Identifying Context-Dependent Translations for Evaluation Set Production

Rachel Wicks[1,2] and **Matt Post**[1–3]
[1]Human Language Technology Center of Excellence, Johns Hopkins University
[2]Center of Language and Speech Processing, Johns Hopkins University
[3]Microsoft
rewicks@jhu.edu, mattpost@microsoft.com

Abstract

A major impediment to the transition to context-aware machine translation is the absence of good evaluation metrics and test sets. Sentences that require context to be translated correctly are rare in test sets, reducing the utility of standard corpus-level metrics such as COMET or BLEU. On the other hand, datasets that annotate such sentences are also rare, small in scale, and available for only a few languages. To address this, we modernize, generalize, and extend previous annotation pipelines to produce CTXPRO, a tool that identifies subsets of parallel documents containing sentences that require context to correctly translate five phenomena: gender, formality, and animacy for pronouns, verb phrase ellipsis, and ambiguous noun inflections. The input to the pipeline is a set of handcrafted, per-language, linguistically-informed rules that select contextual sentence pairs using coreference, part-of-speech, and morphological features provided by state-of-the-art tools. We apply this pipeline to seven languages pairs (EN into and out-of DE, ES, FR, IT, PL, PT, and RU) and two datasets (OpenSubtitles and WMT test sets), and validate its performance using both overlap with previous work and its ability to discriminate a contextual MT system from a sentence-based one. We release the CTXPRO pipeline and data as open source.[1]

1 Introduction

Neural machine translation (NMT) systems can produce high-quality, fluent output which are nearly indistinguishable from human translations, when evaluated at the sentence level. This human-level parity has been shown to disappear, however, when evaluated in context (Läubli et al., 2018; Toral et al., 2018). This is unsurprising, because sentences are nearly always written by humans in some contextual setting, and are translated by translators in the same fashion. Dismissing this context may create ambiguities that do not exist in the document as a whole, and in some cases, may make it impossible to correctly interpret the sentence.

	GENDER	AUXILIARY	INFLECTION	FORMALITY	ANIMACY	LANGS
Müller et al.	✓					de
Lopes et al.	✓					fr
Voita et al.	✓	✓	✓	✓		ru
Nadejde et al.				*		de, es, fr, hi, it, ja
Currey et al.	†					ar, fr, de, hi, it, pt, ru, es
This work	✓	✓	✓	✓	✓	de, fr, ru, pl, pt, it, es

Table 1: This work expands evaluation set coverage to new document phenomena and languages. (*) Note that Nadejde et al. (2022) does not include contextual information. (†) Currey et al. (2022) focuses on natural, rather than grammatical, gender.

Translation to another language must address ambiguities where the semantic or grammatical granularity of two sentences is imbalanced or mismatched. Probably the most widely-known of these is grammatical gender, i.e., when translating referential pronouns from a grammatically non-gendered language to a gendered one. For example, when translating from English to French, the pronoun *it* must be translated to *il* or *elle* depending on the grammatical gender of the antecedent noun, which may not be available in the same sentence.

The obvious path forward in addressing these issues is to move to contextual machine translation, in which sentences are no longer translated in isolation but with their source-side context. Recent work has shown that transformers (Vaswani et al., 2017) are capable of handling longer sequences and improving performance on context-based evaluation (Sun et al., 2022; Post and Junczys-Dowmunt, 2023). However, general contextual translation has

[1]https://github.com/rewicks/ctxpro

Proceedings of the Eighth Conference on Machine Translation (WMT), pages 452–467
December 6–7, 2023.

	English	Target
AUXILIARY	I just figured you need to know. *And now you **do.***	(fr) Je pensais que tu méritais de savoir. *Et maintenant tu **sais.***
	I can't lose my voice. *You **won't.***	(pl) Nie mogę stracić głosu. *Nie **stracisz.***
INFLECTION	Mostly work with the Knicks right now. *And other **athletes.***	(ru) В основном работаю с "Никс". И с другими спортсменами.
GENDER	You think migraines are a sign of weakness, don't want anyone to know. *I used to get **them**, too.*	(it) Lei pensa che le emicranie siano segno di debolezza, e non vuole che si sappia. ***Le** prendevo anch'io.*
	This pain? *I long for **it.***	(pt) A dor? *Anseio por **ela.***
ANIMACY	Et il y a eu cette rose aussi pour toi. *Tu sais, **elle** se distingue des autres.*	(en) Also, uh, this rose came for you. *You know, **it** stands out in front of all the others.*
	La felicidad es un mito. *Y vale la pena luchar por **ella.***	(en) Happiness is a myth. *And **it's** worth fighting for.*
FORMALITY	*We'll call **you** if something happens, huh?*	(de) *Wir rufen **euch** an, wenn etwas passiert.*
	*Well, uh, I was an obstetrician before, and I most definitely owe **you.***	(es) *Bueno, era obstetra antes, y definitivamente se **los** debo.*

Table 2: An example of the extracted ambiguities with their preceding contexts for each language pair. The ambiguous sentence is denoted in *italics* and the ambiguous word is **bolded**. Note the dialectal use of the "usted" accusative form "los". Language denoted in parentheses.

a number of obstacles, foremost is the lack of available evaluation resources. There are essentially two kinds of contextual evaluations: general metrics, which can theoretically be applied to any test set, and fixed test sets. There is relatively little work in the former setting (Vernikos et al., 2022; Jiang et al., 2022), and while they correlate with human judgments, they have not been proven capable of discriminating sentence-based from known-high-quality contextual systems. For the latter, a number of high-quality evaluation sets exist (Müller et al., 2018; Lopes et al., 2020; Bawden et al., 2018; Voita et al., 2019, Table 1), but they are limited both in language coverage and scope of phenomena.

In this work, we address this lack of evaluation data by extending coverage of existing datasets to more languages and contextual phenomena. We:

- develop a pipeline that makes use of broad-language-coverage annotation tools and hand-developed rules to identify context-based phenomena in any test set;
- construct rules for five context-based phenomena (§ 2) and seven language pairs (§ 3): DE, ES, FR, IT, PL, PT, and RU with EN; and
- apply this toolchain to multiple datasets.

We show that this dataset, called CTXPRO, is capable of discriminating high-quality contextual systems from sentence-level ones.

2 Contextual phenomena

A number of context-based phenomena which create ambiguities are common. We display some examples in Table 2. Humans easily handle these ambiguities during translation, which nearly always takes place in context, so a machine translation system which ignores these issues will never reach human-level parity. Some, such as lexical cohesion or fluency, are hard to quantify, while others, for example pronoun translation accuracy or word sense disambiguation, are easier. These phenomena all present difficulties and even impossibilities to systems that translate sentences in isolation. Our goal is to identify as many of these phenomena we can in a general way, such that we can create a general pipeline for isolating them, that can be reliably applied to any test set.

We describe each phenomena for comprehension and then provide our extraction methodology in order to identify when these ambiguities arise.

