Controlling Translation Formality Using Pre-trained Multilingual Language Models

Elijah Rippeth* and **Sweta Agrawal*** and **Marine Carpuat**
Department of Computer Science
University of Maryland
{erip,sweagraw,marine}@cs.umd.edu

Abstract

This paper describes the University of Maryland's submission to the Special Task on Formality Control for Spoken Language Translation at IWSLT, which evaluates translation from English into 6 languages with diverse grammatical formality markers. We investigate to what extent this problem can be addressed with a *single multilingual model*, simultaneously controlling its output for target language and formality. Results show that this strategy can approach the translation quality and formality control achieved by dedicated translation models. However, the nature of the underlying pre-trained language model and of the finetuning samples greatly impact results.

1 Introduction

While machine translation (MT) research has primarily focused on preserving meaning across languages, linguists and lay-users alike have long known that pragmatic-preserving communication is an important aspect of the problem (Hovy, 1987). To address one dimension of this, several works have attempted to control aspects of formality in MT (Sennrich et al., 2016; Feely et al., 2019; Schioppa et al., 2021). Indeed, this research area was formalized as formality-sensitive machine translation (FSMT) by Niu et al. (2017), where the translation is not only a function of the source segment but also the desired target formality. The lack of gold translation with alternate formality for supervised training and evaluation has lead researchers to rely on manual evaluation and synthetic supervision in past work (Niu and Carpuat, 2020). Additionally, these works broadly adapt to formal and informal registers as opposed to specifically controlling grammatical formality.

The Special Task on Formality Control on Spoken Language Translation provides a new benchmark by contributing high-quality training datasets for diverse languages (Nădejde et al., 2022). In this task, a source segment in English is paired with two references which are minimally contrastive in grammatical formality, one for each formality level (formal and informal; Table 1). Training and test samples are provided in the domains of "telephony data" and "topical chat" (Gopalakrishnan et al., 2019) for four language pairs (English-{German (DE), Spanish (ES), Hindi (HI), Japanese(JA)}) and a test dataset for two additional "zero-shot" (ZS) language pairs (EN-{Russian (RU), Italian (IT)}). Markers of grammatical formality vary across these languages. Personal pronouns and verb agreement mark formality in many Indo-European languages (e.g., DE, HI, IT, RU, ES), while in JA, Korean (KO) and other languages, distinctions can be more extensive (e.g., using morphological markers) to express polite, respectful, and humble speech.

Source: **Do you like**$_1$ Legos? **did you**$_2$ ever play with them as a child or even later?
German Informal: **Magst du**$_1$ Legos? **Hast du**$_2$ jemals als Kind mit ihnen gespielt oder sogar später?
German Formal: **Mögen Sie**$_1$ Legos? **Haben Sie**$_2$ jemals als Kind mit ihnen gespielt oder sogar später?

Table 1: Contrastive formal and informal translations into German. Grammatical formality markers are bolded and aligned via indices.

In this work, we investigate how to control grammatical formality in MT for many languages with minimal resources. Specifically, we ask whether a single multilingual model can be finetuned to translate in the appropriate formality for any of the task languages. We introduce additive vector interventions to encode style on top of mT5-large (Xue et al., 2021) and mBART-large (Liu et al., 2020), and investigate the impact of finetuning on varying types of gold and synthetic samples to minimize reliance on manual annotation.

* equal contribution.

Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT 2022), pages 327 - 340

2 Method

Given an input sequence x, we design a *single model* that produces an output

$$y^* = \arg\max p(y|x, l, f; \theta_{LM}, \theta_F)$$

for any language l and formality level f considered in this task. The bulk of its parameters θ_{LM} are initialized with a pre-trained multilingual language model. A small number of additional parameters θ_F enable formality control. All parameters are finetuned for formality-controlled translation.

2.1 Multilingual Language Models

We experiment with two underlying multilingual models: 1) **mT5-large**[1] — a multilingual variant of T5 that is pre-trained on the Common Crawl-based dataset covering 101 languages and 2) **mBART-large**[2] — a Transformer encoder-decoder which supports multilingual machine translation for 50 languages. While mBART-large is pre-trained with parallel and monolingual supervision, mT5-large uses only monolingual dataset during the pre-training phase. Following standard practice, mT5 controls the output language, l, via prompts ("Translate to German"), and mBART replaces the beginning of sequence token in the decoder with target language tags (<2xx>).

2.2 Additive Formality Control

While large-scale pre-trained language models have shown tremendous success in multiple monolingual and multilingual controlled generation (Zhang et al., 2022) and style transfer tasks, their application to controlled cross-lingual text generation have been limited. Few-shot style-transfer approaches (Garcia et al., 2021; Riley et al., 2021; Krishna et al., 2022) hold the promise of minimal supervision but perform poorly on low-resource settings and their outputs lack diversity.

A popular way of introducing control when generating text with a particular style attribute is *tagging*, where the desired control tags (e.g., <2formal>) are appended to the source or the target sequence. However, as discussed in Schioppa et al. (2021), this approach has several limitations, including but not limited to the necessity of including the control tokens in the vocabulary at the start of the training, which restricts the enhancement of pre-trained models with controllability.

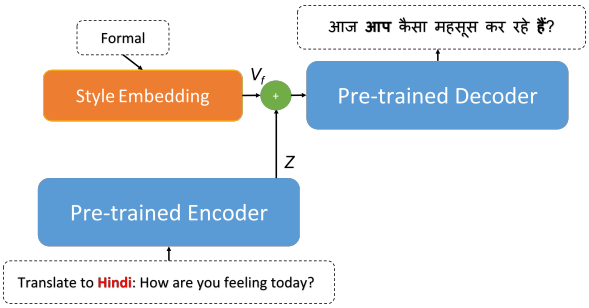


Figure 1: Controlling the output formality of a multilingual language model with additive interventions.

We introduce formality control by adapting the vector-valued interventions proposed by Schioppa et al. (2021) for machine translation (MT), as illustrated in Figure 1. Formally, given source text x, a formality level f, an encoder E and decoder D, parameterized by θ_{LM}, and a style embedding layer (Emb) parameterized by θ_F with the same output dimension as E, we have

$$Z = E(x), \quad V = \text{Emb}(f)$$
$$y = D(Z + V)$$

Our formality levels can take values corresponding to formal, informal, and "neutral" translations, the last of which is used to generate "generic" translations in which there is no difference in the grammatical formality of the translation of the source if translated formally or informally. Unlike Schioppa et al. (2021) who use a zero-vector as their neutral vector, we learn a separate vector.

