When Does Translation Require Context? A Data-driven, Multilingual Exploration

Anonymous ACL submission

Abstract

Although proper handling of discourse phenomena significantly contributes to the quality of machine translation (MT), improvements on these phenomena are not adequately measured in common translation quality metrics. Recent works in context-aware MT attempt to target a small set of these phenomena during evaluation. In this paper, we propose a methodology to identify translations that require context systematically, and use this methodology to both confirm the difficulty of previously studied phenomena as well as uncover new ones that have not been addressed in previous work. We then develop the **Mu**ltilingual **D**iscourse-**A**ware (MuDA) benchmark, a series of taggers for these phenomena in 14 different language pairs, which we use to evaluate context-aware MT. We find that commonly studied context-aware MT models make marginal improvements over context-agnostic models, which suggests these models do not handle these ambiguities effectively. We will release code and data to invite the MT research community to increase efforts on translation on discourse phenomena and languages that are currently overlooked.

1 Introduction

In machine translation (MT), information from previous utterances has been found crucial to adequately translate a number of discourse phenomena including anaphoric pronouns, lexical cohesion, and discourse markers (Guillou et al., 2018; Läubli et al., 2018; Toral et al., 2018). However, while generating proper translations of these phenomena is important, they represent only a small portion of the words in natural language data. Because of this, common metrics such as BLEU (Papineni et al., 2002) do not provide a clear picture of whether they are appropriately captured or not.

Recent work on neural machine translation (NMT) models that attempt to incorporate extra-sentential context (Tiedemann and Scherrer, 2017; Miculicich et al., 2018; Maruf and Haffari, 2018, *inter alia*) often perform targeted evaluation of certain discourse phenomena, mostly focusing on ellipsis, formality (Voita et al., 2019b,a), and pronoun translation (Müller et al., 2018; Bawden et al., 2018; Lopes et al., 2020). However, only a limited set of discourse phenomena for a few language pairs have been studied (see summary in Table 1). The difficulty of broadening these studies stems from the reliance of previous work on introspection and domain knowledge to identify the relevant discourse phenomena, frequently involving expert speakers, which then requires engineering complex language-specific methods to create test suites or manually designing data for evaluation.

Dataset	Lang.	Phenomena
Müller et al. (2018)	EN → DE	Pronouns
Bawden et al. (2018)	EN → FR	Pronouns, Coherence Lexical Consistency
Voita et al. (2018) Voita et al. (2019b)	EN → RU	Pronouns Deixis, Ellipsis Lexical Consistency
Jwalapuram et al. (2020)	DE → EN FR → EN RU → EN	Pronouns, Coherence Lexical Consistency Discourse Connectives
Our Work	14 Pairs (§5)	Pronouns, Ellipsis Formality Lexical Consistency Verb Forms

Table 1: Some representative works on contextual machine translation that perform evaluation on discourse phenomena, contrasted to our work. For a more complete review see Maruf et al. (2021).

In this paper, we fill this gap by proposing a *data-driven, semi-automatic methodology for identifying salient phenomena* that require context for translation, and we apply this method to create a *multilingual benchmark testing these discourse phenomena*. This is done through several steps. First, we develop P-CXMI (§2) as a metric to identify when context is helpful in MT, or more broadly text generation in general. Then, we perform a systematic analysis of words with high P-CXMI to find categories of translations where context is useful (§3). This allows us to identify novel discourse phenomena that to our knowledge have not been addressed previously (e.g. consistency of verb forms), without requiring a-priori language-specific knowledge. Finally, we design a series of methods to automatically tag words belonging to the identified classes of ambiguities (§4) and we evaluate existing translation models for different categories of ambiguous translations (§5).

We perform our study on a parallel corpus spanning 14 language pairs, measuring translation ambiguity and model performance. We find that the context-aware methods, while improving on standard evaluation metrics, only perform better than the context-agnostic baselines for certain discourse phenomena in our benchmark, while on other phenomena, context-aware models do not observe significant improvements. Our benchmark therefore provides a more fine-grained evaluation of translation models and reveals the weaknesses of context-aware models, such as verb form cohesion. We also find that DeepL, a commercial document-level translation system, does better in our benchmark than its sentence-level ablation and Google Translate. We hope that the released benchmark and code, as well as our findings, will spur targeted evaluation of discourse phenomena in MT to cover more languages and more phenomena in the future.

2 Measuring Context Usage

2.1 Cross-Mutual Information

While document-level MT models can be compared using standard translation metrics such as BLEU (Papineni et al., 2002), they do not provide a clear picture of whether models are performing better due to improvements in processing context or other improvements (Kim et al., 2019). Another common evaluation paradigm is *contrastive evaluation*, which evaluates contextual models' ability to distinguish between correct and incorrect translations of specific discourse phenomena, such as anaphora resolution (Müller et al., 2018) and lexical cohesion (Bawden et al., 2018). However, this provides only a limited measure of context usage on a limited set of ambiguous phenomena defined by the creators of the dataset, not capturing other unanticipated ways in which the model might need context (Vamvas and Sennrich, 2021). We are therefore interested in devising a metric that is able to capture *all* context usage by a model, beyond a predefined set.

Conditional Cross-Mutual Information (CXMI) (Bugliarello et al., 2020; Fernandes et al., 2021) measures the influence of context on model predictions. CXMI is defined as:

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

where X and Y are a source and target sentence, respectively, C is the context, $\mathrm{H}_{q_{MT_A}}$ is the entropy of a *context-agnostic* MT model, and $\mathrm{H}_{q_{MT_C}}$ refers to a *context-aware* MT model. This quantity can be estimated over a held-out set with N sentence pairs and the respective context as:

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

Importantly, the authors find that training a *single* model q_{MT} as both the context-agnostic and context-aware model ensures that non-zero CXMI values are due to context and not other factors (see Fernandes et al. (2021) and §3.1 for details).

2.2 Context Usage Per Sentence and Word

CXMI measures the context usage by a model by comparing the log-likelihood ratio of samples across *the whole corpus*. However, for our purposes, we are interested in measuring how much the context is helpful for single sentences or even just particular words in a sentence.

Pointwise Mutual Information (P-MI) (Church and Hanks, 1990) measures the association between two random variables for *specific* outcomes. Mutual information can be seen as the expected value of P-MI over all possible outcomes of the variables. Taking inspiration from this, we define the **Pointwise Cross-Mutual Information** (P-CXMI) for a source, target, context triplet (x, y, C) as:

$$\text{P-CXMI}(y,x,C) = -\log\frac{q_{MT_A}(y|x)}{q_{MT_C}(y|x,C)}$$

Intuitively, P-CXMI measures how much more (or less) likely a target sentence y is when it is given context C, compared to not being given that context. Note that this is estimated *according to the models* q_{MT_A} *and* q_{MT_C} since, just like CXMI, this measure depends on their learned distributions.