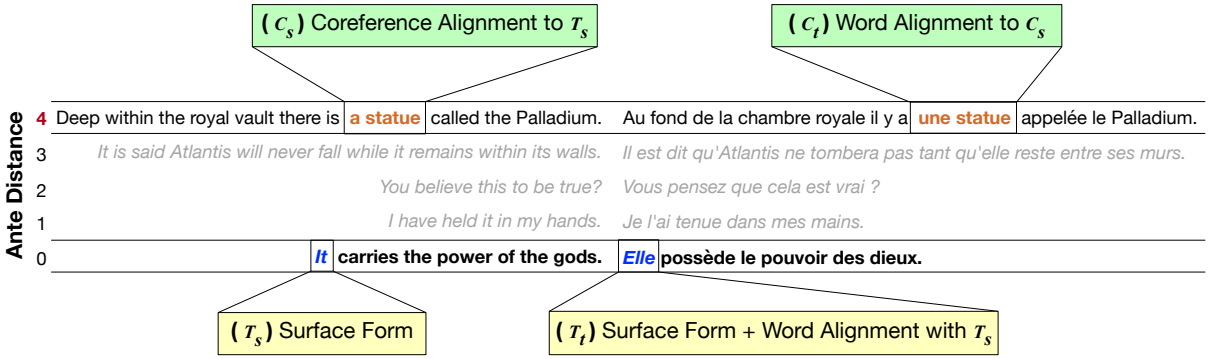

Figure 1: A diagram showing how the four key words for GENDER identification are identified. The antecedent distance is determined by what sentence C_e is found in. In order to be considered, T_e, T_t, C_e, C_t would also have to pass morphological feature tests similar to those shown in Table 3.

2.1 Anaphoric pronouns

Pronouns are a general descriptor that function as a placeholder for a noun phrase, providing the speaker with a more succinct form instead of repeatedly identifying an established referent.

In grammatical contexts, anaphora refers to the use of a pronoun to refer to a previously mentioned word or entity. Pronouns for which the referent noun can be found in preceding contexts are called *anaphora*; in contrast, *cataphora* denotes situations where the referent noun follows the pronoun. We do not consider cataphora in this paper.

2.1.1 Gender

Languages with gendered nouns require agreement with the appropriate gendered pronoun. English, which makes no such distinction for inanimate objects, will use the pronoun "it." In order to correctly translate "it" into Spanish, it is necessary to know what "it" refers to. If "it" refers to a school, it would be translated differently (*una* escuela) than if it refers to a heart (*un* corazón).

Apart from a few exceptions, English does not make use of grammatical gender. Machine translation often centers around translating either into or out of English with most of the paired languages expressing genders (masculine, feminine, and neuter), so there is a clear need to evaluate the translation of gender. Further, removing English from the equation does not resolve the problem. Gender assignment of inanimate objects is arbitrary which means that translating between two gendered languages is non-trivial. In extreme cases, a language may exhibit *"noun classes"* which behave similarly to gender, but may correlate more heavily with meaning. A noun in Swahili is not grouped via an arbitrary *gender* assignment, but is instead somewhat assigned to groups based on other labels such as *animacy, items, plants*, or *tools*. These classes affect morphological agreement in ways that English does not express. In any case, translating a pronoun that refers to a previously mentioned noun requires resolving this coreference in order to correctly generate the new pronoun.

2.1.2 Animacy

Humans and animals are often treated differently grammatically than inanimate objects. As stated, English makes no gender distinction for inanimate objects, though it does have gendered pronouns for *animate* objects. *She* and *he* are English pronouns used for humans and often animals but are rarely used to refer to inanimate objects.[2] This results in an ambiguity when translating pronouns into English from languages that do not make this distinction. For example, in English, *she is in the kitchen* clearly refers to a person while *it is in the kitchen* refers to a non-person. In French, the word *elle* would be used in both situations, requiring an MT system to make a choice.

2.1.3 Formality

Social expectations dictate language usage. In many languages, this is explicitly lexicalized with different second-person pronouns and verb conjugations that distinguish intimate or familiar relationships from formal ones. Examples include the *tu/vous* distinction in French and *du/Sie* in German.

Over time, English has lost its formal register in pronouns (often called the T-V distinction) which other languages frequently employ. A common sentence "Where are you?" may have multiple

[2]A small exception occurs when inanimate objects are personified. A frequent example is boats, which are often referred to as *she* in English.

interpretations determined by the addressee, but subtle cues in preceding context may indicate the level of formality or familiarity of the speaker—a "sir", the domain, or profession mentioned can clarify this. When translating this sentence into French, the system must choose a register to produce either "Où êtes-vous?" or "Où es-tu?" There is often insufficient information to make the correct choice from just a single sentence.

2.2 Verb Phrase Ellipsis

Verb phrases can be dropped for emphasis, style, or brevity. The manner in which they are ellipsed will follow the rules of syntax of the specific language.

2.2.1 Isolated Auxiliaries

English auxiliaries ("do", "will", "would") can occur as standalone verbs by taking the place of a verb phrase. The question "Will you walk with me?" can be answered with a short "I will." Many target languages require translation of the original head of the verb phrase rather than the modal or auxiliary. Simply, "I will" must be translated as "I will walk" or rather "I walk" inflected in the future tense. We limit this work to the aforementioned auxiliaries as they rarely have direct translations.

2.2.2 Inflection of Verb-less Nouns

Extreme ellipsis may remove entire portions of a sentence and render it a *phrase*. English word order conveys grammatical role of nouns. When elements of the original sentence, such as the verb, are ellipsed, it may be impossible to infer the grammatical case of any remaining nouns which have no inflection. Translation into languages with case systems suffers. Voita et al. (2019) exemplifies using the phrase: "You call her your friend but have you been to her home? Her work?" To translate this phrase into Russian, it is necessary to know that "her work" has the same grammatical case as "her home" in the previous sentence.

3 Extraction Pipeline

Our pipeline functions by identifying up to four key tokens and ensuring each token matches a set of predefined criteria. The four components are: (1) The source (English) token defined as T_s, the target (non-English) token defined as T_t, the source token which conveys the contextual information required to resolve the ambiguity defined as C_s, and the target token aligned to C_s defined as C_t. These relationships are illustrated in Figure 1. Contextual information is defined by a contextual relationship, Q, which has an associated solver. The predefined criteria is a set of rules, R.

We can identify ambiguous sentences by:

1. For each source–target sentence pair, apply word alignment. Each aligned pair of words forms a potential T_s–T_t pair.
2. Ensure T_s meets all criteria R_{T_s}
3. Ensure T_t meets all criteria R_{T_t}
4. Apply a solver for the contextual relationship, Q to the English token T_s and its preceding context to identify C_s.
5. Ensure C_s meets all criteria R_{C_s}.
6. Identify the target token C_t via word alignment to C_s. If translation conveys semantic symmetry, this token *also* has a contextual relationship with T_t.
7. Ensure C_t meets all criteria R_{C_t}

Consider the ambiguity of pronoun resolution. Müller et al. (2018) first proposed a pipeline for extracting ambiguous translations of English "it" to German nominatives ("er", "es", and "sie"). We can explain their methodology[3] via the aforementioned definition. The following identifies all ambiguities where the English "it" is translated as "sie."

1. For each source-target sentence pair, apply word alignment. Each aligned pair of words forms a potential T_s–T_t pair.
2. Ensure T_s is the word "it"
3. Ensure T_t is the word "sie"
4. The contextual information to resolve the ambiguity is its antecedent—expressed via a coreference relationship. Apply a coreference resolver (Q) to identify C_s.
5. Ensure C_s is a noun (not another pronoun).
6. Identify C_t via word alignment.
7. Ensure C_t is a feminine, singular noun.

The same criteria could be enumerated for the masculine and neuter equivalents, appropriately changing gender and surface form checks.

To extract a specific phenomenon and language, a "rule" (R) must be written which specifies features that T_s, T_t, C_s, and C_t must have. These features can range from part-of-speech, lemma, gender, case, plurality or others. The manner in which

[3] Müller et al. (2018) performs an extra coreference check on the target side that we do not.