2.3 Finetuning

Finetuning each multilingual model requires triplets of the form (x, y, f) for each available target language, l, where x, y and f are the source text, the reference translation and the formality label corresponding to the reference translation respectively. The loss function is then given by:

$$\mathcal{L} = \sum_{(x,y,l,f)} \log p(y|x, l, f; \theta_{LM}, \theta_F) \quad (1)$$

Given *paired contrastive* training samples of the form (X, Y_f, Y_{if}), as provided by the shared task, the loss decomposes into balanced formal and informal components, but does not explicitly exploit

[1] 24 layers with 1024 sized embeddings, 2816 FFN embedding dimension, and 16 heads for both encoder and decoder.

[2] 12 layers with 1024 sized embeddings, 4096 FFN embedding dimension, and 16 heads for both encoder and decoder.

Language	Size		Length			Style		
	Train	Test	Source	Formal	Informal	Avg. TER	# Phrasal	# Neutral
EN-DE	400	600	22.78	24.68	24.57	0.126	1.89	23
EN-ES	400	600	22.72	22.64	22.60	0.089	1.56	49
EN-HI	400	600	22.90	25.92	25.92	0.068	1.57	68
EN-JA	1000	600	24.61	32.43	30.80	0.165	2.47	20

Table 2: Shared Task Data Statistics: We use "13a" tokenization for all languages except Japanese for which we use "ja-mecab" implemented in the sacrebleu library.

the fact that Y_i and Y_f align to the same input:

$$\mathcal{L} = \sum_{(x,y_f,l)} \log p(y_f|x,l,f;\theta_{LM},\theta_F) + \sum_{(x,y_{if},l)} \log p(y_{if}|x,l,if;\theta_{LM},\theta_F) \quad (2)$$

2.4 Synthetic Supervision

Since paired contrastive samples are expensive to obtain, we explore the use of synthetic training samples to replace or complement them. This can be done either by automatically annotating naturally occurring bitext for formality, which yields formal and informal samples, and additionally by rewriting the translation to alter its formality to obtain paired contrastive samples. The second approach was used by Niu and Carpuat (2020) to control the register of MT output. However, since this shared task targets grammatical formality and excludes other markers of formal vs. informal registers, we focus on the first approach, thus prioritizing control on the nature of the formality markers in the output over the tighter supervision provided by paired contrastive samples.

Given a translation example (x, y), we predict a silver-standard formality label (f) for the target y using two distinct approaches:

- Rules (ES, DE, IT, RU): We label formality using heuristics based on keyword search, dependency parses, and morphological features. We use spaCy (Honnibal et al., 2020) to (non-exhaustively) retrieve documents that imply a necessarily formal, necessarily informal, or ambiguously formal label. In the case of an ambiguously formal label, we treat it as unambiguously formal (for examples, see B). The complete set of rules for each of the languages are included in the Appendix Table 12. While simple to implement, these heuristics privilege precision over recall, and risk biasing the synthetic data to the few grammatical aspects they encode.
- Classifiers (HI, JA, IT, RU): We train a binary formal vs. informal classifier on the shared task data (HI, JA) and on the synthetic data (IT, RU). Unlike rules, they can also be transferred in a zero-shot fashion to new languages, and might be less biased toward precision when well-calibrated.

3 Evaluation Settings

Data The shared task provides English source segments paired with two contrastive reference translations, one for each formality level (informal and formal) for four language pairs: EN-{DE, ES, JA, HI} in the *supervised* setting and two language pairs: EN-{RU, IT} in the *zero-shot* setting. The sizes and properties of the datasets for the supervised language pairs are listed in Table 2. Formal texts tend to be longer and more diverse than informal texts for JA compared to other language pairs. The percentage of neutral samples (same formal and informal outputs) vary from 2% (in JA) to 17% (in HI). In the *zero-shot* setting, 600 test samples are released for the two language pairs (RU, IT).

During development, the last 50 paired contrastive examples from each domain are set aside as a validation set for each of the supervised languages (TASK DEV) and use the remaining samples for training (TASK TRAIN).

Metrics We evaluate the translation quality of the detruecased detokenized outputs from each systems using **BLEU** (Papineni et al., 2002) and **COMET** (Rei et al., 2020). We use the 13A tokenizer to report SACREBLEU[3] scores for all languages, except Japanese, for which we use the JA-MECAB. We also report the official **formality accuracy** (ACC.). Given a set of hypotheses H, sets of corresponding phrase-annotated formal references F and informal

[3] https://pypi.org/project/sacrebleu/2.0.0/

Model	Target Language	Size	Source
Synthetic Finetuned	JA	15K	JParaCrawl (Morishita et al., 2020)
	HI	13K	CCMatrix (Schwenk et al., 2021b)
	IT, RU	15K	Paracrawl v8 (Bañón et al., 2020)
	DE	15K	CommonCrawl, Europarl v7 (Koehn, 2005)
	ES	15K	CommonCrawl, Europarl v7, UN (Ziemski et al., 2016)
Bilingual Baselines	DE,ES,IT,RU	20M	Paracrawl v9
	HI	0.7M	CCMatrix
	JA	3.2M	Wikimatrix (Schwenk et al., 2021a), JESC (Pryzant et al., 2018)

Table 3: Data sources from which unlabeled formality parallel examples are sampled for EN-X for training the *Synthetic Finetuned* and the *Bilingual* baselines.

references IF, and a function ϕ yielding phrase-level contrastive terms from a reference, the task-specific evaluation metric is defined as follows:

$$match_f = \sum_j \mathbb{1}\left[\phi(F_j) \in H_j \wedge \phi(IF_j) \notin H_j\right]$$

$$match_i = \sum_j \mathbb{1}\left[\phi(F_j) \notin H_j \wedge \phi(IF_j) \in H_j\right]$$

$$acc_j = \frac{match_j}{match_f + match_i}, \quad j \in \{f, i\}$$

We note that the task accuracy is a function of the number of *matches* in the hypotheses, not the number of *expected* phrases, i.e. $match_f + match_{if} \leq \|H\|$ and discuss the implications in the Appendix (Section C).