We can also apply P-CXMI at the *word level* (as opposed to the sentence level) to measure how much more likely a particular word in a sentence is when it is given the context, by leveraging the autoregressive property of the neural decoder. Given

Avelile's mother had HIV virus. Avelile had the virus, she was born with the virus. 阿维利尔的母亲是携有艾滋病病毒。 阿维利尔也有艾滋病病毒。她一生下来就有。	Lexical Cohesion
Your daughter? Your niece? *Votre fille ?* Votre nièce ?	Formality (T-V)
Roger. I got'em. Two-Six, this is Two-Six , we're mobile. 了解 捕捉した。 2-6 こちら移動中だ。	Formality (Honorifics)
Our tools today don't look like shovels and picks. They look like the stuff we walk around with. *As ferramentas de hoje não se parecem com pás e picaretas.* Elas se parecem com as coisas que usamos.	Pronouns
Louis XIV had a lot of people working for him. They made his silly outfits, like this. *Luis XIV tenía un montón de gente trabajando para él.* Ellos hacían sus trajes tontos, como éste.	Verb Form
They're the ones who know what society is going to be like in another generation. I don't. *Ancak onlar başka bir nesilde toplumun nasıl olacağını biliyorlar.* Ben bilmiyorum.	Ellipsis

Table 2: Examples of high P-CXMI tokens and corresponding linguistic phenomena. Contextual sentences are *italicized*. The high P-CXMI target token is highlighted in pink, source and contextual target tokens related to the high P-CXMI token are highlighted in blue and green respectively.

the triplet (x, y, C) and the word index i, we can measure the P-CXMI for that particular word as:

$$\text{P-CXMI}(i, y, x, C) = -\log \frac{q_{MT_A}(y_i|y_{t<i}, x)}{q_{MT_C}(y_i|y_{t<i}, x, C)}$$

Note that nothing constrains the form of C or even x and P-CXMI can, in principle, be applied to any conditional language modelling problem.

Using this metric, we now ask: what kind of words tend to see their likelihood increase when given the context? Such words should have a high P-CXMI, which we examine in the following §3.

3 Which Translation Phenomena Benefit from Context?

To identify salient translation phenomena that require context, we perform a *thematic analysis* (Braun and Clarke, 2006), examining words with high P-CXMI across different language pairs and manually identifying patterns and categorizing them into phenomena where context is useful for translation. To do so, we systematically examined (1) the mean P-CXMI per POS tag, (2) the vocabulary items with the highest P-CXMI, and (3) the individual tokens with the highest P-CXMI.

3.1 Data & Model

To compare linguistic phenomena that arise during document-level translation across various language pairs, we need a dataset that is document-level, rich in context-dependent discourse phenomena, and parallel in multiple languages. We, therefore, perform our study on transcripts of TED talks and their translations (Qi et al., 2018). We choose to study translation between English and Arabic, German, Spanish, French, Hebrew, Italian, Japanese, Korean, Dutch, Portuguese, Romanian, Russian, Turkish and Mandarin Chinese. These 14 target languages are chosen for their high availability of TED talks and linguistic tools, as well as for the diversity of language types in our comparative study (Table 8 in Appendix A). For each language pair, our dataset contains 113,711 parallel training sentences from 1,368 talks, 2,678 development sentences from 41 talks, and 3,385 testing sentences from 43 talks.

To obtain the P-CXMI for words in the data, we train a small Transformer (Vaswani et al., 2017) model for every target language and incorporate the target context by concatenating it to the current target sentence (Tiedemann and Scherrer, 2017). We train the model with *dynamic* context size (Fernandes et al., 2021), by sampling between 0 and 3 target context sentences and estimate P-CXMI by using this model both q_{MT_A} and q_{MT_C} (more training details in Appendix D).

3.2 Analysis Procedure

We adopt a top-down approach and start our analysis by studying POS tags with high mean P-CXMI. In Appendix B, we report the mean P-CXMI for selected POS tags on our test data. Some types of ambiguity, such as dual form pronouns (§3.3), can be linked to a single POS tag and be identified at this step, whereas others require finer inspection.

Next, we inspect the vocabulary items with high mean P-CXMI. At this step, we can detect phenomena that are reflected by certain lexical items that consistently benefit from context for translation.

Finally, we examine individual tokens that obtain the highest P-CXMI. In doing so, we identify patterns that do not depend on lexical features, but rather on syntactic constructions for example. In Table 2, we provide selected examples of tokens that have high P-CXMI and the discourse phenomenon we have identified from them.

3.3 Identified Phenomena

Through our thematic analysis of P-CXMI, we identified various types of translation ambiguity. Unlike previous work, our method requires no prior knowledge of the languages to find relevant discourse phenomena and easily scales to new languages (§4.4).

First, we find high P-CXMI for second-person pronouns (PRON.2) in languages with T-V distinction (Appendix A, "Pronouns Politeness"). While English uses the same second-person pronouns for everyone, in these languages, certain pronouns depend on the level of **formality** and relationship between the speaker and addressee. Furthermore, languages such as Japanese and Korean use honorifics to indicate formality. In Japanese, vocabulary items such as "ございい" / "じゃ" that control formality have high mean P-CXMI (0.42 / 0.34).

In English, only the 3rd person singular pronoun is gendered and gender is assigned based solely on semantic rules (Appendix A, "Gendered Pronouns", "Gender Assignment"). We find several languages with high P-CXMI on pronouns (PRON), and these languages use gendered pronouns for pronouns other than the 3rd person singular or assign gender using formal rules (German, French, Hebrew, Italian, Portuguese, Russian, and Chinese). When translating a gender-neutral English pronoun to a gendered target pronoun, context is therefore needed to determine the gender of the antecedent.

We find high P-CXMI for certain **verb forms**, such as the imperfect form in Spanish Italian and Romanian (VERB.Imp). While English verbs may have five forms (e.g. *write, writes, wrote, written, writing*), other languages often have a more fine-grained verb morphology. For example, English has only a single form for the past tense, while the Spanish past tense consists of six verb forms. Verbs must be translated using the verb form that reflects the tone, mood and cohesion of the document.

When we inspect vocabulary items with the highest mean P-CXMI scores, we often find names of entities (e.g. the Japanese translation of Mandela " マンデラ " has mean P-CXMI of 0.36). As in the first row of Table 2, proper nouns may have multiple possible translations, but the same entity should be referred to by the same word in a translated document for **lexical cohesion** (Carpuat, 2009).