	English (T_e)			German (T_t)			Coref English (C_e)	Coref German (C_t)		
Rule	Form	POS	Case	Form	POS	Case	POS	POS	Gender	Number
NOM.FEM.SING	it	PNOUN	*	sie	PNOUN	Nom.	NOUN	NOUN	Fem.	Sing.
NOM.MASC.SING	it	PNOUN	*	er	PNOUN	Nom.	NOUN	NOUN	Masc.	Sing.
NOM.NEUT.SING	it	PNOUN	*	es	PNOUN	Nom.	NOUN	NOUN	Neut.	Sing.
ACC.FEM.SING	it	PNOUN	*	sie	PNOUN	Acc.	NOUN	NOUN	Fem.	Sing.
ACC.MASC.SING	it	PNOUN	*	ihn	PNOUN	Acc.	NOUN	NOUN	Masc.	Sing.
ACC.NEUT.SING	it	PNOUN	*	es	PNOUN	Acc.	NOUN	NOUN	Neut.	Sing.
DAT.FEM.SING	it	PNOUN	*	ihr	PNOUN	Dat.	NOUN	NOUN	Fem.	Sing.
DAT.MASC.SING	it	PNOUN	*	ihm	PNOUN	Dat.	NOUN	NOUN	Masc.	Sing.
DAT.NEUT.SING	it	PNOUN	*	ihm	PNOUN	Dat.	NOUN	NOUN	Neut.	Sing.
NOM.INFORM.SING	you	PNOUN	*	du	PNOUN	Nom.	-	-	-	-
NOM.FORM+PLUR	you	PNOUN	*	Sie	PNOUN	Nom.	-	-	-	-
NOM.INFORM.PLUR	you	PNOUN	*	ihr	PNOUN	Nom.	-	-	-	-
ACC.INFORM.SING	you	PNOUN	*	dich	PNOUN	Acc.	-	-	-	-
ACC.FORM+PLUR	you	PNOUN	*	Sie	PNOUN	Acc.	-	-	-	-
ACC.INFORM.PLUR	you	PNOUN	*	euch	PNOUN	Acc.	-	-	-	-
DAT.INFORM.SING	you	PNOUN	*	dir	PNOUN	Dat.	-	-	-	-
DAT.FORM+PLUR	you	PNOUN	*	ihnen	PNOUN	Dat.	-	-	-	-
DAT.INFORM.PLUR	you	PNOUN	*	euch	PNOUN	Dat.	-	-	-	-

Table 3: German criteria for all pronouns. We expand from Müller et al. (2018) to consider more cases (Accusative and Dative). English case is not used since the German annotations are more precise (English does not label Dative). PNOUN check in some cases is required to eliminate determiners (possessive adjectives instead of possessive pronouns)

	English (T_e)	French (T_t)
Rule	Lemma	Illegal Lemmas
DO.ELL	do	faire, aller
WOULD.ELL	would	faire, pouvoir
WILL.ELL	will	aller, faire

Table 4: French ellipsis Rules. English must have specified lemma. French alignment cannot have a lemma in the specified list.

these four components are identified creates the adaptability for each phenomena.

Gender Following previous works, we retrieve T_s and T_t based on surface form and word alignment. C_s is a noun discovered via coreference chain. If the coreference is a noun phrase, the head of the phrase is used. C_t is retrieved via word alignment. C_t must match the same morphological features present in T_t (e.g., gender and number).

Animacy As explained in Section 2.1.2, the animacy ambiguity that we consider occurs when translating from the gendered languages *into* English (whereas the gender ambiguity occurs when translating *out-of* English). To extract these examples, we use the same rules as GENDER, but we reverse the language direction for inference.

Formality The distinction of formality is the lack of a consistent or discrete C_s which informs the level of formality. Translating between English and a T-V language is always ambiguous with respect to the second person so we forgo using a contextual resolver Q to identify the appropriate context.

Auxiliary T_s is extracted from a pre-constructed list of auxiliaries—similar to those mentioned in Section 2.2.1. T_t, identified via word alignment, cannot occur in a pre-constructed list of forbidden translations. These translations are meant to prevent valid translations of auxiliaries, rather than the ambiguous ellipsed forms. For example, "to do" translated as a form of "faire" in French, is a direct translation, and is likely not representative of an ellipsed form. Contrarily, "to do" translated as a form of "savoir" in French is not a direct translation and is indicative of a previous occurrence of English "to know." C_t can be identified by finding the most recent occurrence of the same verb T_t, and C_s is retrieved from word alignment with C_t.

Inflection T_s and T_t can be of any form and any case. Any aligned noun pair (T_s and T_t) that occurs without an accompanying verb is ambiguous. C_t is identified as the most recent occurrence of *any noun* occurring in the same case as T_t. We assume the verb phrase surrounding C_t was ellipsed when generating T_t. We align C_t to find C_s.

We use FastCoref (Otmazgin et al., 2022) to perform English coreference resolution, simalign (Jalili Sabet et al., 2020) to perform cross-lingual

	DE	FR	RU	PL	PT	IT	ES
GENDER	147k	291k	113k	117k	127k	36k	96k
ANIMACY *	80k	145k	66k	39k	38k	20k	84k
FORMALITY	3.9M	5.7M	3.6M	1.7M	857k	833k	10.1M
AUXILIARY	4414	27.6k	39.1k	34.2k	30.2k	17.5k	29.6k
INFLECTION	-	-	2.6M	3.2M	-	-	-
# LINES	22.5M	41.9M	25.9M	77.2M	33.2M	35.2M	61.4M
% EXTRACTED	18%	14%	25%	6.6%	3.1%	2.5%	16.7%
%-COREFERENCE	0.7%	0.8%	0.6%	0.2%	0.5%	0.2%	0.2%

Table 5: OpenSubtitles2018 Extraction Statistics for each category. # LINES indicates the total number of lines in OpenSubtitles for the EN-XX language pair. % EXTRACTED indicates the percent of the dataset that was extracted. %-COREFERENCE indicates the classes that require a strict antecedent (GENDER and AUXILIARY). (*) ANIMACY was created by reversing a subset of the GENDER class so it is not used to calculate EXTRACTED because of the overlap.

word alignment, and SpaCy[4] to extract all other morphological features. We provide a larger list of our criteria in Appendix A.

3.1 Application to OpenSubtitles

We apply our extractor to the OpenSubtitles2018 dataset (Lison and Tiedemann, 2016) following previous work (Müller et al., 2018; Lopes et al., 2020). It comprises conversational dialog extracted from film and television subtitles. The conversational nature means plenty of context-based phenomenon occur. In Table 5, we present the total number of instances we extracted from Open Subtitles.

The fraction of the dataset that contains the phenomenon we target varies from language to language. This stems from the number of forms in each language, the number of genders, as well as translation standards. German, for instance, has very few AUXILIARY examples. We speculate this is due to German having similar auxiliary features as English so many examples were filtered out due to our "forbidden translation" criteria.

Some categories are extremely common. FORMALITY is invoked every time the second-person is used, which is frequent in conversational speech. INFLECTION also has high occurrences since there was relatively little filtering on the extracted examples. GENDER and AUXILIARY are *very rarely* extracted—less than 1% of the time in all languages. A 1% error rate is extreme when deploying at scale. Further, test sets, in nature, are small. If only 1% of the test set challenges contextual models, the results may be insignificant.

To form the dev, devtest, and test splits, we apply the following approach. For each label within a category, we ensure there are at least 100 examples. If there are fewer, we keep all examples for test. If there are more, we split the most recent years of OpenSubtitles into a 1:1:5 ratio for dev:devtest:test, limiting the test set's maximum size to 5000 examples per label. One label is roughly one surface form, but corresponds to one "rule" (a set of criteria R) or one row as shown in Table 3.

4 Quantitative Evaluation

Our goal is to show that our test sets can usefully discriminate between sentence-level and context-aware systems. An impediment to this goal is the lack of contextual machine translation models across languages for use in comparison and evaluation, and the difficulty in building them. Consequently, we turn to a commercial system, DeepL, which is alone among commercial providers in advertising contextual translation.[5] We translate with document-context by providing DeepL with context when translating, and compare to the same model translating without context at the sentence level. We show that a contextual system appropriately benefits from the additional context and gains significance performance on this test set.