4 Experimental Conditions

We compare multilingual models, where a single model is used to generate formal and informal translations for all languages with bilingual models trained for each language pair, as detailed below.

4.1 Multilingual Models

Data We consider three finetuning settings:

- **Gold finetuned**: the model is finetuned only on *paired contrastive* shared task data (400 to 1000 samples per language pair).
- **Synthetic finetuned**: the model is finetuned on *synthetic silver-labelled triplets* (up to 7500 samples per formality level and language as described below).
- **Two-pass finetuned**: the *Synthetic finetuned* model is further finetuned on a mixture of gold data and 1000 examples re-sampled from the synthetic training set for unseen languages, which we use to avoid catastrophic forgetting from the silver finetuning stage.

Synthetic samples are drawn from multiple data sources (3), sampling at most 7500 examples for each language and formality level. [4] The formality labels are predicted as described in 2.4. Rule-based predictors directly give a label. With classifiers, we assign the formal label if $P(\text{formal}|y) \geq 0.85$ and informal if $P(\text{formal}|y) \leq 0.15$.

We additionally compare with the translations generated from the base mBART-large model with no finetuning, referred to as the "*formality agnostic* m**BART**-*large*".

Training settings We finetune mT5-large and mBART-large with a batch size of 2 and 8 respectively for 10 and 3 epochs respectively. We mask the formality labels used to generate vector-valued interventions with a probability of 0.2. The mT5-large model — "*synthetic finetuned* m**T5**-*large*" — is trained for an additional 5 epochs, with a batch size of 2 on a mixture of task data for seen languages and a subset of the sampled synthetic data for unseen languages. Again, we mask the formality tag with probability 0.2 except in the case of unseen languages where the formality tag is masked with probability 1.0, resulting in the "*two-pass finetuned* m**T5**-*large*" model.

Formality Classifiers Following Briakou et al. (2021), we finetune XLM-R on binary classification between formal and informal classes, using the shared task datasets for each of the supervised language pairs (DE, ES, JA, HI) and synthetic datasets for zero-shot language pairs (RU, IT). We treat the "neutral" samples as both "formal" and "informal" when training the classifiers. We use the Adam optimizer, a batch size of 32, and a learning rate of 5×10^{-3} to finetune for 3 epochs. We report

[4] We do not experiment with varying the sizes of the synthetic dataset due to the time constraints and leave it to the future work.

SAMPLES	TO	EN-DE		EN-HI		EN-JA		EN-ES	
		BLEU	ACC.	BLEU	ACC.	BLEU	ACC.	BLEU	ACC.
Paired Contrastive	F	35.0	**100**	28.7	98.7	33.1	95.3	32.6	**100**
Unpaired Triplets	F	**35.5**	**100**	**31.6**	**100**	**39.6**	**100**	**35.5**	**100**
Paired Contrastive	IF	32.7	98.5	26.4	98.3	32.3	**100**	33.8	**100**
Unpaired Triplets	IF	**35.9**	**98.6**	**30.9**	**98.4**	**40.3**	**100**	**39.6**	97.9

Table 4: Results on the TASK DEV split when training *Additive m*T5-*large* with and without contrastive examples: Sample diversity from Unpaired triplets improve BLEU and Accuracy over paired contrastive samples.

DATA	EN-DE	EN-HI	EN-JA	EN-ES
Paired Contrastive	0.397	0.371	0.421	0.505
Unpaired Triplets	0.459	0.415	0.460	0.580

Table 5: Results on the TASK DEV split: TER between generated formal and informal sentences.

the accuracy of the learned classifiers trained on the TASK TRAIN dataset in Appendix Table 14.

4.2 Bilingual Models

We consider two types of bilingual models:

1. **Formality Agnostic:** These models were released by the shared task organizers. Each model is bilingual and trained on a sample of 20 million lines from the Paracrawl Corpus (V9) using the Sockeye NMT toolkit. Models use big transformers with 20 encoder layers, 2 decoder layers, SSRU's in place of decoder self-attention, and large batch training.

2. **Formality Specific (Gold):** We finetune the models in [1] to generate a formal model and an informal model for each language pair (except the zero-shot language pairs).

The effective capacity of the bilingual, formality specific models is 3.14B parameters.Each model has 314M parameters, resulting in $(314 \times 2 \times 4) = 2.5$B parameters for the four supervised languages (DE, ES, HI, JA) and two pre-trained models $(314 \times 2) = 628$M parameters for the unseen languages (RU, IT).This is significantly larger than the capacities of our single multilingual models (Additive mT5-large: 1.25B, Additive mBART-large: 610M).

5 System Development Results

During system development, we explore the impact of different types of training samples and finetuning strategies on translation quality and formality accuracy on TASK DEV.

Contrastive Samples We estimate the benefits of fine-tuning on informal vs. formal translations of the same inputs for this task. We train two variants of the `gold finetuned mT5-large` model using 50% of the paired contrastive samples and 100% of the unpaired triplets (i.e., selecting one formality level per unique source sentence) from the TASK TRAIN samples (Table 4). Results show that sample diversity resulting from unpaired triplets leads to better translation quality as measured by BLEU (Average Gain: Formal +3.2. Informal +5.38), without compromising on the formality accuracy. Training with paired samples result in lower TER between formal and informal output compared to unpaired triplets (Table 5), suggesting that the outputs generated by the model trained on paired samples are more contrastive. This further motivates our `two-pass finetuned` model which uses gold contrastive samples on the final stage of finetuning to bias the model towards generating contrastive MT outputs.

While TASK DEV is too small to make definitive claims, we report our system development results in Tables 6 and 7. We observe that finetuning on gold contrastive examples (`gold-finetuned`) improves the translation quality and accuracy of the translation models (`formality-agnostic`), highlighting the importance of limited but high-quality in-domain supervision on the resulting models. Further, each of the `mT5-large` models improves in translation quality with additional data and training. While the results are dramatic due to size of both TASK TRAIN and TASK DEV, the trends validate the approach to augment both mBART-large and the mT5-large with additive interventions to control formality.