Finally, among the individual tokens with the highest P-CXMI, we find that many are due to **ellipsis** in the English sentence that does not occur on the target side. For example, in the last row of Table 2, the English text does not repeat the verb *know* in the second sentence as it can be understood from the previous sentence. However, in Turkish, there is no natural way to translate the verb-phrase ellipsis and must infer that "don't" refers to "don't *know*", and translate the verb accordingly.

	pronouns	formality	verb form	lexical	ellipsis
ar	90	0	0	116	982
de	398	1000	0	19	1356
es	245	86	409	15	1496
fr	1591	839	1938	48	1586
he	0	0	468	122	1210
it	182	118	484	31	1320
ja	245	3328	0	94	990
ko	0	221	0	71	373
nl	0	783	1060	27	1590
pt_br	372	515	0	27	1677
ro	60	407	792	53	1002
ru	0	466	2091	41	668
tr	0	30	47	137	704
zh_cn	0	526	0	49	1092

Table 3: Number of MuDA tags on TED test data.

Although this procedure may tend to find phenomena that are intuitive to the annotators, the data-driven approach makes confirmation bias less severe than prior works relying on introspection to identify phenomena. Hence, our procedure can allow us to discover relevant phenomena that have not been previously addressed, such as verb forms.

4 Cross-phenomenon MT Evaluation

After identifying a set of linguistic phenomena where context is useful to resolve ambiguity during translation, we develop a series of methods to automatically tag tokens belonging to these classes of ambiguous translations and propose the **Mu**ltilingual **D**iscourse-**A**ware (MuDA) benchmark for context-aware MT models.

4.1 MT Evaluation Framework

Given a pair of parallel source and target documents (X, Y), our MuDA tagger assigns a set of discourse phenomena tags $\{t_i^1, \cdots, t_i^n\}$ to each target token $y_i \in Y$. Then, using the `compare-mt` toolkit (Neubig et al., 2019), we compute the mean word f-measure of system outputs compared to the reference for each tag. This allows us to identify which discourse phenomena models can translate more or less accurately.

4.2 Automatic Tagging

In this section, we describe our taggers for each discourse phenomenon we identified. In doing so, we create more reliable and informative taggers for each phenomenon, rather than using P-CXMI directly to identify ambiguous words, as P-CXMI is fairly noisy and uninterpretable. For the formality, pronoun choice and verb forms tags, we created language-specific word lists that were verified by native speakers, and these tags are only applicable to certain target langauges that contain the associated discourse phenomenon.

Lexical Cohesion To tag words that require lexical cohesion, we first extract word alignments from a parallel corpus $D = \{(X_1, Y_1), \cdots, (X_{|D|}, Y_{|D|})\}$, where (X_m, Y_m) denote the source and target reference document pair. We use the AWESOME aligner (Dou and Neubig, 2021) to obtain:

$$A_m = \{\langle x_i, y_j \rangle \mid x_i \leftrightarrow y_j, x_i \in X_m, y_j \in Y_m\},$$

where each x_i and y_j are the lemmatized content source and target words and $\leftrightarrow$ denotes a bidirectional word alignment. Then, for each target word y_j that is aligned to source word x_i, if the alignment pair $\langle x_i, y_j \rangle$ occurred at least 3 times already in the current document, excluding the current sentence, we tag y_j for lexical cohesion.

Formality For languages with T-V distinction, we tag the target pronouns containing formality distinction in their various forms, if there has previously been a word pertaining to the same formality level in the same document. Some languages such as Spanish often drop the subject pronoun, and T-V distinction is instead reflected in the verb form. For these languages, we use spaCy (Honnibal and Montani, 2017) and Stanza (Qi et al., 2020) to find POS tags and detect verbs with a second-person subject in the source, and conjugated in the second (T) or third (V) person in the target. For languages with a more complex honorifics system, such as Japanese, we construct a word list of common honorifics-related words to tag (details in Appendix C).

Pronoun Choice To find pronouns in English that have multiple translations, we manually construct a list $P_\ell = \{\langle p_s, \mathbf{p}_t \rangle\}$ for each language (Appendix C), where each p_s is an English pronoun and $\mathbf{p}_t$ the list of possible translations of p_s in the language ℓ. Then, for each aligned token pair $\langle x_i, y_j \rangle$, if x_i, y_j are both pronouns with $\langle x_i, \mathbf{p}_t | y_j \in \mathbf{p}_t \rangle \in P_\ell$, and the antecedent of x_i is *not* in current sentence, we tag y_j as an ambiguous pronoun. To obtain antecedents, we use AllenNLP (Gardner et al., 2017)'s coreference resolution module. This procedure is similar to Müller et al. (2018).

Verb Form For each target language, we define a list $V_\ell = \{v_1, \cdots, v_k\}$ of verb forms (Appendix C) where $v_i \in V_\ell$ if there exists a verb form in English u_j and an alternate verb form $v_k \neq v_i$ in the target language such that an English verb with form u_j may be translated to a target verb with form v_i or v_k depending on the context. Then, for each target token y_j, if y_j is a verb of form $v_j \in V_\ell$, and another verb with form v_j has appeared previously in the same document, we tag y_j as ambiguous.

Ellipsis To detect translation ambiguity due to VP and NP ellipsis, we look for instances where the ellipsis occurs on the source side, but not on the target side, which means that the ellipsis must be resolved during translation. Since existing ellipsis models are limited to specific types ellipsis, we first train an English (source-side) ellipsis detection model. To do so, we extract an ellipsis dataset from the English data in the Penn Treebank (Marcus et al., 1993) and train a BERT text classification model (Devlin et al., 2019), which achieves 0.77 precision and 0.73 recall (see Appendix C for training details). Then, for each sentence pair where the source sentence is predicted to contain an ellipsis, we tag the word y_j in the target sentence Y_m if: (1) y_j is a verb, noun, proper noun or pronoun; (2) y_j has occurred in the previous target sentences of the same document; (3) y_j is not aligned to any source words, that is, $\nexists x_i \in X_m$ s.t. $\langle x_i, y_j \rangle \in A_m$.

4.3 Evaluation of Automatic Tags

We apply the MuDA tagger to the reference translations of our TED talk data. We thus obtain an evaluation set of 3,385 parallel sentences for each of the 14 language pairs. In Appendix B we report the mean P-CXMI for each language and MuDA tag. Overall, we find higher P-CXMI on tokens with a tag compared to those without, which provides empirical evidence that models indeed rely on context to predict words with MuDA tags.

Table 3 shows that the frequency of tags varies significantly across languages. Overall, ellipses are infrequent, as only 4.5% of the English sentences have been marked for ellipsis which gives an upper bound for the number of ellipsis tags. We suggest our tagger to be applied on a large evaluation set to contain enough examples of ellipsis. Further, languages from a different family than English have a relatively high number of ellipsis tags. Korean and especially Japanese have more formality tags than languages with T-V distinction, which is aligned

	lexical	formality	pronouns	verb form	ellipsis
es	1.00	0.92	1.00	1.00	0.53
fr	1.00	1.00	1.00	0.94	0.43
ja	1.00	1.00	1.00	–	0.41
ko	1.00	0.94	–	–	0.26
pt	0.99	0.88	1.00	–	0.31
ru	1.00	1.00	–	1.00	0.50
tr	1.00	1.00	–	1.00	0.57
zh	1.00	1.00	–	–	0.78

Table 4: Precision of MuDA tags on 50 utterances.

with our intuition that register is more often important when translating to languages with honorifics.