Many works release their evaluation sets with the assumption of contrastive evaluation (Müller et al., 2018; Lopes et al., 2020; Voita et al., 2019), where the test is whether a model assigns a higher score to correct data than to linguistically-manipulated counterparts. This assumption ignores the fact that machine translation is a generative problem and should be evaluated as such. Recent work (Post and Junczys-Dowmunt, 2023) confronts this problem

[4] https://spacy.io/usage/models#languages

[5] https://www.deepl.com/docs-api/general/working-with-context

		Generative Accuracy (%)							COMET						
		DE	ES	FR	IT	PL	PT	RU	DE	ES	FR	IT	PL	PT	RU
GENDER	sent.	48.1	34.6	40.2	51.1	32.8	44.3	35.9	0.23	0.50	0.33	0.43	0.51	0.52	0.36
	doc.	**73.3**	**47.4**	**59.0**	**68.3**	**50.2**	**64.3**	**51.8**	**0.31**	**0.52**	**0.43**	**0.48**	**0.54**	**0.57**	**0.42**
		+25.2	+12.8	+18.8	+17.2	+17.4	+20.0	+15.9	+0.08	+0.02	+0.09	+0.05	+0.03	+0.05	+0.06
ANIMACY	sent.	61.0	84.4	68.0	81.4	57.6	64.1	55.4	0.27	0.53	0.40	0.42	0.25	0.43	0.19
	doc.	**74.1**	**87.8**	**75.2**	**86.1**	**70.5**	**79.5**	**71.6**	**0.38**	**0.58**	**0.49**	**0.46**	**0.31**	**0.55**	**0.34**
		+13.1	+3.4	+7.2	+4.7	+12.9	+15.4	+16.2	+0.11	+0.05	+0.09	+0.04	+0.06	+0.12	+0.15
FORMALITY	sent.	44.0	31.7	40.6	38.9	25.3	40.1	55.4	**0.32**	0.54	0.45	0.47	**0.51**	**0.59**	0.57
	doc.	**53.6**	**35.9**	**51.5**	**46.1**	**31.6**	**47.2**	**62.5**	**0.32**	**0.55**	**0.48**	**0.48**	**0.51**	**0.59**	**0.58**
		+9.6	+4.2	+10.9	+7.2	+6.3	+7.1	+7.1	+0.0	+0.01	+0.03	+0.01	+0.0	+0.0	+ 0.01
AUXILIARY	sent.	7.8	3.3	1.3	4.0	8.2	9.2	5.7	-0.27	-0.06	-0.34	-0.02	0.10	0.03	-0.09
	doc.	**40.0**	**52.0**	**32.2**	**40.7**	**49.9**	**53.8**	**49.0**	**0.04**	**0.54**	**0.20**	**0.38**	**0.53**	**0.60**	**0.49**
		+32.2	+48.7	+30.9	+36.7	+41.7	+44.6	+43.3	+0.31	+0.60	+0.54	+0.40	+0.43	+0.57	+0.58
INFLECTION	sent.	-	-	-	-	41.3	-	34.6	-	-	-	-	0.57	-	0.47
	doc.	-	-	-	-	**53.2**	-	**48.3**	-	-	-	-	**0.68**	-	**0.56**
		-	-	-	-	+11.9	-	+13.7	-	-	-	-	+0.11	-	+0.09

Table 6: Generative evaluation percent accuracy scores (left section) evaluation ability to produce expected form; COMET scores (right section) evaluate the translation quality of this model; *sent.* denotes that no additional context was given while *doc.* was given five consecutive sentences for context. All translations made using DeepL commercial API. ANIMACY is *into* English. All others are *out of* English

and proposes generative evaluation as an alternative, showing a wide gap between contextual and sentence-level systems that is only observed under generative evaluation. Translations are counted as correct if the *expected surface form* is present anywhere in the model's output—matching the entire word and not simply a substring. We follow this approach in our evaluation.

We validate our data by showing it (1) adequately addresses context-based phenomena and (2) is sufficiently challenging. We demonstrate the former by showing that a context-aware translation model consistently outperforms a context-less equivalent. We see the latter is true as the contextual model does not solve the problem. There is still significant context-aware work to be done.

4.1 Accuracy

We begin by translating sentences both with and without context, using at most five sentences of context. To limit API calls, we run a subsample of our produced evaluation sets. We limit each category (GENDER, ANIMACY, FORMALITY, AUXILIARY, and INFLECTION) to approximately 10k total examples, divided evenly amongst the categories labels. To extract the final sentence for scoring purposes, we apply segmentation using the ERSATZ segmenter (Wicks and Post, 2022).

The results in Table 6 clearly show that the DeepL model with additional context *far outperforms* its sentence-level equivalent.[6] Many of these evaluation examples have specific preceding context that needs to be used in order to correctly translate the ambiguity. FORMALITY is a slight exception. There is little to no guarantee that explicit cues are given to convey the nature of the relationship between the speaker and addressee, yet preceding context still benefits an average of 9 percentage points across all languages. AUXILIARY is a task of translating verbs. A random guess would equate to sampling from the distribution of verbs in a language–which results in low success rates. Translating AUXILIARY with context increases from nearly never correct to a roughly 50% accuracy rate. Translating ANIMACY has higher sentence-level baselines than some of the other categories. We attribute this to other semantic cues towards ANIMACY which are less arbitrary than something such as GENDER assignment. For instance, if a noun *talks*, it is likely animate, while a noun that *is thrown* is likely inanimate. Similarly, INFLECTION may have some sentence-internal cues. Certain nouns may have a majority class, or preceding prepositions ((“with”, “for”, “in”, etc.) may indicate case. This is similar to the intrasentential coreference found with pronouns, which makes some occurrences easier than oth-

[6] Ideally we would make the same comparison between document- and sentence-level translation with other commercial systems, but there is no way to prevent them from applying sentence-level segmentation to the document-context string.

ers. Nonetheless, additional context aids the model. In every category, the context-aware model shows consistent gains over its context-less variant.

4.2 Automatic metric

We also present COMET scores (Rei et al., 2020) in Table 6. Across all categories and language pairs, COMET shows improvement when the system leverages additional context. The consistent improvement in COMET reinforces the trends we see with the generative evaluation metric. The one exception is the FORMALITY class which has minimal differences between the sentence-level and contextual inputs. COMET rewards synonyms and we suspect formal and informal surface forms have more similar encodings in COMET models than these other grammatical forms. A surface-based metric would better capture the gains that can be seen from the accuracy scores, which is indeed what we find (Table 18 in Appendix A).

5 Qualitative Evaluation

Our extraction pipeline relies on handbuilt rules applied to the outputs of automatic tools. As a result, the process is noisy and may be susceptible to errors. The previous section showed that a contextual system does better on our test sets than its sentence-based counterpart, and there is no reason we can think of to suspect that errors would systematically benefit the contextual system. However, in the interest of completeness, we took a more qualitative look at the data. This includes a systematic manual review (§ 5.1), direct comparison with prior work (§ 5.2), and an error analysis (§ 5.3).

5.1 Manual review

Previous work in automatic test set production has not typically included a manual analysis of rule quality. To build confidence in these automatic extraction methodologies, we sampled 100 random test examples from the extracted English–French GENDER set and manually reviewed and annotated them for errors. We find that 92 of the extracted examples are correct. Three more were questionably incorrect—with correct translations and alignments—yet had atypical coreference resolutions that were difficult for our human reviewer to understand. Of the remaining five, two had a non-referential pronoun. One such example "What is it?" was used in the sense of "What's wrong?" rather than "What is that?" In the former, "it" has no valid antecedent, yet it was extracted.