6 Official Results

Submissions We submit five variants of multilingual models (numbered **`[1-5]`** in Tables 8-11),

Model	EN-DE			EN-ES			EN-JA			EN-HI		
	BLEU	COMET	Acc.	BLEU	COMET	Acc.	BLEU	COMET	Acc.	BLEU	COMET	Acc.
Bilingual												
Formality Agnostic	33.2	0.432	33.8	41.3	0.675	24.5	13.0	-0.093	25.6	27.8	0.464	96.5
Formality Specific (Gold)	49.1	0.539	100.0	56.0	0.790	100.0	26.0	0.242	89.1	37.5	0.694	100.0
Multilingual Model												
mBART-large												
Formality Agnostic	33.3	0.295	68.9	27.0	0.120	56.5	18.3	-0.016	71.9	20.7	0.340	88.4
Gold Finetuned	42.8	0.462	95.9	41.1	0.548	97.7	24.7	0.326	89.4	29.6	0.678	95.6
mT5-large												
Gold Finetuned	53.3	0.260	**100.0**	53.5	0.427	**100.0**	49.8	0.645	98.1	43.5	0.359	**100.0**
Synthetic Finetuned	64.5	0.557	**100.0**	50.7	0.345	**100.0**	58.5	0.837	97.7	61.2	0.844	**100.0**
Two-pass Finetuned	**86.8**	**0.824**	**100.0**	**88.2**	**1.070**	**100.0**	**68.3**	**0.980**	**100.0**	**70.4**	**0.975**	**100.0**

Table 6: Results on the Task dev split in the ***formal*** *supervised* setting. Acc.: *formal* accuracy.

Model	EN-DE			EN-ES			EN-JA			EN-HI		
	BLEU	COMET	Acc.	BLEU	COMET	Acc.	BLEU	COMET	Acc.	BLEU	COMET	Acc.
Bilingual												
Formality Agnostic	37.2	0.470	66.2	45.8	0.691	75.5	13.5	-0.096	74.4	23.7	0.436	3.5
Formality Specific (Gold)	48.4	0.557	98.5	55.1	0.813	95.7	22.6	0.182	97.8	36.3	0.675	91.5
Multilingual Model												
mBART-large												
Formality Agnostic	29.3	0.262	31.1	26.3	0.101	43.5	16.2	-0.036	28.1	18.7	0.330	11.6
Gold Finetuned	39.6	0.456	76.5	40.4	0.582	95.3	21.6	0.289	72.7	27.7	0.631	82.8
mT5-large												
Gold Finetuned	52.8	0.232	**100.0**	53.8	0.513	**100.0**	47.3	0.617	**100.0**	41.7	0.144	**100.0**
Synthetic Finetuned	66.0	0.563	**100.0**	57.6	0.530	**100.0**	59.0	0.813	98.5	57.7	0.761	**100.0**
Two-pass Finetuned	**86.6**	**0.843**	**100.0**	**87.7**	**1.081**	**100.0**	**69.5**	**0.976**	**100.0**	**70.1**	**0.957**	**100.0**

Table 7: Results on the Task dev split in the ***informal*** *supervised* setting. Acc.: *informal* accuracy.

and compare them to the bilingual models built on top of the organizers' baselines. We first discuss results on the official test split for the *supervised* setting (Tables 8, 9). To better understand the degree of overall control afforded, we also report the average scores of the formal and informal settings in Table 10 before turning to the *zero-shot* setting in Table 11.

Multilingual Approach The best multilingual models (`[1]` & `[4]`) consistently outperform the `bilingual formality-agnostic` baselines, improving both translation quality (Worst-case gain in Average BLEU: Formal (+1.67), Informal: (+3.7)) and formality accuracy (Worst-case gain in Average Acc.: Formal (+40.38), Informal: (+31.6)). They approach the quality of formal and informal bilingual systems, but the gap in translation quality and formality accuracy varies across languages. While for DE and ES, there is a large difference in translation quality (approx. 10 BLEU points) between the multilingual models and the bilingual baselines, the multilingual models consistently get higher formality accuracy across language pairs and style directions and also perform comparably with the bilingual models in matching the translation quality for HI and JA. We attribute these differences to the amount of training data used across the language pairs (HI: 0.7M to DE 20M). This is an encouraging result, since the bilingual approach uses a much larger language-specific parameter budget and bitext for training than the all purpose multilingual models, which can benefit from transfer learning across languages.

mBART vs. mT5 The `gold finetuned mBART-large` model achieves the best overall translation quality among the multilingual variants as expected given that mBART-large is pre-trained on parallel text. Its translation quality is higher than that of mT5-large models according to BLEU and COMET for all languages except HI (informal), which could be attributed to the nature and amount of pre-training data used for HI. Its formality accuracy is in the 90's and within 5 percentage

	EN-DE			EN-ES			EN-JA			EN-HI		
	BLEU	COMET	ACC.	BLEU	COMET	ACC.	BLEU	COMET	ACC.	BLEU	COMET	ACC.
Bilingual Models												
Formality Agnostic	33.0	0.472	53.6	37.5	0.646	37.9	14.9	-0.102	23.3	**26.5**	**0.519**	98.8
Formal Gold Finetuned	**45.9**	**0.557**	**100.0**	**48.6**	**0.734**	**98.4**	**26.0**	**0.290**	**87.1**	23.0	0.303	**98.9**
Multilingual Models												
mBART-large												
Formality Agnostic	35.1	0.344	83.6	26.9	0.210	67.8	18.3	0.051	**93.4**	20.1	0.383	93.5
`[4]` Gold Finetuned	**38.6**	**0.484**	93.6	38.3	**0.549**	96.7	**26.1**	**0.397**	78.2	29.7	**0.691**	98.5
mT5-large												
`[3]` Gold Finetuned	7.9	-1.472	**100.0**	5.2	-1.340	97.0	8.9	-0.792	88.5	3.9	-1.152	**99.6**
`[2]` Synthetic Finetuned	22.1	0.076	92.4	28.1	0.274	86.5	16.3	-0.086	84.5	22.6	0.305	99.5
`[1]` Two-pass Finetuned	37.0	0.302	99.4	**38.6**	0.509	**99.5**	24.7	0.273	86.3	**29.9**	0.471	99.4

Table 8: Results on the official test split in the ***formal*** *supervised* setting. Best scores from multilingual and bilingual systems are **bolded**. Our official submissions to the shared task are numbered `[1-4]`.

