the multi-encoder 1-to-2 model with different probabilities for CoWord dropout. Averaged across three runs for each method.

Implications / Future Work

- New, architecture-agnostic metric to measure context usage in context-aware MT models
- Simple regularization technique to increase context usage by models
- Applicable to most recently proposed context-aware models
- Can be used in future work
  - For training to increase context usage
  - For evaluation of context usage

Discussion Question:

Paper 2 suggests a metric to capture context usage in models. Do you think this metric truly tells us about the context captured in the models? What might be another way to measure the context captured?

Evaluation Paper

[1912.08494] A Survey on Document-level Neural Machine Translation: Methods and Evaluation

Table 5: Overview of works which introduce techniques to evaluate discourse phenomena in MT. Some of the referenced works do not consider inter-sentence context information.

Evaluation Type	Discourse Phenomena	Dependency (Resource or Language)	Reference
Automatic Metric	Pronouns	Alignments, Pronoun lists	[40, 84]
		English in target (anaphoric)	[53]
	Lexical Cohesion	Lexical cohesion devices	[133]
		Topic model, Lexical chain	[31]
	Discourse Connectives	Alignments, Dictionary	[37]
		Discourse parser	[36, 49, 111]
Test Suite	Pronouns	En→Fr	[34]
		En→Fr (anaphora)	[8]
		En→De (anaphora)	[85]
	Cohesion	En→Fr	[8]
		En→Ru	[128]
	Coherence	En→Fr	[8]
		En↔De, Cs↔De, En→Cs	[130]
		En→Cs	[98]
	Conjunction	En/Fr→De	[91]
	Deixis, Ellipsis	En→Ru	[128]
	Grammatical Phenomena	En→De	[103]
		De→En	[3]
	Word Sense Disambiguation	De→En/Fr	[96, 95]
		En↔De/Fi/Lt/Ru, En→Cs	[93]

Summary

- Sentence-level NMT assumes conditional independence between sentences resulting in issues related to pronoun resolution and lexical cohesion.
- Context-aware MT takes source and/or target-side context into account to improve translation quality.
- Monolingual data can be used to repair translations outputted by NMT models.
- We can also develop new training methods for context-aware MT to further improve translation quality.
- Evaluation of context-aware NMT is difficult as standard metrics like BLEU fail to measure context usage

Neural Machine Translation: Transformer

- Tensor2Tensor: Library of deep learning models and datasets with official Transformer implementation
- Quick start Colab notebook: https://colab.research.google.com/github/tensorflow/tensor2tensor/blob/master/tensor2tensor/notebooks/hello_t2t.ipynb
- Pytorch Implementation: https://colab.research.google.com/github/graykode/nlp-tutorial/blob/master/5-1.Transformer/Transformer.ipynb

Context-aware NMT

- CADec (Context-aware Decoder) and DocRepair (Context-aware Monolingual Repair): https://github.com/lena-voita/good-translation-wrong-in-context
- Measuring and Increasing Context Usage in Machine Translation (Paper 2): https://github.com/neulab/contextual-mt