Manual Evaluation To evaluate our tagger, we asked native speakers with computational linguistics backgrounds to manually verify MuDA tags for 8 languages on 50 randomly selected utterances as well as all words tagged with *ellipsis* in our corpus. We paid them 20$/hour. This allows us to measure how many automatic tags violate the given definition of the linguistic tag. Table 4 reports the tags' precision.

For all languages, we obtain high precision for all tags except *ellipsis*, confirming that the methodology can scale to languages where no native speakers were involved in developing the tags. For *ellipsis*, false positives often come from one-to-many or non-literal translations, where the aligner does not align all target words to the corresponding source word. We believe that the *ellipsis* tagger is still useful in selecting difficult examples that require context for translation; despite the low precision, we find a significantly higher P-CXMI on *ellipsis* words for many languages (Appendix B).[1]

4.4 Extension to New Languages

While MuDA currently supports 14 language pairs, our methodology can be easily extended to new languages. The *lexical* and *ellipsis* tags can be directly applied to other languages provided a word aligner between English and the new target language. The *formality* tag can be extended by adding a list of pronouns or verb forms related to formality in the new language. Similarly, the *pronouns* and *verb forms* tag can also be extended by providing a list of ambiguous pronouns and verb forms.

Exhaustively listing all relevant phenomena in document-level MT is extremely complex and beyond the scope of our paper. To identify new discourse phenomena on other languages, our thematic analysis can be reused as follows: (1) Train a model with dynamic context size on translation between the new language pair; (2) Use the model to compute P-CXMI for words in a parallel document-level corpus of the language pair; (3) Manually analyze the POS tags, vocabulary items and individual tokens with high P-CXMI; (4) Link patterns of tokens with high P-CXMI to particular discourse phenomena by consulting linguistic resources.

5 Exploring Context-aware MT

Next, we use our MuDA benchmark to perform an initial exploration of context usage across 14 languages pairs and 4 models, including those we trained ourselves and commercial systems.

5.1 Trained Models

We train a sentence-level and document-level concatenation-based small transformer (`base`) for every target language. While conceptually simple, concatenation approaches have been shown to outperform more complex models when properly trained. For the context-aware model, the major difference from §3.1 is that we use a *static* context size of 3, since we are not using these models to measure P-CXMI. (Lopes et al., 2020).

To evaluate stronger models, we additionally train a large transformer model (`large`) that was pretrained on a large, sentence-level corpora, for German, French, Japanese and Chinese. Further training details can be found in Appendix D.

5.2 Commercial Models

To assess if commercially available machine translation engines are able to leverage context and therefore do well in the MuDA Benchmark, we consider two engines:[2] (1) the *Google Cloud Translation* v2 API. In early experiments, we assessed that this model only does sentence-level translation, but included it due to its widespread usage and recognition; (2) the *DeepL* v2 API. This model advertises its usage of context as part of their translations and our experiments confirm this. Early experimentation with other providers (Amazon and Azure) indicated that these are not context-aware so we refrained from evaluating them.

To obtain provider translations, we feed the documents into an API request. To re-segment the translation into sentences, we include special marker tokens in the source that are preserved during translation and split the translation on those tokens. We

[1]Also note that wrongly assigned tags should also not penalize a system greatly as it should give a low score only if the translation does not match the falsely tagged word.

[2]`translate.google.com`, `deepl.com`

		ar	de	es	fr	he	it	ja	ko	nl	pt	ro	ru	tr	zh
BLEU	no-context	17.25	28.02	35.72	37.74	32.70	32.30	7.10	6.80	32.22	39.03	25.36	17.00	12.32	15.96
	context	16.92	28.24	36.00	37.23	32.92	32.11	4.48	3.77	32.67	39.10	25.37	17.14	11.97	15.01
	context-gold	18.61	28.60	36.27	37.96	33.41	32.37	5.96	6.92	32.73	39.55	28.49	17.70	12.49	16.05
COMET	no-context	0.0002	0.1841	0.3809	0.3087	0.0948	0.2608	-0.5366	-0.0275	0.3105	0.4562	0.3826	0.0033	0.2113	-0.1419
	context	-0.0066	0.1846	0.3875	0.2811	0.0887	0.2496	-0.7728	-0.3339	0.3238	0.4444	0.3747	-0.0190	0.1831	-0.1917
	context-gold	0.0025	0.1886	0.3879	0.2821	0.0922	0.2467	-0.6827	-0.1000	0.3218	0.4506	0.3805	-0.0173	0.1871	-0.1274
ellipsis	no-context	0.374	0.387	0.210	0.400	0.439	0.259	0.123	0.169	0.400	0.342	0.333	0.255	0.165	0.145
	context	0.325	0.323	0.333	0.406	0.389	0.400	0.021	0.033	0.471	0.450	0.270	0.292	0.240	0.135
	context-gold	0.388	0.296	0.300	0.435	0.371	0.381	0.025	0.150	0.444	0.450	0.306	0.226	0.187	0.154
formality	no-context	–	0.607	0.370	0.792	–	0.429	0.443	0.399	0.682	0.599	0.434	0.464	0.097	0.691
	context	–	0.639	0.351	0.791	–	0.462	0.414	0.397	0.694	0.600	0.405	0.469	0.083	0.695
	context-gold	–	0.661	0.443	0.803	–	0.464	0.431	0.425	0.697	0.622	0.440	0.492	0.182	0.741
lexical	no-context	0.639	0.762	0.819	0.826	0.723	0.766	0.615	0.574	0.821	0.853	0.661	0.624	0.671	0.645
	context	0.630	0.736	0.833	0.830	0.722	0.772	0.572	0.524	0.825	0.851	0.689	0.624	0.647	0.644
	context-gold	0.675	0.737	0.832	0.832	0.727	0.773	0.614	0.593	0.828	0.857	0.713	0.625	0.647	0.676
pronouns	no-context	0.660	0.613	0.576	0.774	–	0.548	0.473	–	–	0.452	0.356	–	–	–
	context	0.691	0.614	0.538	0.771	–	0.549	0.377	–	–	0.451	0.414	–	–	–
	context-gold	0.700	0.624	0.550	0.788	–	0.530	0.428	–	–	0.485	0.432	–	–	–
verb tense	no-context	–	–	0.263	0.435	0.227	0.308	–	–	0.477	–	0.292	0.215	0.128	–
	context	–	–	0.287	0.442	0.229	0.282	–	–	0.479	–	0.292	0.215	0.094	–
	context-gold	–	–	0.272	0.435	0.229	0.285	–	–	0.487	–	0.328	0.238	0.120	–