MODEL	EN-DE			EN-ES			EN-JA			EN-HI		
	BLEU	COMET	ACC.	BLEU	COMET	ACC.	BLEU	COMET	ACC.	BLEU	COMET	ACC.
Bilingual Models												
Formality Agnostic	32.3	0.476	46.4	40.4	0.672	62.1	15.5	-0.094	76.7	20.8	0.493	1.2
Formality Specific (Gold)	**43.5**	**0.559**	**90.0**	**48.2**	**0.762**	**92.9**	**23.5**	**0.272**	**98.7**	**31.2**	**0.724**	**92.1**
Multilingual Models												
*m*BART-*large*												
Formality Agnostic	28.4	0.299	16.4	25.3	0.205	32.2	16.2	0.032	6.6	16.7	0.370	6.5
`[4]` Gold Finetuned	**36.1**	**0.472**	77.4	**38.3**	**0.549**	82.7	**22.8**	**0.346**	88.0	27.6	**0.670**	64.7
*m*T5-*large*												
`[3]` Gold Finetuned	7.3	-1.424	96.0	5.9	-1.295	**96.1**	7.2	-0.795	**98.9**	2.7	-1.205	96.5
`[2]` Synthetic Finetuned	21.7	0.046	91.4	28.2	0.337	91.6	13.6	-0.135	83.3	17.8	0.277	8.3
`[1]` Two-pass Finetuned	35.9	0.301	**96.5**	38.0	0.539	93.2	22.3	0.252	97.5	**29.2**	0.439	**98.7**

Table 9: Results on the official test split in the ***informal*** *supervised* setting. Best scores from multilingual and bilingual systems are **bolded**. Our official submissions to the shared task are numbered `[1-4]`.

points to the highest score for all languages except Japanese (78.2%) in the formal direction. In the informal direction, the gap between mBART-large and the best system on formality accuracy is larger across the board (Average Acc.: +19.3), suggesting that finetuning on gold data cannot completely recover an informal translation despite generally strong performance in formal translations.

Finetuning strategies Results show that the combination of synthetic and gold data is crucial to help the mT5-large-based model learn to translate and mark formality appropriately. Finetuning only on the gold data leads to overfitting: the model achieves high formality accuracy scores, but poor translation quality (BLEU < 10). Manual inspection of mT5-large-based system outputs suggests that translations often include tokens in the wrong language (Appendix Table 13). Finetuning on synthetic data improves translation quality substantially compared to gold data only (Average gain in BLEU: Formal (+15.8), Informal (+14.6)). Two-pass finetuning improves formality control (Average gain in ACC.: Formal (+5.43), Informal (+27.85)), with additional translation quality improvement across the board over synthetic-finetuned model (Average gain in BLEU: Formal (+10.27), Informal (+11.03); COMET: Formal (+0.247), Informal (+0.252)). While we primarily focused on the impact of synthetic supervision on mT5-large, we believe a similar investigation using mBART-large would yield interesting results and leave this as future work.

Performance across languages While the higher resource language pairs (DE, ES) achieve better translation quality (in BLEU and COMET) over the relatively lower resource languages (HI, JA), the formality accuracy is more comparable across the language pairs for the multilingual models

	Model	EN-DE			EN-ES			EN-JA			EN-HI		
		BLEU	COMET	Acc.	BLEU	COMET	Acc.	BLEU	COMET	Acc.	BLEU	COMET	Acc.
	Bilingual Models												
	Formality Agnostic	32.7	0.474	50.0	39.0	0.659	50.0	15.2	-0.100	50.0	23.7	0.506	50.0
	Formality Specific (Gold)	**44.7**	**0.558**	**95.0**	**48.4**	**0.748**	**95.7**	**24.8**	**0.281**	**92.9**	**27.1**	**0.513**	**95.5**
	Multilingual Models												
	*m*BART-*large*												
	Formality Agnostic	31.8	0.322	50.0	26.1	0.207	50.0	17.3	0.041	50.0	18.4	0.377	50.0
`[4]`	Gold Finetuned	**37.4**	**0.478**	85.5	**38.3**	**0.549**	89.7	**24.5**	**0.371**	83.1	28.7	**0.680**	81.6
	*m*T5-*large*												
`[3]`	Gold Finetuned	7.6	-1.448	**98.0**	5.6	-1.317	**96.6**	8.1	-0.794	**93.7**	3.3	-1.179	98.1
`[2]`	Synthetic Finetuned	21.9	0.061	91.9	28.2	0.305	89.1	15.0	-0.111	83.9	20.2	0.291	53.9
`[1]`	Two-pass Finetuned	36.5	0.301	**98.0**	**38.3**	0.524	96.4	23.5	0.263	91.9	**29.6**	0.455	**99.1**

Table 10: Averaged formal and informal results on the official test split in the *supervised* setting. Best scores from multilingual and bilingual systems are **bolded**. Our official submissions to the shared task are numbered `[1-4]`.

	Model	To Formal						To Informal					
		EN-IT			EN-RU			EN-IT			EN-RU		
		BLEU	COMET	Acc.	BLEU	COMET	Acc.	BLEU	COMET	Acc.	BLEU	COMET	Acc.
	Bilingual baselines	37.0	0.557	4.5	27.9	0.220	93.3	44.2	0.618	95.5	22.0	0.169	6.7
`[1]`	mT5-large (ZS)	27.6	0.306	32.8	22.7	0.123	**100.0**	32.6	0.379	**97.9**	17.0	0.058	1.1
`[4]`	mBART-large (ZS)	**30.2**	**0.545**	38.7	**26.2**	**0.275**	**100.0**	**35.0**	**0.597**	95.9	**20.8**	**0.203**	**13.8**
`[5]`	mT5-large (FS)	27.1	0.302	**49.7**	20.7	0.007	**100.0**	31.2	0.346	93.3	15.5	-0.050	1.8

Table 11: Results on the official test split for the *zero-shot* setting. Our official submissions to the shared task are numbered `[1-5]`.

