Controlling Formality in Low-Resource NMT with Domain Adaptation and Re-Ranking: SLT-CDT-UoS at IWSLT2022

Sebastian T. Vincent, Loïc Barrault, Carolina Scarton
Department of Computer Science, University of Sheffield
Regent Court, 211 Portobello, Sheffield, S1 4DP, UK
{stvincent1, l.barrault, c.scarton}@shef.ac.uk

Abstract

This paper describes the SLT-CDT-UoS group's submission to the first Special Task on Formality Control for Spoken Language Translation, part of the IWSLT 2022 Evaluation Campaign. Our efforts were split between two fronts: data engineering and altering the objective function for best hypothesis selection. We used language-independent methods to extract formal and informal sentence pairs from the provided corpora; using English as a pivot language, we propagated formality annotations to languages treated as zero-shot in the task; we also further improved formality controlling with a hypothesis re-ranking approach. On the test sets for English-to-German and English-to-Spanish, we achieved an average accuracy of .935 within the constrained setting and .995 within unconstrained setting. In a zero-shot setting for English-to-Russian and English-to-Italian, we scored average accuracy of .590 for constrained setting and .659 for unconstrained.

1 Introduction

Formality-controlled machine translation enables the system user to specify the desired formality level at input so that the produced hypothesis is expressed in a formal or informal style. Due to discrepancies between different languages in formality expression, it is often the case that the same source sentence has several plausible hypotheses, each aimed at a different audience; leaving this choice to the model may result in an inappropriate translation.

This paper describes our team's submission to the first Special Task on Formality Control in SLT at IWSLT 2022 (Anastasopoulos et al., 2022), where the objective was to achieve control over binary expression of formality in translation (enable the translation pipeline to generate formal or informal translations depending on user input). The task evaluated translations from English (EN) into German (DE), Spanish (ES), Russian (RU), Italian (IT), Japanese (JA) and Hindi (HI). Among these, EN-{RU,IT} were considered zero-shot; for other pairs, small paired formality-annotated corpora were provided. The task ran in two settings: **constrained** (limited data and pre-trained model resources) and **unconstrained** (no limitations on either resource). Submissions within both the constrained and unconstrained track were additionally considered in two categories: full supervision and zero-shot.

Our submission consisted of four primary systems, one for each track/subtrack combination, and we focused on the EN-{DE,ES,RU,IT} language directions. We were interested in leveraging the provided formality-annotated triplets $(src, tgt_{\text{formal}}, tgt_{\text{informal}})$ to extract sufficiently large annotated datasets from the permitted training corpora, without using language-specific resources or tools. We built a multilingual translation model in the given translation directions and fine-tuned it on our collected data. Our zero-shot submissions used fine-tuning data only for the non-zero-shot pairs. To boost the formality control (especially within the constrained track), we included a formality-focused hypothesis re-ranking step. Our submissions to both tracks followed the same concepts, with the unconstrained one benefitting from larger corpora, and thus more fine-tuning data.

In Section 2 we describe our submission to the constrained track, including the data extraction step (Section 2.2, 2.3). Our approach begins with extending this small set to cover more samples by extracting them from the allowed corpora. We use a language-independent approach of domain adaptation for this. Then, we extract samples for the zero-shot pairs (EN-{RU,IT}) based on data collected for (EN-{DE,ES}). We then experiment with re-ranking the top n model hypotheses with a formality-focused objective function. Within our systems, we provide the formality information as a *tag* appended to the input of the model. Throughout the paper we use $\mathbb{F}$ to denote the *formal* style

Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT 2022), pages 341 - 350

and $\mathbb{I}$ to denote the *informal* style.

All our models submitted to the "supervised" subtracks achieved an average of $+.284$ accuracy point over a baseline for all EN-{DE,ES,RU,IT} test sets, while the "zero-shot" models achieved an average improvement of .124 points on the EN-{RU,IT} test sets. Our work highlights the potential of both data adaptation and re-ranking approaches in attribute control for NMT.