Table 5: BLEU, COMET, and Word f-measure per tag for `base` context-aware models. BLEU, COMET and word f-measures statistically significantly higher than no-context ($p < 0.05$) are underlined.

		de	fr	ja	zh
BLEU	no-context	36.09	45.64	15.55	22.15
	context	35.86	45.40	12.68	22.68
	context-gold	36.69	46.60	16.60	22.98
COMET	no-context	0.5256	0.6332	0.0602	0.1160
	context	0.5337	0.6425	0.0753	0.2705
	context-gold	0.5427	0.6529	0.1808	0.2809
ellipsis	no-context	0.429	0.462	0.126	0.254
	context	0.518	0.393	0.068	0.230
	context-gold	0.444	0.444	0.144	0.209
formality	no-context	0.642	0.824	0.510	0.747
	context	0.640	0.810	0.513	0.739
	context-gold	0.692	0.820	0.537	0.739
lexical	no-context	0.773	0.864	0.704	0.661
	context	0.776	0.868	0.699	0.671
	context-gold	0.796	0.875	0.740	0.696
pronouns	no-context	0.633	0.790	0.493	–
	context	0.635	0.795	0.541	–
	context-gold	0.665	0.801	0.536	–
verb tense	no-context	–	0.526	–	–
	context	–	0.532	–	–
	context-gold	–	0.534	–	–

Table 6: Word f-measure per tag for `large` models. BLEU, COMET, word f-measures statistically significantly higher than no-context ($p < 0.05$) are underlined.

also evaluate a *sentence-level* version of DeepL where we feed each sentence separately to compare with its document-level counterpart.

5.3 Results and Discussion

Table 5 shows the results for `base` models, trained either without context (`no-context`) or with context, and for the latter with either *predicted* context (`context`) or *reference* context (`context-gold`) during decoding. Results are reported with respect to standard MT metrics such as BLEU (Papineni et al., 2002) and COMET (Rei et al., 2020), as well as the MuDA benchmark.

First, we find that BLEU are highest for `context-gold` models for most language pairs, but context-agnostic models have higher COMET scores. Moreover, in terms of mean word f-measure overall, we do not find significant differences between the three systems. It is therefore difficult to see which system performs the best on document-level ambiguities using only corpus-level metrics.

For words tagged by MuDA as requiring context for translation, context-aware models often achieve higher word f-measure than context-agnostic models on certain tags such as *ellipsis* and *formality*, but on other tags such as *lexical* and *verb form*, they do not significantly outperform the context-agnostic models. This demonstrates how MuDA allows us to identify what kind of inter-sentential ambiguities context-aware models are able to resolve or not.

For the pretrained `large` models (Table 6), context-aware models perform better than the context-agnostic on corpus-level metrics, especially COMET. On words tagged with MuDA, context-aware models generally obtain the highest f-measure as well, particularly when given reference context, especially on phenomena such as *lexical* and *pronouns*, but the improvements are less pronounced than on corpus-level evaluation.

Among commercial engines (Table 7), DeepL seems to outperform Google on most metrics and language pairs. Also, the sentence-level ablation of DeepL performs worse than its document-level system for most MuDA tags, which further suggests DeepL is able to process context to some extent.

Overall, current context-aware MT systems seem to translate some inter-sentential discourse phenomena well, but they are still unable to consistently ob-

		ar	de	es	fr	he	it	ja	ko	nl	pt	ro	ru	tr	zh
BLEU	Google	11.73	34.76	43.47	30.77	10.77	31.34	12.98	8.77	38.51	38.49	28.54	24.79	18.22	28.92
	DeepL (sent)	x	34.29	42.00	42.57	x	35.41	14.88	x	37.58	37.37	28.98	25.67	x	27.94
	DeepL (doc)	x	36.75	43.06	43.43	x	36.04	15.66	x	38.29	37.76	29.79	26.53	x	27.34
COMET	Google	0.3862	0.5480	0.7694	0.6655	0.3666	0.6707	0.2116	0.4721	0.6401	0.7925	0.7437	0.5121	0.7254	0.3697
	DeepL (sent)	x	0.5750	0.7680	0.7121	x	0.6951	0.2973	x	0.6321	0.7513	0.8026	0.5501	x	0.3739
	DeepL (doc)	x	0.5848	0.7882	0.7267	x	0.7049	0.2343	x	0.6357	0.7572	0.8121	0.5495	x	0.2453
ellipsis	Google	0.343	0.667	0.500	0.306	0.359	0.468	0.279	0.352	0.389	0.632	0.405	0.367	0.236	0.323
	DeepL (sent)	x	0.417	0.400	0.422	x	0.500	0.275	x	0.500	0.421	0.458	0.385	x	0.303
	DeepL (doc)	x	0.435	0.526	0.493	x	0.553	0.208	x	0.500	0.359	0.532	0.385	x	0.295
formality	Google	–	0.621	0.404	0.738	–	0.458	0.489	0.300	0.638	0.633	0.479	0.512	0.367	0.599
	DeepL (sent)	–	0.641	0.419	0.733	–	0.455	0.487	x	0.610	0.625	0.533	0.533	x	0.729
	DeepL (doc)	–	0.670	0.446	0.785	–	0.503	0.520	x	0.641	0.614	0.526	0.534	x	0.664
lexical	Google	0.665	0.786	0.854	0.827	0.697	0.794	0.602	0.611	0.825	0.860	0.700	0.635	0.677	0.693
	DeepL (sent)	x	0.773	0.840	0.860	x	0.805	0.657	x	0.799	0.848	0.714	0.653	x	0.660
	DeepL (doc)	x	0.776	0.841	0.872	x	0.812	0.640	x	0.802	0.846	0.713	0.649	x	0.657
pronouns	Google	0.670	0.648	0.626	0.757	–	0.511	0.486	–	–	0.488	0.326	–	–	–
	DeepL (sent)	x	0.608	0.538	0.737	–	0.543	0.526	–	–	0.483	0.394	–	–	–
	DeepL (doc)	x	0.706	0.588	0.789	–	0.551	0.557	–	–	0.513	0.472	–	–	–
verb tense	Google	–	–	0.415	0.529	0.311	0.450	–	–	0.554	–	0.358	0.314	0.167	–
	DeepL (sent)	–	–	0.390	0.553	x	0.478	–	–	0.562	–	0.400	0.327	x	–
	DeepL (doc)	–	–	0.426	0.562	x	0.445	–	–	0.567	–	0.411	0.349	x	–

Table 7: Scores for commercial models. DeepL (doc) BLEU, COMET and word f-measures statistically significantly higher than DeepL (sent) are underlined.

tain considerable improvements over their context-agnostic counterparts on challenging MuDA data.

6 Related Work

To target evaluation on discourse phenomena, several works resort to measuring the performance of context-aware models targeted to discourse phenomena that require context.