(standard deviation: mT5-large (4), mBART-large (10)). We can observe that the task accuracy is lowest ($< 90\%$) when translating to formal Japanese. By inspection, we observe three broad classes of errors: 1) lexical choice, 2) cross-script matching, 3) ambiguity in politeness levels (Feely et al., 2019). Lexical choice is invariant in machine translation and is occasionally a valid error in the case of mistranslation, so we focus on the latter two error cases. Japanese has three writing systems and false positives in formality evaluation can occur when surface forms do not match as in the case of 面白い which can also be written as おもしろい (gloss: 'interesting'). Finally, there are cases in which the system and reference formality mismatch but can both be interpreted as formal (e.g., 働きます vs. 働く; gloss: 'work' (polite) vs. 'work' (formal)).

Zero-Shot We observe limited zero-shot transfer of grammatical formality to unseen languages (Table 11). For both `mBART-large` and `mT5-large` models, the EN-IT performance is biased towards informal translations, while EN-RU is biased in the formal direction. In the case of EN-IT, both mBART-large and mT5-large almost always interpret the English second person pronoun as second person *plural* when translating to formal, exploiting the ambiguity of English on the source side. By contrast, when generating informal translations, pronouns are typically preserved as singular. In comparison, with mT5-large-based translations into RU, we see almost unanimous preference toward the formal, likely due to sampling bias when curating the synthetic training set. We also observe that mBART-large prefers to translate in a formal manner irrespective of desired target. In addition, when mBART-large fails to account for the target formality, it often generates paraphrases of the formal target. These strong preferences might be symptoms of systematic differences in formality across languages in the training data of these models. Finally, the use of silver standard formality labels ("fully supervised" setting (FS)) does not improve over the zero-shot approach, with similar observations of mT5-large-based translations as outlined above. We observe that in the case of EN-RU, there is a higher incidence of code-switched translations. This may indicate noise introduced in the automatic labeling process and requires further examination in future work.

7 Related Work

Most MT approaches only indirectly capture the style properties of the target text. While efforts have been made to generate better outputs in their pragmatic context via controlling formality (Sennrich et al., 2016; Feely et al., 2019; Niu and Carpuat, 2020; Schioppa et al., 2021), complexity (Marchisio et al., 2019; Agrawal and Carpuat, 2019), gender (Rabinovich et al., 2017), these studies only focus a single language pair. Due to the paucity of style annotated corpora, zero-shot style transfer within and across languages has received a lot of attention. However, adapting pre-trained large-scale language models during inference using only a few examples (Garcia et al., 2021; Riley et al., 2021; Krishna et al., 2022) limits their transfer ability and the diversity of their outputs. While prior works use pre-trained language models like BERT, GPT to intialize θ_{LM} for improving translation quality (Guo et al., 2020; Zhu et al., 2019), in this work, we focus on adapting sequence-to-sequence multilingual models for controlled generation of a desired formality and study style transfer in multilingual supervised and zero-shot settings.

8 Conclusion

We present the University of Maryland's submission which examines the performance of a single multilingual model allowing control of both target language and formality. Results show that while multilingual FSMT models lag behind large, bilingual, formality-specific models in terms of MT quality, they show stronger formality control performance across all the language pairs. Furthermore, while synthetic unpaired triplets help `mT5-large` with FSMT performance and the two-stage finetuning process improves MT quality and contrastive task performance, `mBART-large` still outperforms this class of models, likely due to its large amount of pre-training supervision.

In future work, we suggest a deeper investigation of potentially confounding roles in the study of FSMT, such as the impact of formal register as compared to grammatical formality in training data. We also suggest a thorough analysis of *what* is transferred in the zero-shot setting. Finally, we recommend an audit of underlying pre-training and finetuning data sources for pre-trained multilingual models, which we believe hinder zero-shot formality transfer for EN-IT and EN-RU in which a single formality is strongly preferred.

References

Sweta Agrawal and Marine Carpuat. 2019. Controlling text complexity in 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 1549–1564, Hong Kong, China. Association for Computational Linguistics.

Marta Bañón, Pinzhen Chen, Barry Haddow, Kenneth Heafield, Hieu Hoang, Miquel Esplà-Gomis, Mikel L. Forcada, Amir Kamran, Faheem Kirefu, Philipp Koehn, Sergio Ortiz Rojas, Leopoldo Pla Sempere, Gema Ramírez-Sánchez, Elsa Sarrías, Marek Strelec, Brian Thompson, William Waites, Dion Wiggins, and Jaume Zaragoza. 2020. ParaCrawl: Web-scale acquisition of parallel corpora. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4555–4567, Online. Association for Computational Linguistics.

Eleftheria Briakou, Sweta Agrawal, Joel Tetreault, and Marine Carpuat. 2021. Evaluating the evaluation metrics for style transfer: A case study in multilingual formality transfer. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1321–1336, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Weston Feely, Eva Hasler, and Adrià de Gispert. 2019. Controlling Japanese honorifics in English-to-Japanese neural machine translation. In *Proceedings of the 6th Workshop on Asian Translation*, pages 45–53, Hong Kong, China. Association for Computational Linguistics.

Xavier Garcia, Noah Constant, Ankur Parikh, and Orhan Firat. 2021. Towards continual learning for multilingual machine translation via vocabulary substitution. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1184–1192, Online. Association for Computational Linguistics.

Karthik Gopalakrishnan, Behnam Hedayatnia, Qinlang Chen, Anna Gottardi, Sanjeev Kwatra, Anu Venkatesh, Raefer Gabriel, and Dilek Hakkani-Tür. 2019. Topical-Chat: Towards Knowledge-Grounded Open-Domain Conversations. In *Proc. Interspeech 2019*, pages 1891–1895.

Junliang Guo, Zhirui Zhang, Linli Xu, Hao-Ran Wei, Boxing Chen, and Enhong Chen. 2020. Incorporating bert into parallel sequence decoding with adapters. *Advances in Neural Information Processing Systems*, 33:10843–10854.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

Eduard Hendrik Hovy. 1987. *Generating Natural Language under Pragmatic Constraints*. Ph.D. thesis, USA. AAI8729079.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of Machine Translation Summit X: Papers*, pages 79–86, Phuket, Thailand.