2 Constrained Track

The MuST-C textual corpus (Di Gangi et al., 2019) with quantities listed in Table 1 was the only data source allowed within the constrained track, alongside the IWSLT corpus of formality-annotated sentences (Nadejde et al., 2022). MuST-C is a collection of transcribed TED talks, all translated from English. The IWSLT data itself came from two domains: telephone conversations and topical chat (Gopalakrishnan et al., 2019). The data was additionally manually annotated at phrase level for formal and informal phrases, and the organisers provided an evaluation tool `scorer.py` which, given a set of hypotheses, used these annotations to match sought formal or informal phrases, yielding an accuracy score when the number of correct matches is greater than the number of incorrect matches[1]. This scorer skips test cases where no matches are found in the hypotheses.

In all our experiments we used the multilingual Transformer model architecture provided within `fairseq` (Ott et al., 2019). For our pre-training data we used the full MuST-C corpus. We applied SentencePiece (Kudo and Richardson, 2018) to build a joint vocabulary of 32K tokens across all languages. We list the model specifications in Table 2. Pre-training lasts 100K iterations or 63 epochs. We average checkpoints saved at roughly the last 10 epochs.

2.1 Formality Controlling

Once the model was pre-trained, we fine-tuned it on the supervised data to control the desired formality of the hypothesis with a *tagging* approach (Sennrich et al., 2016), whereby a formality-indicating tag is appended to the source input. This method has been widely used in research in various controlling tasks (e.g. Johnson et al., 2017; Vanmassenhove et al., 2018; Lakew et al., 2019).

[1] https://github.com/amazon-research/contrastive-controlled-mt/blob/main/IWSLT2022/scorer.py, accessed 8 April 2022.

2.2 Automatic Extraction of Formal and Informal Data

Since our approach was strongly dependent on the availability of labelled data, our initial efforts focused on making the training corpus larger by extracting sentence pairs with formal and informal target sentences from the provided MuST-C corpus. We made the assumption that similar sentences would correspond to a similar formality level. Thus, we decided to use the data selection approach to select the most similar sentence pairs from the out-of-domain corpus (MuST-C) to both the formal and informal sides of the IWSLT corpus, which we consider our in-domain data (each side separately).

Specifically, let $G = (G_{src}, G_{tgt})$ be the out-of-domain corpus (MuST-C), and let $S_{\mathbb{F}} = (S_{src}, S_{tgt,\mathbb{F}})$ and $S_{\mathbb{I}} = (S_{src}, S_{tgt,\mathbb{I}})$ be the in-domain corpora (IWSLT). For simplicity, let us focus on adaptation to $S_{\mathbb{F}}$.

Our adaptation approach focuses on the target-side sentences because the IWSLT corpus is paired (for each English sentence there is a formal and informal variant in the target language). The approach builds a vocabulary of non-singleton tokens from $S_{tgt,\mathbb{F}}$, then builds two language models: LM_S from $S_{tgt,\mathbb{F}}$ and LM_G from a random sample of 10K sentences from G_{tgt}; both language models use the originally extracted vocabulary. Then, we calculate the sentence-level perplexity $PP(LM_G, G_{tgt})$ and $PP(LM_S, G_{tgt})$. Finally, the sentence pairs within G are ranked by

$$PP(LM_S, G_{tgt}) - PP(LM_G, G_{tgt}).$$

Let $G_{sorted\_by\_\mathbb{F}}$, $G_{sorted\_by\_\mathbb{I}}$ denote the resulting corpora sorted by the perplexity difference. The intuition behind this approach is that sentences which use a certain formality will naturally rank higher on the ranked list for that formality, due to similarities in the used vocabulary.