The first example of discourse phenomena evaluations was done by Hardmeier et al. (2010), which evaluated automatically the precision and recall of pronoun translation in statistical MT systems. Jwalapuram et al. (2019) proposed evaluating models on pronoun translation based on a pairwise comparison between translations that were generated with and without context, and later Jwalapuram et al. (2020) extended this work to include more languages and phenomena in their automatic evaluation/test set creation. While these works rely on prior domain knowledge and intuitions to identify context-aware phenomena, we instead take a systematic, data-driven approach and find additional phenomena in doing so.

Most works have focused on evaluating performance in discourse phenomena through the use of *contrastive datasets* instead. Müller et al. (2018) automatically create a dataset for anaphoric pronoun resolution to evaluate MT models in EN → DE. Bawden et al. (2018) manually creates a dataset for both pronoun resolution and lexical choice in EN → FR. Voita et al. (2018, 2019b) creates a dataset for anaphora resolution, deixis, ellipsis and lexical cohesion in EN → RU. However, Yin et al. (2021) suggest that the task of *translating* and *disambiguating* between two contrastive choices are inherently different, which motivates our approach in measuring direct translation performance through evaluation of word f-measure.

7 Conclusions and Future Work

In this work, we investigate the types of ambiguous translations where MT models benefit from context using our proposed P-CXMI metric. Our data-driven thematic analysis helps us identify context-sensitive discourse phenomena, some of which (such as *verb forms*) have not been addressed in prior works on context-aware MT, for 14 language pairs. The advantages of our approach is that it is systematic and does not require a-priori language-specific knowledge to identify these phenomena, so we believe that our methodology can be easily extended to other language pairs. P-CXMI can also be used to identify types of context-dependent words for tasks outside MT. Based on our findings, we then construct the MuDA benchmark that tags words in a given parallel corpus and evaluate models on 5 context-dependent discourse phenomena. We find that *ellipsis* is the most challenging to tag with high precision and we leave improvements to model cross-lingual ellipsis for future work.

Our evaluation using MuDA reveals that both context-aware and commercial translation systems achieve small improvements over context-agnostic models on many discourse-aware translations, and we encourage using MuDA to benchmark the development of models that address these ambiguities.

References

Rachel Bawden, Rico Sennrich, Alexandra Birch, and Barry Haddow. 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.

Ann Bies, Mark Ferguson, Karen Katz, Robert MacIntyre, Victoria Tredinnick, Grace Kim, Mary Ann Marcinkiewicz, and Britta Schasberger. 1995. Bracketing guidelines for treebank ii style penn treebank project. *University of Pennsylvania*, 97:100.

Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative research in psychology*, 3(2):77–101.

Emanuele Bugliarello, Sabrina J. Mielke, Antonios Anastasopoulos, Ryan Cotterell, and Naoaki Okazaki. 2020. It's easier to translate out of English than into it: Measuring neural translation difficulty by cross-mutual information. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1640–1649, Online. Association for Computational Linguistics.

Marine Carpuat. 2009. One translation per discourse. In *Proceedings of the Workshop on Semantic Evaluations: Recent Achievements and Future Directions (SEW-2009)*, pages 19–27, Boulder, Colorado. Association for Computational Linguistics.

Kenneth Ward Church and Patrick Hanks. 1990. Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1):22–29.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Zi-Yi Dou and Graham Neubig. 2021. Word alignment by fine-tuning embeddings on parallel corpora. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2112–2128, Online. Association for Computational Linguistics.

Miquel Esplà, Mikel Forcada, Gema Ramírez-Sánchez, and Hieu Hoang. 2019. ParaCrawl: Web-scale parallel corpora for the languages of the EU. In *Proceedings of Machine Translation Summit XVII Volume 2: Translator, Project and User Tracks*, pages 118–119, Dublin, Ireland. European Association for Machine Translation.

Patrick Fernandes, Kayo Yin, Graham Neubig, and André F. T. Martins. 2021. Measuring and increasing context usage in context-aware machine translation. In *Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL-IJCNLP)*, Virtual.

Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson F. Liu, Matthew Peters, Michael Schmitz, and Luke S. Zettlemoyer. 2017. Allennlp: A deep semantic natural language processing platform.

Liane Guillou, Christian Hardmeier, Ekaterina Lapshinova-Koltunski, and Sharid Loáiciga. 2018. A pronoun test suite evaluation of the English–German MT systems at WMT 2018. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 570–577, Belgium, Brussels. Association for Computational Linguistics.

Christian Hardmeier, Marcello Fondazione, and Bruno Kessler. 2010. Modelling pronominal anaphora in statistical machine translation.

Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.

Prathyusha Jwalapuram, Shafiq Joty, Irina Temnikova, and Preslav Nakov. 2019. Evaluating pronominal anaphora in machine translation: An evaluation measure and a test suite. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2964–2975, Hong Kong, China. Association for Computational Linguistics.

Prathyusha Jwalapuram, Barbara Rychalska, Shafiq R. Joty, and Dominika Basaj. 2020. Can your context-aware MT system pass the dip benchmark tests? : Evaluation benchmarks for discourse phenomena in machine translation. *CoRR*, abs/2004.14607.

Yunsu Kim, Duc Thanh Tran, and Hermann Ney. 2019. When and why is document-level context useful in neural machine translation? In *Proceedings of the Fourth Workshop on Discourse in Machine Translation (DiscoMT 2019)*, pages 24–34, Hong Kong, China. Association for Computational Linguistics.

Samuel Läubli, Rico Sennrich, and Martin Volk. 2018. Has machine translation achieved human parity? a case for document-level evaluation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4791–4796, Brussels, Belgium. Association for Computational Linguistics.

António Lopes, M. Amin Farajian, Rachel Bawden, Michael Zhang, and André F. T. Martins. 2020. Document-level neural MT: A systematic comparison. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 225–234, Lisboa, Portugal. European Association for Machine Translation.

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.

Sameen Maruf and Gholamreza Haffari. 2018. Document context neural machine translation with memory networks. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1275–1284, Melbourne, Australia. Association for Computational Linguistics.

Sameen Maruf, Fahimeh Saleh, and Gholamreza Haffari. 2021. A survey on document-level neural machine translation: Methods and evaluation. *ACM Comput. Surv.*, 54(2).

Lesly Miculicich, Dhananjay Ram, Nikolaos Pappas, and James Henderson. 2018. Document-level neural machine translation with hierarchical attention networks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2947–2954, Brussels, Belgium. Association for Computational Linguistics.

Makoto Morishita, Jun Suzuki, and Masaaki Nagata. 2020. JParaCrawl: A large scale web-based English-Japanese parallel corpus. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 3603–3609, Marseille, France. European Language Resources Association.

Mathias Müller, Annette Rios, Elena Voita, and Rico Sennrich. 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.