Kalpesh Krishna, Deepak Nathani, Xavier Garcia, Bidisha Samanta, and Partha Talukdar. 2022. Few-shot controllable style transfer for low-resource multilingual settings.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Kelly Marchisio, Jialiang Guo, Cheng-I Lai, and Philipp Koehn. 2019. Controlling the reading level of machine translation output. In *Proceedings of Machine Translation Summit XVII: Research Track*, pages 193–203, Dublin, Ireland. European Association for Machine Translation.

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.

Xing Niu and Marine Carpuat. 2020. Controlling neural machine translation formality with synthetic supervision. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8568–8575.

Xing Niu, Marianna Martindale, and Marine Carpuat. 2017. A study of style in machine translation: Controlling the formality of machine translation output. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2814–2819, Copenhagen, Denmark. Association for Computational Linguistics.

Maria Nădejde, 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*, Seattle, USA. 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.

Reid Pryzant, Youngjoo Chung, Dan Jurafsky, and Denny Britz. 2018. JESC: Japanese-English subtitle corpus. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Ella Rabinovich, Raj Nath Patel, Shachar Mirkin, Lucia Specia, and Shuly Wintner. 2017. Personalized machine translation: Preserving original author traits. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 1074–1084, Valencia, Spain. 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.

Parker Riley, Noah Constant, Mandy Guo, Girish Kumar, David Uthus, and Zarana Parekh. 2021. TextSETTR: Few-shot text style extraction and tunable targeted restyling. 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 3786–3800, Online. Association for Computational Linguistics.

Andrea Schioppa, David Vilar, Artem Sokolov, and Katja Filippova. 2021. Controlling machine translation for multiple attributes with additive interventions. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6676–6696, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Holger Schwenk, Vishrav Chaudhary, Shuo Sun, Hongyu Gong, and Francisco Guzmán. 2021a. WikiMatrix: Mining 135M parallel sentences in 1620 language pairs from Wikipedia. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1351–1361, Online. Association for Computational Linguistics.

Holger Schwenk, Guillaume Wenzek, Sergey Edunov, Edouard Grave, Armand Joulin, and Angela Fan. 2021b. CCMatrix: Mining billions of high-quality parallel sentences on the web. 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 6490–6500, Online. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Controlling politeness in neural machine translation via side constraints. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human*

Language Technologies, pages 35–40, San Diego, California. Association for Computational Linguistics.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Hanqing Zhang, Haolin Song, Shaoyu Li, Ming Zhou, and Dawei Song. 2022. A survey of controllable text generation using transformer-based pre-trained language models. *arXiv preprint arXiv:2201.05337*.

Jinhua Zhu, Yingce Xia, Lijun Wu, Di He, Tao Qin, Wengang Zhou, Houqiang Li, and Tieyan Liu. 2019. Incorporating bert into neural machine translation. In *International Conference on Learning Representations*.

Michał Ziemski, Marcin Junczys-Dowmunt, and Bruno Pouliquen. 2016. The United Nations parallel corpus v1.0. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3530–3534, Portorož, Slovenia. European Language Resources Association (ELRA).

A Rules for Synthetic Data Curation

LANG	Formal	Informal
en-de	(P=2 ∈ M and Num=Plural ∈ M) or PP=Sie	P=2 ∈ M and Num=Plural ∉ M
en-es	P=2 ∈ M and Form=Polite ∈ M	P=2 ∈ M and Num=Singular ∈ M and Form=Polite ∉ M
en-it	PP=voi or PP=lei	PP=tu
en-ru	PP=Вы	PP=ты

Table 12: Rules for extracting formal and informal sentences for each language pair from existing bitext. P: Person; PP: Personal pronoun; N: Number; $x \in M$ indicates that some token within the sentence has morphological features matching x as produced by spaCy.

B Glosses

B.1 Necessarily formal

Appropriate pronouns with accompanying conjugation imply the sentence is grammatically formal.

(1) ¿Cuándo nació usted? (Spanish)
When born you (form.)?
'When were you (form.) born?'

(2) Woher kommen Sie? (German)
Where from come you (form.)?
'Where are you (form.) from?'

B.2 Necessarily informal

Appropriate pronouns with accompanying conjugation imply the sentence is grammatically informal. Note that Spanish is pro-drop, which relaxes the requirement on personal pronouns.

(3) ¿Cuándo naciste (tú)? (Spanish)
When born you (inf.)?
'When were you (inf.) born?'

(4) Woher kommst du? (German)
Where from come you (inf.)?
'Where are you (inf.) from?'

B.3 Ambiguously formal

Because Spanish is pro-drop, personal pronouns can be omitted depending on context. Since formal conjugations are shared with neutral third person subjects, this leaves ambiguity when the pronoun is dropped. For sake of gloss, we use ∅ to indicate a dropped pronoun.

(5) ¿Cuándo nació ∅?
When born {you (form.), he, she, it}?
'When {were you (form.), was {he, she, it}} born?'

C Official Evaluation

We report the number of examples labeled as FORMAL, INFORMAL, NEUTRAL, OTHER by the formality scorer for the best multilingual models (`[1, 4]`) and the baseline systems for each language pair and formality direction. As described in 3, the accuracy is computed based on *realized* matches, which excludes examples labelled as NEUTRAL and OTHER. Figure 2 shows that the number of these excluded NEUTRAL samples can range from 15% to 43%.

D Example Outputs

Source: Wow, that's awesome! Who is your favorite Baseball team? I like my Az team lol

German Formal Hypothesis: Wow, das ist toll! Wer ist Ihr Lieblings- Baseballteam? Ich mag meine Az-Team lol.

German Formal Reference: Wow, das ist fantastisch! Welches ist Ihr Lieblingsbaseballteam? Ich stehe auf mein AZ-Team lol.

German Informal Hypothesis: Wow, das ist toll! Wer ist dein Lieblings野球team? Ich mag meine Az Team lol.

German Informal Reference: Wow, das ist fantastisch! Welches ist dein Lieblingsbaseballteam? Ich stehe auf mein AZ-Team lol.