To obtain the formal and informal corpora from the sorted data, we needed to decide on a criterion. Let $\mathbb{F}_{pos}$ and $\mathbb{I}_{pos}$ be the position of a sentence pair in the formal/informal ranking, respectively. Our first approach was simple: let $\mathcal{C}$ denote the size of the out-of-domain corpus; we implemented an $Assign_\theta$ function which, for a $\theta \in [0, \mathcal{C})$, assigned a label to the sentence pair (src, tgt), using the following rules:

$$Assign_\theta \begin{cases} \mathbb{F}, & \text{if } \mathbb{F}_{pos} < \theta < \mathbb{I}_{pos}; \\ \mathbb{I}, & \text{if } \mathbb{I}_{pos} < \theta < \mathbb{F}_{pos}; \\ None, & \text{otherwise.} \end{cases}$$

Corpus	EN-DE		EN-ES		EN-IT		EN-RU	
MuST-C (v1.2)	229.7K		265.6K		253.6K		265.5K	
IWSLT-22	0.8K		0.8K		–		–	
Formality-annotated	$\mathbb{F}$	$\mathbb{I}$	$\mathbb{F}$	$\mathbb{I}$	$\mathbb{F}$	$\mathbb{I}$	$\mathbb{F}$	$\mathbb{I}$
INFEREASY	8.6K	8.6K	6.7K	6.7K	36.6K	36.6K	38.3K	38.3K
INFERFULL	13.7K	9.5K	10.5K	4.5K	11.4K	13.5K	12.0K	14.1K
+ZERO SHOT ON EN-{RU,IT}	13.7K	9.5K	10.5K	4.5K	0K	0K	0K	0K
+IWSLT-22	14.1K	9.9K	10.9K	4.9K	11.4K	13.5K	12.0K	14.1K

Table 1: Corpora containing training data used in the constrained track. Values indicate number of sentence pairs after preprocessing.

```
CUDA_VISIBLE_DEVICES 0,1,2,3
-finetune-from-model *
-max-update *
-ddp-backend=legacy_ddp
-task multilingual_translation
-arch multilingual_transformer_iwslt_de_en
-lang-pairs en-de,en-es,en-ru,en-it
-encoder-langtok tgt
-share-encoders
-share-decoder-input-output-embed
-optimizer adam
-adam-betas '(0.9, 0.98)'
-lr 0.0005
-lr-scheduler inverse_sqrt
-warmup-updates 4000
-warmup-init-lr '1e-07'
-label-smoothing 0.1
-criterion label_smoothed_cross_entropy
-dropout 0.3
-weight-decay 0.0001
-save-interval-updates *
-keep-interval-updates 10
-no-epoch-checkpoints
-max-tokens 1000
-update-freq 2
-fp16
```

Table 2: Parameters of `fairseq-train` for pre-training and fine-tuning all models. The starred (*) parameters depend on the track/subtrack and can be found in the paper description or in the implementation.

We condition assignment on both positional lists since common phrases such as (*Yes! – Ja!*) may rank high on both sides, but should not get included in either corpus. We determine θ empirically by selecting a value that yields the most data as a result. These values were selected dynamically for each language pair, and resulted in $\theta = 0.45\mathcal{C}$ for EN-DE and $\theta = 0.5\mathcal{C}$ for EN-ES. We refer to this approach as INFEREASY.

We quickly observed that the selection method needed to take into account the relative ranking of a sentence pair for both formalities. To illustrate this, let $\theta = 50$, the number of sentences $n = 100$; a sentence pair with rankings $\mathbb{F}_{pos} = 49, \mathbb{I}_{pos} = 51$ will get included in the formal corpus, but with $\mathbb{F}_{pos} = 1, \mathbb{I}_{pos} = 50$ it will not, because $\mathbb{I}_{pos}$ is in the top k for the informal set, even though the relative difference between the two positions is large. To amend this, we introduced a classification by *relative position difference*: for any sentence pair with positions $(\mathbb{F}_{pos}, \mathbb{I}_{pos})$ we classify it as formal if $\mathbb{F}_{pos} - \mathbb{I}_{pos} > \alpha$. We determine α empirically: using $0.05\mathcal{C}$ and $0.2\mathcal{C}$ as the lower and upper bound, respectively, for several values α in range we compute a language model from the resulting data and calculate average perplexity $PP(\text{LM}_{Corpus(\alpha)}, \text{IWSLT})$. We select the α value which minimises this perplexity. We refer to this approach as INFERFULL.