Graham Neubig, Zi-Yi Dou, Junjie Hu, Paul Michel, Danish Pruthi, and Xinyi Wang. 2019. compare-mt: A tool for holistic comparison of language generation systems. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 35–41, Minneapolis, Minnesota. Association for Computational Linguistics.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53, Minneapolis, Minnesota. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 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.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108, Online. Association for Computational Linguistics.

Ye Qi, Devendra Sachan, Matthieu Felix, Sarguna Padmanabhan, and Graham Neubig. 2018. When and why are pre-trained word embeddings useful for neural machine translation? In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 529–535, New Orleans, Louisiana. Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Jörg Tiedemann and Yves Scherrer. 2017. Neural machine translation with extended context. In *Proceedings of the Third Workshop on Discourse in Machine Translation*, pages 82–92, Copenhagen, Denmark. Association for Computational Linguistics.

Antonio Toral, Sheila Castilho, Ke Hu, and Andy Way. 2018. Attaining the unattainable? reassessing claims of human parity in neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 113–123, Brussels, Belgium. Association for Computational Linguistics.

Jannis Vamvas and Rico Sennrich. 2021. On the limits of minimal pairs in contrastive evaluation. In *Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 58–68, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Elena Voita, Rico Sennrich, and Ivan Titov. 2019a. Context-aware monolingual repair for neural machine translation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 877–886, Hong Kong, China. Association for Computational Linguistics.

Elena Voita, Rico Sennrich, and Ivan Titov. 2019b. When a good translation is wrong in context: Context-aware machine translation improves on deixis, ellipsis, and lexical cohesion. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1198–1212, Florence, Italy. Association for Computational Linguistics.

Elena Voita, Pavel Serdyukov, Rico Sennrich, and Ivan Titov. 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.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Kayo Yin, Patrick Fernandes, Danish Pruthi, Aditi Chaudhary, André F. T. Martins, and Graham Neubig. 2021. Do context-aware translation models pay the right attention? In *Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL-IJCNLP)*, Virtual.

Language	Family	Word Order	Pronouns Politeness	Gendered Pronouns	Gender Assignment
Arabic	Afro-Asiatic	VSO	None	1 and/or 2 and 3	Semantic-Formal
English	Indo-European	SVO	None	3.Sing	Semantic
German	Indo-European	SOV/SVO	Binary	3.Sing	Semantic-Formal
Spanish	Indo-European	SVO	Binary	1 and/or 2 and 3	Semantic-Formal
French	Indo-European	SVO	Binary	3.Sing	Semantic-Formal
Hebrew	Afro-Asiatic	SVO	None	1 and/or 2 and 3	Semantic-Formal
Italian	Indo-European	SVO	Binary	3.Sing	Semantic-Formal
Japanese	Japonic	SOV	Avoided	3	None
Korean	Koreanic	SOV	Avoided	3.Sing	None
Dutch	Indo-European	SOV/SVO	Binary	3.Sing	Semantic-Formal
Portuguese	Indo-European	SVO	Binary	3.Sing	Semantic-Formal
Romanian	Indo-European	SVO	Multiple	3.Sing	Semantic-Formal
Russian	Indo-European	SVO	Binary	3.Sing	Semantic-Formal
Turkish	Turkic	SOV	Binary	None	None
Mandarin	Sino-Tibetan	SVO	Binary	3.Sing	None

Table 8: Properties of the languages in our study.

A Language Properties

Table 8 summarizes the properties of the languages analyzed in this work.

B P-CXMI Results

Table 9 presents the average P-CXMI value per POS tag and per MuDA tag.

C Tagger Details

C.1 Formality Words

Table 10 gives the list of words related to formality for each target language.

C.2 Ambiguous Pronouns

Table 11 provides English pronouns and the list of possible target pronouns.

C.3 Ambiguous Verbs

Table 12 lists verb forms that may require disambiguation during translation.

C.4 Ellipsis Classifier

We train a BERT text classification model (Devlin et al., 2019) on data from the Penn Treebank, where we labeled each sentence containing the tag '*?*' as containing ellipsis (Bies et al., 1995). We obtain 248,596 sentences total, with 2,863 tagged as ellipsis. Then, our model using HuggingFace Transformers (Wolf et al., 2020). To address the imbalance in labels, we up-weight the loss for samples tagged as ellipsis by a factor of 100.

D Training details

The *transformer-small* model has hidden size of 512, feedforward size of 1024, 6 layersa and 8 attention heads. The *transformer-large* model has hidden size of 1024, feedforward size of 4096, 6 layers, 16 attention heads.

As in Vaswani et al. (2017), we train using the Adam optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.98$ and use an inverse square root learning rate scheduler, with an initial value of 10^{-4} for `large` model and 5×10^{-4} for the `base` and `multi` models, with a linear warm-up in the first 4000 steps.

For the pretrained models we used Paracrawl (Esplà et al., 2019) for German and French, JParacrawl (Morishita et al., 2020) for Japanese and the Backtranslated News from WMT2021 for Chinese.