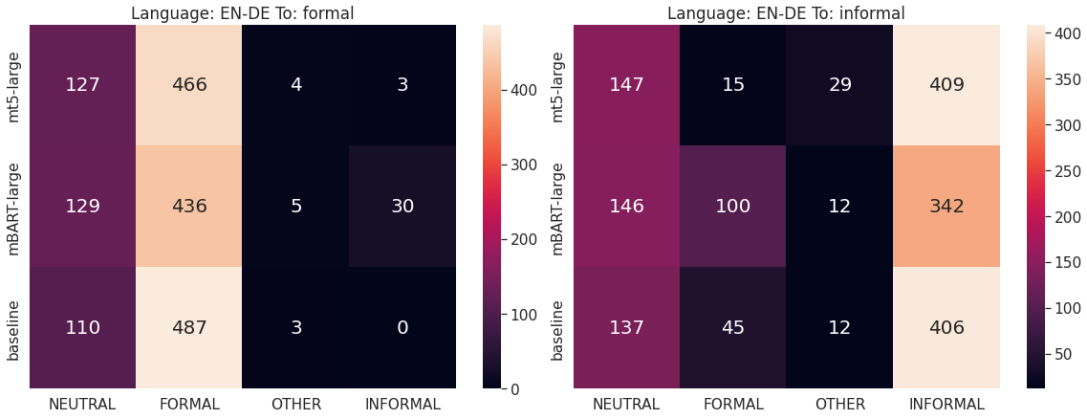
Table 13: Contrastive outputs from English-German. Note that there is not only variety in lexical choice between references and hypotheses, but also between hypotheses of varying formality (i.e., 野球 is "baseball" in Japanese)

E Accuracy of Formality Classifiers

We report the accuracy of the learned classifiers on the TASK TRAIN dataset in Table 14.

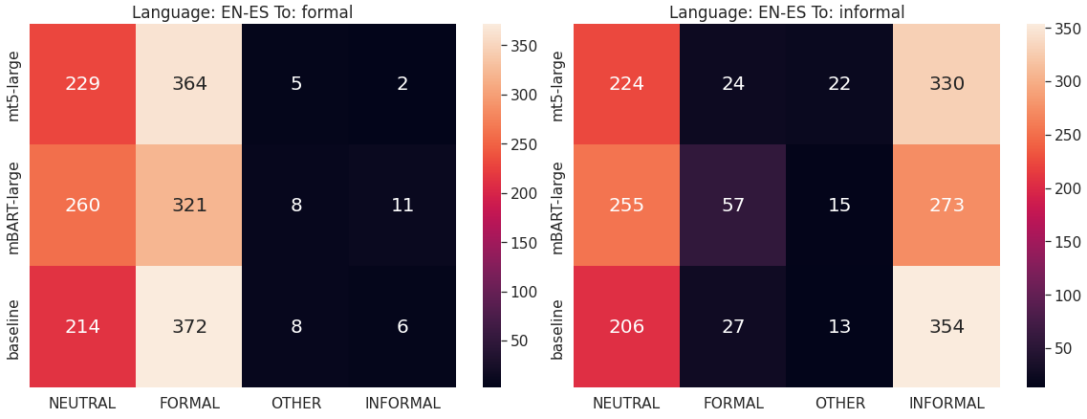
LANGUAGE	Accuracy	
	Formal	Informal
en-de	98%	99%
en-es	99%	92%
en-ja	98%	98%
en-hi	96%	95%

Table 14: Accuracy of trained formality classifiers on the TASK DEV dataset.



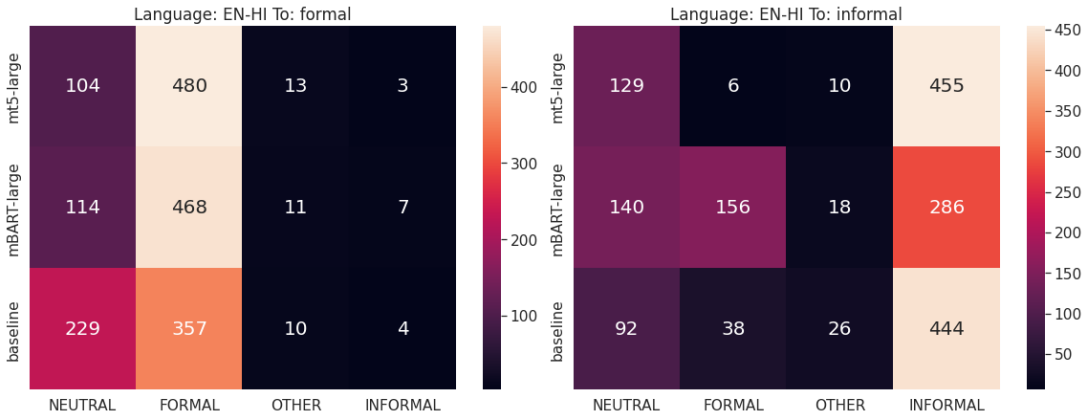
(a)

(b)



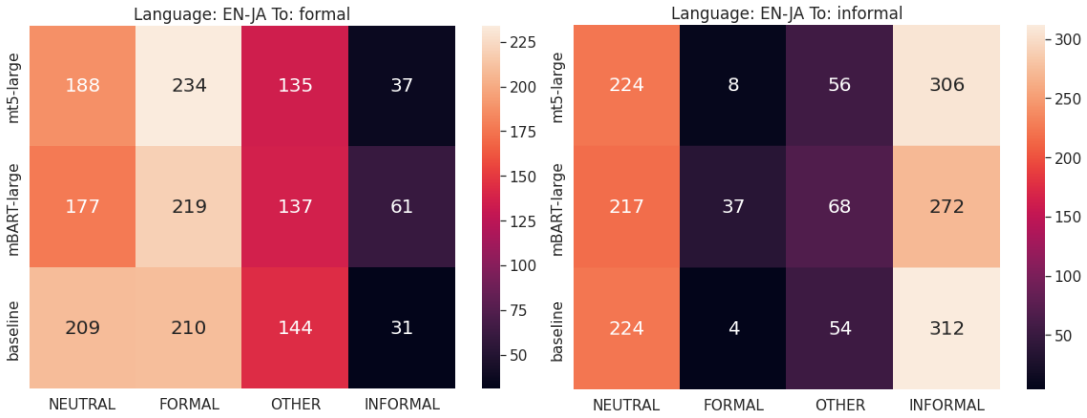
(c)

(d)



(e)

(f)



(g)

(h)

Figure 2: Class Distribution for the baseline, mBART-large and mT5-large systems for all the supervised language pairs.