We present the remaining three errors in Table 7, where they demonstrate where errors arise at each step in the pipeline. The Coreference Error is a clear mistake. "They don't want us to know what they're working on" refers to the people being talked to, and not "these guys"—who instead seemed to be criminals who broke into a company. The Alignment Error is an unfortunate combination of a bad alignment and inconsistent translation. "the discipline" is aligned to the word "espionnage." "discipline" in French is a feminine noun, while "espionnage" is masculine. The French "il" is masculine, and thus has "espionnage" as an antecedent despite the English having "the discipline." This coincidental error caused this example to still be extracted. Lastly, one of these examples seemed to have a typo in the English transcript. The word "signatures" seemed to be incorrect. We suspect the correct transcription word was "serial killers." Given the inconsistent context on the English side, we suspect the neural coreference model had difficulties resolving this.

5.2 Comparison with prior work

Since our extraction framework is largely based on that of Müller et al. (2018), we expect to have a similar quality of extracted rules (or better, since the underlying annotations tools have improved). We thus undertake a comparison to the data that they released. When applying our pipeline to the German–English OpenSubtitles data, we extract 147,211 sentences that have ambiguous pronoun usage. Müller did not report their raw extraction numbers, but their release includes 12,000 examples, balanced across gender (but not distance). We therefore focus our analysis on this subset.

Since their pipeline contained a target-side coreference check that we do not have, one might think their pipeline would be a stricter selection process, but we find the opposite to be true. Our pipeline's selection overlaps with only half of ContraPro (6,003 sentences), rejecting the other half (5,997 sentences). An analysis of this rejected portion of ContraPro turns up some explanations. ContraPro extracts three categories of German pronouns corresponding to neuter, masculine, and feminine genders. For *er* and *sie*, we rejected roughly 25% of the ContraPro examples; however, we rejected over 75% of the neuter examples from ContraPro. Upon review, we found a substantial number of

Error Type	English	French
Coreference	We got any ideas what these ***guys*** were after?	Une idée de ce que voulaient ces ***gars***?
	No, CEO is on his way down to talk to us now.	Non, le PDG arrive pour nous le dire.
	So far, everyone we've talked to hasn't really given us much.	Tous ceux à qui on a parlés ne nous ont rien appris.
	Makes sense.	C'est logique.
	They don't want us to know what they're working on here.	***Ils*** ne veulent pas qu'on sache ce qu'ils font.
Alignment	As you know the ***discipline*** of media espionage is a new one.	Comme vous le savez, l'***espionnage*** médiatique est une nouvelle discipline.
	Oh yes, ***it*** is everywhere.	***Il*** est partout.
Translation	You know more about ***signatures*** than most of ***them*** put together.	Vous en savez plur sur ces ***tueurs*** (`en: killers`) qu'***eux*** tous réunis

Table 7: In a sample of 100 extracted items, 8 errors were found. This table shows 3 of these errors made by the extraction pipeline on the French Gender set. The ***indicated*** words show the pronouns in French and English, as well as their antecedents. Some examples fit into multiple categories, but these show the most evident error type. `en:` indicates the English translation of French word.

non-referential instances. These examples include sentences such as "It was your duty.", "It would have been all right if it wasn't for you." and "It was one of those California Spanish houses" that all have either a non-specified referent or have a passive construction. The inclusion of these examples points to inaccurate coference chains, likely explained by their use of older corefence tools.

Our extraction employs strict criteria to find the head of a span during coreference and alignment. The head is used for the gender, person, and number checks included in the definition of R (§ 3). From our understanding of Müller's work, they did not include this check. Mistakes are inherent to any automatic process, and likely persist in our dataset as well. Our analysis here lends some confidence to the belief that tighter selection criteria and improved underlying tools result in better data.

5.3 Model analysis

Absent sufficient information, the translation of ambiguous words will regress to their proportions in the training data. For pronouns, this would be the neuter or masculine class; for auxiliaries, the direct translation (the "Illegal Lemma" in R).

We examine the English–German model outputs. Our evaluation sets have balanced counts across genders, so a correct model would produce a neuter pronoun roughly one-third of the time. Instead, this sentence-level model produces either "es" or "ihm" (the German neuter pronouns) closer to two-thirds of the time. This contextual model has better performance producing the neuter pronouns about 40% of the time. This problem is well-known, but other issues are not as well documented.

The auxiliary category had the worst scores, both in terms of how low the sentence-level model was performing as well as the absolute increase from adding context. The cause of these scores becomes obvious as we examine the model outputs. To generate the rules for the AUXILIARY class, we enumerated illegal lemmas that represent the most common direct translations of English modals as described in Section 3. Ideally, a model would never generate these verbs for our evaluation set unless part of a larger verb phrase construction. We find the sentence-level model generates a translation that contains a form of one of these lemmas approximately two-thirds of the time. Conversely, the contextual model generates these closer to one-third of the time.

6 Analysis of WMT test sets

As previously earlier, this pipeline is easily applied to new data and test sets. We demonstrate this by applying it to the 2019–2022 WMT newswire test sets (Barrault et al., 2019, 2020; Akhbardeh et al., 2021; Kocmi et al., 2022). In so doing, we find

	DE	RU	PL
GENDER	135	64	13
FORMALITY	540	416	4
AUXILIARY	1	0	0
INFLECTION	-	14	1
WMT # LINES	6454	7038	1000

Table 8: Counts on the number of extracted examples from WMT 2019-2022 (when available) test sets.

phenomena in a similar proportion of sentences to OpenSubtitles, but with a different distribution; there is a higher rate of GENDER but smaller of FORMALITY and AUXILIARY. In Table 8, we present the total number of examples discovered in WMT 2019-2022 in `en-de`, `en-ru`, and `en-pl` (when available). The newswire text hardly ever contains the AUXILIARY type of ambiguity. Formality comprises the bulk of the examples, and upon further inspection, we find a severe bias towards the formal register, with a 1 to 7 ratio of informal to formal—likely due to the characteristics of the domain. Further, we suspect the sparseness in contextual ambiguities is important to consider when evaluating these systems.

7 Related Work

Work in contextual machine translation can be divided into three categories: (1) the publication of resources, similar to this work; (2) alterations on the training paradigm via architecture or data input; (3) evaluation metrics.

This work largely follows the path set forth by those who have previously published resources on the detection of gender, pronouns, and formality (Guillou and Hardmeier, 2016; Müller et al., 2018; Bawden et al., 2018; Voita et al., 2019; Lopes et al., 2020). (Currey et al., 2022) produces a gender-based evaluation dataset using human annotators, but covers the complement of this work: gender assigned to humans rather than inanimate objects. In addition to the manual pipelines, recent work has been done to promote the automatic detection of these phenomena. Nadejde et al. (2022) implements a cross-lingual mutual information metric that tags words as needing additional context. The tags were found to often overlap with the variety discussed in this work. Fernandes et al. (2023) also use a mutual-information based score to select data that is then used to derive a similar rule-based extraction approach, but do not release evaluation sets.

A substantial amount of work has been done to allow traditional neural models to handle additional input. Some approaches involve more complex architectures or modifications to training paradigms incorporate longer sequences (Miculicich et al., 2018; Bao et al., 2021), but Sun et al. (2022) showed that unaltered Transformers can handle longer sequences. Other work has focused on leveraging and cleaning the available data, since large-scale document bitext is lacking (Junczys-Dowmunt, 2019; Post and Junczys-Dowmunt, 2023).