2.3 Generalisation for Zero-Shot Language Pairs

For two language pairs (EN-{RU,IT}) no supervised training data was provided, meaning we could only use the IWSLT corpus and our inferred data from EN-{DE,ES} to obtain data for these pairs. We decided to focus on comparisons on the source (EN) side, meaning we could not use the IWSLT corpus as it was paired. One observation we made at this point was that, contrary to intuition, the same source sentences within the MuST-C corpus had different formality expressions in the German and Spanish corpora, respectively.

Let EN-DEXES be a corpus of triplets of sentences $(src_{\text{EN}}, tgt_{\text{DE}}, tgt_{\text{ES}})$ obtained by identifying English sentences which occur in both the EN-DE and EN-ES corpora. Since there are many such sentences in the MuST-C corpus, the EN-DEXES contains 85.72% of sentence pairs from the EN-DE and 74.13% of pairs from the EN-ES corpus. After marking the target sides of the EN-DEXES corpus for formality with INFERFULL, we quantified in how many cases both languages get the same label

(formal of informal), and in how many cases they get a different label (Table 3). Out of all annotated triplets, only 5.8% triplets were annotated in both target languages; this is a significantly smaller fraction than expected. Within that group, almost 60% triplets had matching annotations. This implies that the same English sentence can sometimes (approx. 2 out of 5 times in our case) be expressed with different formality in the target language in the same discourse situation.

EN-DE	EN-ES	Count	% of annotated
$\mathbb{F}$	$\mathbb{F}$	845	2.85%
$\mathbb{I}$	$\mathbb{I}$	233	0.78%
$\mathbb{F}$	$\mathbb{I}$	381	0.95%
$\mathbb{I}$	$\mathbb{F}$	362	1.22%
$\mathbb{F}$	$\emptyset$	10851	36.54%
$\mathbb{I}$	$\emptyset$	7805	26.29%
$\emptyset$	$\mathbb{F}$	6567	22.12%
$\emptyset$	$\mathbb{I}$	2749	9.26%

Table 3: Context combinations for the EN-DEXES triplet extracted from the MuST-C dataset. "$\emptyset$" denotes "no context".

Given the non-zero count of triplets with matching formalities, we make another assumption: namely that the English sentences of the triplets with matching formalities may be of "strictly formal" or "strictly informal" nature, meaning the translations of at least some of those sentences to Russian and Italian may express the same formality. To extract formal and informal sentences for the zero-shot pairs, we adapted the original method, but this time using English as a pivot to convey the formality information. As the in-domain corpus, we used the English sentences whose German and Spanish translations were both labelled as formal or both as informal, respectively (columns $1, 2$ in Table 3). We ranked the EN-RU and EN-IT corpora by their source sentences' similarity to that intersection (using the perplexity difference as before).

To infer the final corpora with the INFERFULL method, we used the α which yielded corpora of similar quantity to the ones for EN-{DE,ES}, since we could not determine that value empirically.

2.4 Relative Frequency Model for Reranking: FORMALITYRERANK

We observed that even when a model gets the formality wrong in its best hypothesis, the correct answer is sometimes found within the *n* best hypotheses, but at a lower position. We hypothesised that by re-ranking the n-best list according to a criterion different from the beam search log probability we could push the hypothesis with the correct formality to the first position.

We performed an oracle experiment with `scorer.py` to obtain an upper bound on what can be gained by re-scoring the n-best list perfectly: we generated k-best hypotheses for $k \in \{1, 5, 10, 20, 30, ..100\}$[2] and from each list of k hypotheses we selected the first hypothesis (if any) which `scorer.py` deemed of correct formality. The results (Table 4) show that as we expand the list of hypotheses, among them we can find more translations of correct formality, up to a .959 average accuracy (+.106 w.r.t. the model) for $k = 100$. The column "# Cases" shows that on average in up to 21 cases a hypothesis of the correct formality could be found with re-ranking. Finally, for any k, selecting the hypotheses with the correct formality (Oracle) in place of the most probable ones does (Model) not decrease translation quality, and may improve it (column "BLEU").