	ar	de	es	fr	he	it	ja	ko	nl	pt	ro	ru	tr	zh
CXMI	0.073	0.008	0.011	0.011	0.021	0.015	0.067	0.035	0.005	0.009	0.051	0.015	0.016	0.081
P-CXMI	0.075	0.005	0.011	0.021	0.023	0.016	0.059	0.038	0.002	0.013	0.049	0.015	0.014	0.057
ADJ	0.017	-0.014	-0.011	0.000	-0.037	-0.008	0.001	-0.002	-0.006	-0.005	0.020	0.015	-0.006	0.007
ADP	0.017	-0.001	-0.004	-0.004	-0.006	-0.005	0.005	0.014	-0.005	-0.001	0.011	-0.003	-0.005	-0.001
ADV	0.038	-0.011	0.008	0.002	0.007	0.005	0.005	-0.006	0.001	0.011	0.062	0.023	-0.013	0.009
AUX	0.053	0.010	0.002	0.010	0.008	0.036	0.012	0.032	0.010	0.010	0.048	0.045	0.055	0.007
CCONJ	0.044	0.025	0.024	0.005	0.012	0.043	0.034	-0.020	0.010	0.009	0.165	0.042	-0.007	-0.023
DET	0.006	0.004	0.006	0.002	-0.004	0.001	0.011	0.043	-0.007	0.002	0.046	0.018	0.011	0.008
INTJ	-0.066		-0.024	0.013	0.010	-0.015	-0.087	0.004	0.037	-0.019	0.031	-0.041	-0.009	
NOUN	0.012	-0.010	0.000	0.010	-0.001	0.000	-0.008	0.003	-0.011	-0.003	0.044	-0.010	-0.006	-0.002
NUM	0.011	-0.005	-0.005	-0.008	0.002	0.017	0.019	-0.046	-0.002	0.009	0.008	0.025	-0.000	0.004
PART	0.025	-0.007	0.029	0.063		-0.718	0.006				0.018	0.016		-0.006
PRON	0.019	0.014	-0.002	0.021	0.039	0.003	-0.009	0.047	0.006	0.013	0.029	0.023	0.000	0.023
PRON.1	0.015	0.011	0.009	0.015	0.043	0.021			0.008	0.015	0.046	0.015	-0.012	0.025
PRON.1.Plur	0.027	0.007	-0.002	0.008	0.082	0.004				0.045	0.012	0.013	-0.022	0.033
PRON.1.Sing	-0.036	0.014	0.017	0.020	0.016	0.037				0.001	0.075	0.015	-0.006	
PRON.2	0.040	0.222	-0.020	0.037	0.108	0.015			0.013	0.171	-0.017	0.103	-0.026	0.009
PRON.2.Plur	0.075	-0.055	-0.019	-0.008	0.088	0.011					-0.008	0.069	-0.024	
PRON.2.Sing	0.009	0.226	-0.021	0.357	0.125	0.052					-0.033	0.412	-0.038	
PRON.3	0.018	0.026	-0.009	0.024	0.031	-0.020			0.004	0.033	0.029	0.042	0.008	0.045
PRON.3.Dual	0.057													
PRON.3.Plur	0.016	0.017	-0.021	0.037	0.050	0.024				0.058	0.062	0.038	0.047	0.038
PRON.3.Sing	0.017	0.032	0.000	0.030	0.026	0.009				0.014	0.046	0.044	-0.001	
PRON.Plur		0.001	0.018	0.096		0.021				0.003		-0.027		
PRON.Sing		0.002	-0.005	0.025	-0.004	0.005				0.002		0.007		
PROPN	0.016	-0.014	-0.002	0.018	0.017	-0.016	-0.018	0.003	-0.005	-0.013	0.007	0.021	-0.014	0.005
PUNCT	0.129	0.007	0.012	0.001	0.019	0.019	0.353	0.017	0.018	0.021	0.005	0.017	0.022	0.106
SCONJ	0.137	-0.001	0.017	0.001	0.007	-0.000	0.004	0.005	0.005	0.003	0.044	-0.001		
SYM	0.050	0.081	0.136	0.152		0.017	-0.034	-0.014	-0.010	-0.071		-0.040		0.015
VERB	0.042	0.006	0.004	0.003	0.007	0.004	0.008	0.036	0.002	0.005	0.047	0.015	0.014	0.015
VERB.Fut			0.043	0.004	0.019	0.008				-0.001		-0.018	0.007	
VERB.Imp			0.039	0.010		0.057				0.029	0.069			
VERB.Past		0.041	0.011	0.009	0.008	0.007			-0.001	0.005	-0.009	0.064	0.010	
VERB.Pres		0.013	0.001	-0.001		-0.006			0.011	0.014	0.039	0.002	0.016	
ellipsis	0.052	-0.053	-0.111	0.055	0.071	0.019	0.020	0.022	0.037	-0.070	0.111	-0.020	-0.041	0.082
formality		0.038	0.077	0.040		0.048	0.036	0.022	0.014	0.008	0.008	0.107	-0.073	0.012
lexical	-0.006	0.003	0.011	-0.001	0.003	0.001	-0.007	-0.008	-0.004	0.002	0.034	-0.002	0.008	0.004
no tag	0.041	0.001	0.003	0.005	0.005	0.006	0.011	0.013	0.002	0.005	0.036	0.009	0.003	0.017
pronouns	0.028	0.068	-0.002	0.055		0.006	-0.027			0.055	0.008			
verb form			0.042	0.009	0.009	0.041			-0.002		0.046	0.065	0.013	
with tag	-0.001	0.024	0.018	0.021	0.005	0.013	0.023	0.005	0.001	0.010	0.034	0.056	0.002	0.009

Table 9: P-CXMI for all POS tags and our ambiguity tags. In the top two rows, CXMI is the average of P-CXMI for each sentence across the corpus, and P-CXMI is the average of P-CXMI over all tokens in the corpus. Per-tag values are the average of P-CXMI for each token with the tag. The 3 highest P-CXMI scores are highlighted in varying intensities of green.

Due to the sheer number of experiments, we use a single seed per experiment.
We base our experiments on the framework *Fairseq* (Ott et al., 2019).

de	du sie
es	tú, tu, tus, ti, contigo, tuyo, te, tuya usted, vosotros, vuestro, vuestra, vuestras, os
fr	tu, ton,ta, tes, toi, te, tien, tiens, tienne, tiennes vous, votre, vos
it	tu, tuo, tua, tuoi lei, suo, sua, suoi
ja	だ, だっ, じゃ, だろう, だ, だけど, だっ ござい, ます, いらっしゃれ, いらっしゃい, ご覧, 伺い, 伺っ, 存知, です, まし
ko	제가, 저희, 나 댁에, 성함, 분, 생신, 식사, 연세, 병환, 약주, 자제분, 뵙다, 저
nl	jij, jouw, jou, jullie, je u, men, uw
pt	tu, tua, teu, teus, tuas, te você, sua, seu, seus, suas, lhe
ro	tu, el, ea, voi, ei, ele, tău, ta, tale, tine dumneavoastră, dumneata, mata,matale,dânsul, dânsa dumnealui,dumneaei, dumnealor
ru	ты, тебя, тебе, тобой, твой, твоя, твои,тебе вы, вас, вам, вами, ваш, ваши
tr	sen, senin siz, sizin
zh	你 您

Table 10: Words related to formality for each target language.

ar	you	انت، انتَ، انتِ، انتى، أنتم ، أنتن، انتو، أنتما، أنتما
	it	هو، هي
	they, them	هم، هن، هما
de	it	er, sie, es
es	it	él, ella
	they, them	ellos, ellas
	this	ésta, éste, esto
	that	esa, ese
	these	estos, estas
	those	aquellos, aquellas, ésos, ésas
fr	it	il, elle, lui
	they, them	ils, elles
	we	nous, on
	this	celle, ceci
	that	celle, celui
	these, those	celles, ceux
it	it	esso, essa
	them	ellos, ellas
	this	questa, questo
	that	quella, quello
	these	queste, questi
	those	quelle, quelli
ja	I	私, 僕, 俺
pt	it	ele, ela, o, a
	them	eles, elas, os, as
	they	eles, elas
	this, that	este, esta, esse, essa
	these, those	estes, estas, esses, essas
ro	it	el, ea
	they, them	ei, ele

Table 11: Ambiguous pronouns w.r.t. English for each target language.

es	Imperfect, Pluperfect, Future
fr	Imperfect, Past, Pluperfect
he	Imperfect, Future, Pluperfect
it	Imperfect, Pluperfect, Future
nl	Past
pt	Pluperfect
ro	Imperfect, Past, Future
ru	Past
tr	Pluperfect

Table 12: Ambiguous verb forms w.r.t. English for each target language.