Lastly, many have realized that BLEU, COMET, or other sentence-level metrics will not address the distinction in document-level performance. Vernikos et al. (2022) proposed a new method for adjusting current methods to adjust for document-level inputs. Jiang et al. (2022) proposed BlonDe, an entirely novel metric for document-level evaluation. We hope this work complements these works and serves to further the field in its aspirations towards true context-aware translation.

8 Summary

Machine translation systems face a performance ceiling that can't be overcome so long as they continue to operate at the sentence level. A major obstacle to that transition is the unavailability of test sets in many languages and for many contextual phenomena. The goal of this work has been to help address that problem. The extraction pipeline proposed in this paper can be used to identify and generate new test sets which contain linguistic phenomena that can only be consistently translated by contextual systems. The application of our pipeline to the OpenSubtitles dataset in seven languages provides a new set of evaluation sets including a wider set of languages and phenomena than were available before. Further, we hope that the extensibility of our pipeline to new phenomena and languages allows for others to build upon this work to expand resources and coverage. The CTXPRO datasets and extraction pipeline are available as open source from `https://github.com/rewicks/ctxpro`.

References

Farhad Akhbardeh, Arkady Arkhangorodsky, Magdalena Biesialska, Ondřej Bojar, Rajen Chatterjee, Vishrav Chaudhary, Marta R. Costa-jussa,

Cristina España-Bonet, Angela Fan, Christian Federmann, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Leonie Harter, Kenneth Heafield, Christopher Homan, Matthias Huck, Kwabena Amponsah-Kaakyire, Jungo Kasai, Daniel Khashabi, Kevin Knight, Tom Kocmi, Philipp Koehn, Nicholas Lourie, Christof Monz, Makoto Morishita, Masaaki Nagata, Ajay Nagesh, Toshiaki Nakazawa, Matteo Negri, Santanu Pal, Allahsera Auguste Tapo, Marco Turchi, Valentin Vydrin, and Marcos Zampieri. 2021. Findings of the 2021 conference on machine translation (WMT21). In *Proceedings of the Sixth Conference on Machine Translation*, pages 1–88, Online. Association for Computational Linguistics.

Guangsheng Bao, Yue Zhang, Zhiyang Teng, Boxing Chen, and Weihua Luo. 2021. G-transformer for document-level machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3442–3455, Online. Association for Computational Linguistics.

Loïc Barrault, Magdalena Biesialska, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Matthias Huck, Eric Joanis, Tom Kocmi, Philipp Koehn, Chi-kiu Lo, Nikola Ljubešić, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Santanu Pal, Matt Post, and Marcos Zampieri. 2020. Findings of the 2020 conference on machine translation (WMT20). In *Proceedings of the Fifth Conference on Machine Translation*, pages 1–55, Online. Association for Computational Linguistics.

Loïc Barrault, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, Christof Monz, Mathias Müller, Santanu Pal, Matt Post, and Marcos Zampieri. 2019. Findings of the 2019 conference on machine translation (WMT19). In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 1–61, Florence, Italy. Association for Computational Linguistics.

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.

Anna Currey, Maria Nadejde, Raghavendra Reddy Pappagari, Mia Mayer, Stanislas Lauly, Xing Niu, Benjamin Hsu, and Georgiana Dinu. 2022. MT-GenEval: A counterfactual and contextual dataset for evaluating gender accuracy in machine translation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4287–4299, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Patrick Fernandes, Kayo Yin, Emmy Liu, André Martins, and Graham Neubig. 2023. When does translation require context? a data-driven, multilingual exploration. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 606–626, Toronto, Canada. Association for Computational Linguistics.

Liane Guillou and Christian Hardmeier. 2016. PROTEST: A test suite for evaluating pronouns in machine translation. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 636–643, Portorož, Slovenia. European Language Resources Association (ELRA).

Masoud Jalili Sabet, Philipp Dufter, François Yvon, and Hinrich Schütze. 2020. SimAlign: High quality word alignments without parallel training data using static and contextualized embeddings. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1627–1643, Online. Association for Computational Linguistics.

Yuchen Jiang, Tianyu Liu, Shuming Ma, Dongdong Zhang, Jian Yang, Haoyang Huang, Rico Sennrich, Ryan Cotterell, Mrinmaya Sachan, and Ming Zhou. 2022. BlonDe: An automatic evaluation metric for document-level machine translation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1550–1565, Seattle, United States. Association for Computational Linguistics.

Marcin Junczys-Dowmunt. 2019. Microsoft translator at wmt 2019: Towards large-scale document-level neural machine translation.

Tom Kocmi, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Thamme Gowda, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Rebecca Knowles, Philipp Koehn, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Michal Novák, Martin Popel, and Maja Popović. 2022. Findings of the 2022 conference on machine translation (WMT22). In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 1–45, Abu Dhabi, United Arab Emirates (Hybrid). 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.

Pierre Lison and Jörg Tiedemann. 2016. OpenSubtitles2016: Extracting large parallel corpora from

movie and TV subtitles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 923–929, Portorož, Slovenia. European Language Resources Association (ELRA).

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.

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.

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.

Maria Nadejde, Anna Currey, Benjamin Hsu, Xing Niu, Marcello Federico, and Georgiana Dinu. 2022. CoCoA-MT: A dataset and benchmark for contrastive controlled MT with application to formality. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 616–632, Seattle, United States. Association for Computational Linguistics.

Shon Otmazgin, Arie Cattan, and Yoav Goldberg. 2022. F-coref: Fast, accurate and easy to use coreference resolution.

Matt Post and Marcin Junczys-Dowmunt. 2023. Escaping the sentence-level paradigm in machine translation.

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.

Zewei Sun, Mingxuan Wang, Hao Zhou, Chengqi Zhao, Shujian Huang, Jiajun Chen, and Lei Li. 2022. Rethinking document-level neural machine translation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3537–3548, Dublin, Ireland. 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.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need.

Giorgos Vernikos, Brian Thompson, Prashant Mathur, and Marcello Federico. 2022. Embarrassingly easy document-level MT metrics: How to convert any pretrained metric into a document-level metric. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 118–128, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

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

Rachel Wicks and Matt Post. 2022. Does sentence segmentation matter for machine translation? In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 843–854, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

A Additional Materials

	English (T_e)	German (T_t)
Rule	Lemma	Illegal Lemmas
DO.ELL	do	machen, tun, haben, können
WOULD.ELL	would	machen, tun, haben
WILL.ELL	will	machen, tun, haben, werden

Table 9: German auxiliary rules. English must have specified lemma. German alignment cannot have a lemma in the specified list.

	English (T_e)	Polish (T_t)
Rule	Lemma	Illegal Lemmas
DO.ELL	do	robić
WOULD.ELL	would	robić, by być, być, by, móc
WILL.ELL	will	robić, by być, być, by, móc, iść

Table 10: Polish auxiliary rules. English must have specified lemma. Polish alignment cannot have a lemma in the specified list.

	English (T_e)	Russian (T_t)
Rule	Lemma	Illegal Lemmas
DO.ELL	do	Делать
WOULD.ELL	would	Делать
WILL.ELL	will	Делать

Table 11: Russian auxiliary rules. English must have specified lemma. Russian alignment cannot have a lemma in the specified list.

	English (T_e)	Portugese (T_t)
Rule	Lemma	Illegal Lemmas
DO.ELL	do	fazer
WOULD.ELL	would	fazer, poder
WILL.ELL	will	fazer, ir

Table 12: Portuguese auxiliary rules. English must have specified lemma. Portuguese alignment cannot have a lemma in the specified list.

	English (T_e)	Italian (T_t)
Rule	Lemma	Illegal Lemmas
DO.ELL	do	fare
WOULD.ELL	would	fare, potere, volere
WILL.ELL	will	fare, andare

Table 13: Italian auxiliary rules. English must have specified lemma. Italian alignment cannot have a lemma in the specified list.