k	Accuracy		δ_{to_best}	# Cases	BLEU	
	Model	Oracle			Model	Oracle
1	.838	.838	0.00	0.00	25.28	25.28
5	.858	.892	1.79	7.00	24.80	24.80
10	.857	.913	2.66	11.50	25.10	25.53
20	.853	.921	3.46	13.75	24.74	25.15
30	.851	.930	5.75	16.00	24.68	25.06
40	.853	.936	7.84	16.75	24.88	25.24
50	.853	.944	9.64	18.25	24.84	25.20
60	.852	.950	11.78	19.75	24.71	25.04
70	.852	.950	12.08	19.75	24.71	25.04
80	.852	.952	12.78	20.25	24.72	25.04
90	.852	.954	13.58	20.50	24.72	25.04
100	.853	.959	14.66	21.25	24.72	25.04

Table 4: Results of the oracle experiment. The used model was *constrained* and trained with the INFERFULL method, provided values are averaged across the development set. δ_{to_best} describes the average distance to the first hypothesis of correct formality for cases where the most probable hypothesis is incorrect. The column "# Cases" quantifies that phenomenon.

To re-rank the hypotheses we built a simple relative frequency model from the IWSLT data. For each term $t_i \in \mathcal{T}$ we calculated its occurrence counts $\mathbb{F}_{count}$ in the *formal* set and $\mathbb{I}_{count}$ in the *informal* set. Let $count(t_i) = \mathbb{F}_{count}(t_i) + \mathbb{I}_{count}(t_i)$. Since we wished to focus on terms differentiating

[2] We capped the search at $k = 100$ due to long inference times for higher k values.

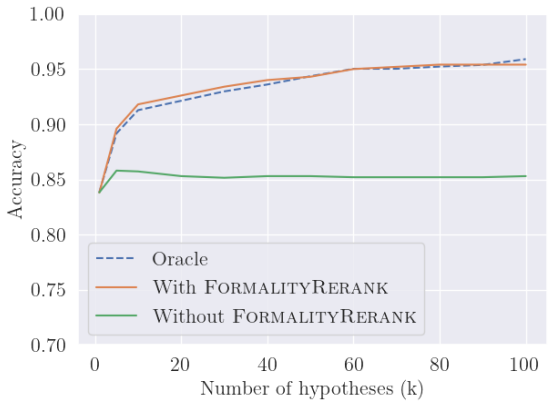

Figure 1: Validation accuracy plot showing the effect of applying FORMALITYRERANK to a list of k model hypotheses.

the two sets, we calculated the count difference ratio and used it as the weight β:

$$\beta(t_i) = \frac{|\mathbb{F}_{count}(t_i) - \mathbb{I}_{count}(t_i)|}{\max_{t_k \in \mathcal{T}} |\mathbb{F}_{count}(t_k) - \mathbb{I}_{count}(t_k)|}$$

We additionally nullified probabilities for terms for which the difference of the number of occurrences in the formal and informal sets was lower than the third of total occurrences:

$$\kappa(t_i) = \begin{cases} 0, & \text{if } \frac{|\mathbb{F}_{count}(t_i) - \mathbb{I}_{count}(t_i)|}{\mathbb{F}_{count}(t_i) + \mathbb{I}_{count}(t_i)} < 0.33^{3}; \\ 1, & \text{otherwise} \end{cases}$$

The probabilities could now be calculated as

$$p(\mathbb{F}|t_i) = \frac{\mathbb{F}_{count}(t_i)}{count(t_i)} * \beta(t_i) * \kappa(t_i)$$
$$p(\mathbb{I}|t_i) = \frac{\mathbb{I}_{count}(t_i)}{count(t_i)} * \beta(t_i) * \kappa(t_i)$$

For a hypothesis Y, a source sentence S and contexts $c, \hat{c} \in \{\mathbb{F}, \mathbb{I}\}, c \neq \hat{c}$, our objective function in translation thus became

$$p(Y|X, c) = p(Y|X) + p(c|Y) - p(\hat{c}|Y)$$

where

$$p(c|Y) = \sum_i p(c|y_i)$$

Figure 1 shows how validation accuracy increases when this method is used, and that the model is now able to match the oracle accuracy for nearly every k. For $k = 100$ the average improvement in accuracy is .102. The effect of model's accuracy sometimes surpassing the oracle accuracy (e.g. for $k = 30$) is a by-product of slight sample size variations: the evaluation script `scorer.py` depends on phrase matches, and a sample is only counted for evaluation if a hypothesis has at least one phrase match against the formality-annotated reference.