	English (T_e)			Spanish (T_t)			Coref English (C_e)	Coref Spanish (C_t)		
Rule	Form	POS	Case	Form	POS	Case	POS	POS	Gender	Number
NOM.FEM.SING	it	PNOUN	Nom.	ella	PNOUN	*	NOUN	NOUN	Fem.	Sing.
NOM.MASC.SING	it	PNOUN	Nom.	él	PNOUN	*	NOUN	NOUN	Masc.	Sing.
NOM.FEM.PLUR	it	PNOUN	Nom.	ellas	PNOUN	*	NOUN	NOUN	Fem.	Plur.
NOM.MASC.PLUR	it	PNOUN	Nom.	ellos	PNOUN	*	NOUN	NOUN	Masc.	Plur.
ACC.MASC.SING	it	PNOUN	Acc.	lo	PNOUN	*	NOUN	NOUN	Masc.	Sing.
ACC.FEM.SING	it	PNOUN	Acc.	la	PNOUN	*	NOUN	NOUN	Fem.	Sing.
ACC.MASC.PLUR	them	PNOUN	Acc.	los	PNOUN	*	NOUN	NOUN	Masc.	Sing.
ACC.FEM.PLUR	them	PNOUN	Acc.	las	PNOUN	*	NOUN	NOUN	Fem.	Sing.
DISJ.MASC.SING	it	PNOUN	-Nom.	él	PNOUN	*	NOUN	NOUN	Masc.	Sing.
DISJ.MASC.SING.ALT	it	PNOUN	-Nom	ello	PNOUN	*	NOUN	NOUN	Masc.	Sing.
DISJ.FEM.SING	it	PNOUN	-Nom	ella	PNOUN	*	NOUN	NOUN	Fem.	Sing.
DISJ.MASC.PLUR	them	PNOUN	-Nom	ellos	PNOUN	*	NOUN	NOUN	Masc.	Plur.
DISJ.FEM.PLUR	them	PNOUN	-Nom	ellas	PNOUN	*	NOUN	NOUN	Fem.	Plur.
NOM.INFORM.SING	you	PNOUN	Nom.	tú	PNOUN	*	-	-	-	-
NOM.FORM.SING	you	PNOUN	Nom.	usted	PNOUN	*	-	-	-	-
NOM.FORM.PLUR	you	PNOUN	Nom.	ustedes	PNOUN	*	-	-	-	-
NOM.INFORM.PLUR.MASC	you	PNOUN	Nom.	vosotros	PNOUN	*	-	-	-	-
NOM.INFORM.PLUR.FEM	you	PNOUN	Nom.	vosotras	PNOUN	*	-	-	-	-
ACC.INFORM.SING	you	PNOUN	Acc.	te	PNOUN	*	-	-	-	-
ACC.FORM.SING.MASC	you	PNOUN	Acc.	lo	PNOUN	*	-	-	-	-
ACC.FORM.SING.FEM	you	PNOUN	Acc.	la	PNOUN	*	-	-	-	-
ACC.FORM.PLUR.MASC	you	PNOUN	Acc.	los	PNOUN	*	-	-	-	-
ACC.FORM.PLUR.FEM	you	PNOUN	Acc.	las	PNOUN	*	-	-	-	-
ACC.INFORM.PLUR	you	PNOUN	Acc.	os	PNOUN	*	-	-	-	-
DISJ.INFORM.SING	you	PNOUN	-Nom.	ti	PNOUN	*	-	-	-	-
DISJ.INFORM.SING.ALT	you	PNOUN	-Nom.	contigo	PNOUN	*	-	-	-	-
DISJ.FORM.SING	you	PNOUN	-Nom.	usted	PNOUN	*	-	-	-	-
DISJ.INFORM.PLUR.MASC	you	PNOUN	-Nom.	vosotros	PNOUN	*	-	-	-	-
DISJ.INFORM.PLUR.FEM	you	PNOUN	-Nom.	vosotras	PNOUN	*	-	-	-	-
DISJ.FORM.PLUR	you	PNOUN	-Nom.	ustedes	PNOUN	*	-	-	-	-

Table 14: Spanish Pronoun Rules

	English (T_e)			French (T_t)			Coref English (C_e)	Coref French (C_t)		
Rule	Form	POS	Case	Form	POS	Case	POS	POS	Gender	Number
NOM.FEM.SING	it	PNOUN	Nom.	elle	PNOUN	*	NOUN	NOUN	Fem.	Sing.
NOM.MASC.SING	it	PNOUN	Nom.	il	PNOUN	*	NOUN	NOUN	Masc.	Sing.
NOM.FEM.PLUR	they	PNOUN	Nom.	elles	PNOUN	*	NOUN	NOUN	Fem.	Plur.
NOM.MASC.PLUR	they	PNOUN	Nom.	ils	PNOUN	*	NOUN	NOUN	Masc.	Plur.
ACC.MASC.SING	it	PNOUN	Acc.	le	PNOUN	*	NOUN	NOUN	Masc.	Sing.
ACC.FEM.SING	it	PNOUN	Acc.	la	PNOUN	*	NOUN	NOUN	Fem.	Sing.
GEN.FEM.SING.1S	mine	PNOUN	*	mienne	PNOUN	*	NOUN	NOUN	Fem.	Sing.
GEN.FEM.SING.1P	ours	PNOUN	*	la nôtre	PNOUN	*	NOUN	NOUN	Fem.	Sing.
GEN.FEM.SING.2S	yours	PNOUN	*	tienne	PNOUN	*	NOUN	NOUN	Fem.	Sing.
GEN.FEM.SING.2P	yours	PNOUN	*	la vôtre	PNOUN	*	NOUN	NOUN	Fem.	Sing.
GEN.FEM.SING.3SM	his	PNOUN	*	sienne	PNOUN	*	NOUN	NOUN	Fem.	Sing.
GEN.FEM.SING.3SF	hers	PNOUN	*	sienne	PNOUN	*	NOUN	NOUN	Fem.	Sing.
GEN.FEM.SING.3N	its	PNOUN	*	sienne	PNOUN	*	NOUN	NOUN	Fem.	Sing.
GEN.FEM.SING.3P	theirs	PNOUN	*	la leur	PNOUN	*	NOUN	NOUN	Fem.	Sing.
NOM.INFORM.SING	you	PNOUN	Nom.	tu	PNOUN	*	-	-	-	-
NOM.FORM+PLUR	you	PNOUN	Nom.	vous	PNOUN	*	-	-	-	-
ACC.INFORM.SING	you	PNOUN	Acc.	te	PNOUN	*	-	-	-	-
ACC.INFORM.SING.LIAS	you	PNOUN	Acc.	t’	PNOUN	*	-	-	-	-
ACC.FORM+PLUR	you	PNOUN	Acc.	vous	PNOUN	*	-	-	-	-
DISJ.INFORM.SING	you	PNOUN	-Nom	toi	PNOUN	*	-	-	-	-

Table 15: A sampling of French pronoun rules (abridged). Some forms left off for space.

	English (T_e)			Italian (T_t)			Coref English (C_e)	Coref Italian (C_t)		
Rule	Form	POS	Case	Form	POS	Case	POS	POS	Gender	Number
NOM.MASC.SING	it	PNOUN	Nom.	lui	PNOUN	*	NOUN	NOUN	Masc.	Sing.
NOM.FEM.SING	it	PNOUN	Nom.	lei	PNOUN	*	NOUN	NOUN	Fem.	Sing.
ACC.MASC.SING	it	PNOUN	Acc.	lo	PNOUN	*	NOUN	NOUN	Masc.	Sing.
ACC.FEM.SING	it	PNOUN	Acc.	la	PNOUN	*	NOUN	NOUN	Fem.	Sing.
ACC.MASC.PLUR	them	PNOUN	Acc.	li	PNOUN	*	NOUN	NOUN	Masc.	Plur.
ACC.FEM.PLUR	them	PNOUN	Acc.	le	PNOUN	*	NOUN	NOUN	Fem.	Plur.
DAT.MASC.SING	it	PNOUN	Acc.	gli	PNOUN	*	NOUN	NOUN	Masc.	Sing.
DAT.FEM.SING	it	PNOUN	Acc.	le	PNOUN	*	NOUN	NOUN	Fem.	Sing.
DISJ.MASC.SING	it	PNOUN	-Nom	lui	PNOUN	*	NOUN	NOUN	Masc.	Sing.
DISJ.FEM.SING	it	PNOUN	-Nom	lei	PNOUN	*	NOUN	NOUN	Fem.	Sing.
GEN.FEM.SING.1S	mine	PNOUN	*	mia	PNOUN	*	NOUN	NOUN	Fem.	Sing.
GEN.FEM.SING.2S	yours	PNOUN	*	tua	PNOUN	*	NOUN	NOUN	Fem.	Sing.
GEN.FEM.SING.3M	his	PNOUN	*	sua	PNOUN	*	NOUN	NOUN	Fem.	Sing.
GEN.FEM.SING.3F	hers	PNOUN	*	sua	PNOUN	*	NOUN	NOUN	Fem.	Sing.
GEN.FEM.SING.3N	its	PNOUN	*	sua	PNOUN	*	NOUN	NOUN	Fem.	Sing.
GEN.FEM.SING.2P	yours	PNOUN	*	vostra	PNOUN	*	NOUN	NOUN	Fem.	Sing.
GEN.FEM.SING.3P	theirs	PNOUN	*	loro	PNOUN	*	NOUN	NOUN	Fem.	Sing.
NOM.INFORM.SING	you	PNOUN	*	tu	PNOUN	*	-	-	-	-
NOM.FORM.SING	you	PNOUN	*	lei	PNOUN	*	-	-	-	-
NOM.INFORM.PLUR	you	PNOUN	*	voi	PNOUN	*	-	-	-	-

Table 16: A sampling of Italian Pronoun Rules. We do not consider the conflated Italian pronouns which combine accusatives and datives which co-occur. English case is used as it is a better model. Accusative is used for dative since the SpaCy models conflate the two in English.

	English (T_e)			Polish (T_t)			Coref English (C_e)	Coref Polish (C_t)		
Rule	Form	POS	Case	Form	POS	Case	POS	POS	Gender	Number
NOM.NEUT.SING	it	PNOUN	*	ono	PNOUN	Nom.	NOUN	NOUN	Neut.	Sing.
NOM.MASC.SING	it	PNOUN	*	on	PNOUN	Nom.	NOUN	NOUN	Masc.	Sing.
NOM.FEM.SING	it	PNOUN	*	ona	PNOUN	Nom.	NOUN	NOUN	Fem.	Sing.
ACC.NEUT.SING	it	PNOUN	*	je	PNOUN	Acc.	NOUN	NOUN	Neut.	Sing.
ACC.NEUT.SING.ALT1	it	PNOUN	*	nie	PNOUN	Acc.	NOUN	NOUN	Neut.	Sing.
ACC.MASC.SING	it	PNOUN	*	je	PNOUN	Acc.	NOUN	NOUN	Masc.	Sing.
ACC.MASC.SING.ALT	it	PNOUN	*	niego	PNOUN	Acc.	NOUN	NOUN	Masc.	Sing.
ACC.FEM.SING	it	PNOUN	*	ją	PNOUN	Acc.	NOUN	NOUN	Fem.	Sing.
GEN.NEUT.SING	it	PNOUN	*	jego	PNOUN	Gen.	NOUN	NOUN	Neut.	Sing.
GEN.NEUT.SING.ALT1	it	PNOUN	*	niego	PNOUN	Gen.	NOUN	NOUN	Neut.	Sing.
GEN.NEUT.SING.ALT2	it	PNOUN	*	go	PNOUN	Gen.	NOUN	NOUN	Neut.	Sing.
GEN.MASC.SING	it	PNOUN	*	je	PNOUN	Gen.	NOUN	NOUN	Masc.	Sing.
GEN.MASC.SING.ALT1	it	PNOUN	*	niego	PNOUN	Gen.	NOUN	NOUN	Masc.	Sing.
GEN.FEM.SING	it	PNOUN	*	jej	PNOUN	Gen.	NOUN	NOUN	Fem.	Sing.
GEN.FEM.SING.ALT1	it	PNOUN	*	niej	PNOUN	Gen.	NOUN	NOUN	Fem.	Sing.
LOC.NEUT.SING	it	PNOUN	*	nim	PNOUN	Loc.	NOUN	NOUN	Neut.	Sing.
LOC.MASC.SING	it	PNOUN	*	nim	PNOUN	Loc.	NOUN	NOUN	Masc.	Sing.
LOC.FEM.SING	it	PNOUN	*	niej	PNOUN	Loc.	NOUN	NOUN	Fem.	Sing.
DAT.NEUT.SING	it	PNOUN	*	jemu	PNOUN	Dat.	NOUN	NOUN	Neut.	Sing.
INS.NEUT.SING	it	PNOUN	*	nim	PNOUN	Ins.	NOUN	NOUN	Neut.	Sing.
NOM.INFORM.SING	you	PNOUN	*	ty	PNOUN	Nom.	-	-	-	-
ACC.INFORM.SING	you	PNOUN	*	ciebie	PNOUN	Acc.	-	-	-	-
NOM.FORM.SING.FEM	you	PNOUN	*	pani	PNOUN	Nom.	-	-	-	-
ACC.FORM.SING.FEM	you	PNOUN	*	panią	PNOUN	Acc.	-	-	-	-

Table 17: A sampling of Polish Pronoun Rules. Some left off for space.

		DE	ES	FR	IT	PL	PT	RU
GENDER	sent.	29.0	35.4	32.6	28.7	23.8	27.8	24.7
	doc.	**33.8**	**38.7**	**37.2**	**32.7**	**27.1**	**31.3**	**27.6**
		+4.8	+4.6	+2.9	+3.3	+3.5	+4.0	+3.3
ANIMACY	sent.	33.3	44.3	37.5	35.1	29.8	40.5	32.1
	doc.	**37.7**	**48.3**	**40.6**	**37.6**	**32.3**	**44.4**	**36.0**
		+4.4	+4.0	+3.1	+2.5	+2.5	+3.9	+3.9
FORMALITY	sent.	26.4	32.0	28.4	21.7	36.1	29.2	34.3
	doc.	**28.4**	**35.6**	**30.2**	**23.4**	**37.1**	**31.3**	**36.1**
		+2.0	+3.6	+1.8	+1.7	+1.0	+2.1	+1.8
AUXILIARY	sent.	17.7	17.3	14.9	17.8	15.3	15.8	19.9
	doc.	**30.1**	**33.4**	**33.6**	**34.7**	**33.1**	**32.5**	**42.2**
		+12.4	+16.1	+18.7	+16.9	+17.8	+16.7	+22.3
INFLECTION	sent.	-	-	-	-	27.3	-	27.7
	doc.	-	-	-	-	**30.7**	-	**29.9**
		-	-	-	-	+2.2	-	+3.4

Table 18: BLEU scores to evaluate the translation quality of this model. Higher is better. *sent.* denotes that no additional context was given while *doc.* was given five consecutive sentences. All translations produced by DeepL commercial API.