2.5 Model Selection: BESTACCAVERAGING

We fine-tuned each model for 100K iterations on the MuST-C corpus with formality tags appended to relevant sentences. We then evaluated every checkpoint (saved each epoch) with `scorer.py` on IWSLT data. Our initial approach to selecting a model assumed averaging the last 10 checkpoints from training. We experimented with an alternative method to finding which checkpoints to average: we first computed the accuracy on the IWSLT dataset for each checkpoint, and then selected a window of 10 consecutive checkpoints with the highest average accuracy (BESTACCAVERAGING).

2.6 Development Results

We report the validation results in Table 5. The first result we observed was that in both language pairs the pre-trained model (a strong baseline) learned a **dominant** formality: formal for EN-DE (.853 accuracy to .147) and informal for EN-ES (.632 accuracy to .368).

We observed that both methods (INFEREASY and INFERFULL) yield consistently better accuracy for dominant formalities than non-dominant ones. Nevertheless, with INFERFULL we obtain an average $+.474$ accuracy points over the baseline for non-dominant formalities; INFEREASY fails to learn meaningful control for non-dominant formalities. Based on these results we focused out later efforts on INFERFULL alone.

Continuing with INFERFULL, we noticed a significant improvement of up to $+.223$ accuracy points for (EN-DE, $\mathbb{I}$) when using FORMALITYRERANK on top of standard beam search ($k = 100$) without impacting the translation quality. Finally, BESTACCAVERAGING helped bring the average accuracy score up to .961 without impacting translation quality.

2.7 Submitted Models

Based on the validation results, we submitted two models to the constrained track: to the *full supervision* subtrack, we submitted the INFERFULL model with FORMALITYRERANK ($k = 100$) and

	MuST-C (BLEU)				IWSLT (Accuracy)				
	EN-DE	EN-ES	EN-RU	EN-IT	EN-DE		EN-ES		Mean
					$\mathbb{F}$	$\mathbb{I}$	$\mathbb{F}$	$\mathbb{I}$	
Pre-trained	**30.7**	39.7	19.5	**31.3**	.853	.147	.368	.632	.500
INFEREASY	30.1	39.3	19.9	31.1	.967	.167	.376	.595	.526
INFERFULL	30.1	**39.8**	19.8	31.2	.978	.637	.854	.963	.858
+FORMALITYRERANK	30.1	**39.8**	19.8	31.2	**1.000**	.860	**.968**	**.990**	.955
+BESTACCAVERAGING	30.3	39.6	**20.0**	31.2	**1.000**	**.899**	.956	**.990**	**.961**

Table 5: Results on the **development** sets for models built within the constrained track.

BESTACCAVERAGING upgrades; for the *zero-shot* subtrack, we fine-tuned an alternative version of the model where we skipped the EN-{RU,IT} fine-tuning data, effectively making inference for these zero-shot pairs[4]. We used the same augments as in *full supervision*.

3 Unconstrained Track

Our submission for the unconstrained track largely copies the constrained track one, but is applied to a larger training corpus.

3.1 Data Collection and Preprocessing

We collect all datasets permitted by the organisers for our selected language pairs, including:

- **MuST-C (v1.2)** (Di Gangi et al., 2019),
- **Paracrawl (v9)** (Bañón et al., 2020),
- **WMT Corpora** (from the News Translation task) (Barrault et al., 2021):
  - **NewsCommentary (v16)** (Tiedemann, 2012),
  - **CommonCrawl** (Smith et al., 2013),
  - **WikiMatrix** (Schwenk et al., 2021),
  - **WikiTitles (v3)** (Barrault et al., 2020),
  - **Europarl (v7, v10)** (Koehn, 2005),
  - **UN (v1)** (Ziemski et al., 2016),
  - **Tilde Rapid** (Rozis and Skadiņš, 2017),
  - **Yandex**[5].

We list data quantities as well as availability for all language pairs in Table 6. We preprocessed the WMT and Paracrawl corpora: for both we first ran a simple rule-based heuristic of removing sentence pairs with sentences longer than 250 tokens, and with a source-target ratio greater than 1.5; removing non-ASCII characters on the English side, pruning some problematic sentences (e.g. links). We normalised punctuation using the script from Moses (Koehn et al., 2007). We removed cases where either sentence is empty or where the source is the same as the target. Finally, we asserted that the case (lower/upper) of the first characters must be the same between source and target and that if either sentence ends in a punctuation mark, its counterpart must end in the same one. As the last step, we removed identical and very similar sentence pairs.

After the initial preprocessing, we ran the *Bi-Cleaner* tool (Ramírez-Sánchez et al., 2020) on each corpus; the algorithm assigns a confidence score $\in [0, 1]$ to each pair, measuring whether the sentences are good translations of each other, effectively removing potentially noisy sentences. We removed all sentence pairs from the corpora which scored below 0.7 confidence. The final training data quantities are reported in Table 6.

3.2 Data Labelling

Before we applied the same method to obtain fine-tuning data for the unconstrained track, we observed that many sentence pairs in this corpus are not dialogue, and hence useless for fine-tuning. As the first step, we used the original perplexity-based re-ranking algorithm to prune the unconstrained corpus. We used the MuST-C corpus as in-domain and all the unconstrained data as out-of-domain. We truncated the unconstrained set to the top 5M sentences most like the MuST-C data. We then applied INFERFULL with α threshold adapted to the data volume. The resulting data quantities can be found in the last row of Table 6.

[4]We labelled a small random sample of training data with a random formality tag so the model learned to recognise the symbol as part of the input.

[5]https://translate.yandex.ru/corpus?lang=en, accessed 4 Apr 2022.

Corpus	EN-DE		EN-ES		EN-IT		EN-RU	
MuST-C (v1.2)	0.23M		0.27M		0.25M		0.27M	
Paracrawl (v9)	278.31M		269.39M		96.98M		5.38M	
NewsCommentary v16	0.40M		0.38M		0.09M		0.34M	
CommonCrawl	2.40M		1.85M		–		0.88M	
WikiMatrix	5.47M		–		–		3.78M	
WikiTitles (v3)	1.47M		–		–		1.19M	
Europarl (v7|v10)	1.83M		1.97M		1.91M		–	
UN (v1)	–		11.20M		–		–	
Tilde Rapid	1.03M		–		–		–	
Yandex	–		–		–		1M	
Total								
Raw	291.14M		285.06M		99.23M		12.84M	
Preprocessed	76.99M		91.29M		36.99M		3.86M	
Formality-annotated	$\mathbb{F}$	$\mathbb{I}$	$\mathbb{F}$	$\mathbb{I}$	$\mathbb{F}$	$\mathbb{I}$	$\mathbb{F}$	$\mathbb{I}$
	216.5K	187.2K	111.8K	129.7K	101.0K	172.0K	195.9K	218.4K

Table 6: Corpora containing training data used in the unconstrained experiments. Values indicate number of sentence pairs after preprocessing.

3.3 Pre-training and Fine-tuning

We used an identical model architecture to the one from the constrained track but extended the training time: we pre-trained for 1.5M iterations (approx. 1.5 epochs) and fine-tuned for 0.25M iterations (approx. 47 epochs). For fine-tuning, we used the MuST-C corpus (to maintain high translation quality) concatenated with the inferred formality-annotated data (to learn formality control). We applied FORMALITYRERANK with $k = 50$, but not BESTACCAVERAGING as we found that the differences in average accuracy for most checkpoints is minimal (and nears 100); instead, we averaged the last 10 checkpoints.

3.4 Development Results

The development results (Table 7) surpassed those achieved in the constrained track, presumably thanks to richer corpora extracted for both formalities. INFERFULL yielded near-perfect accuracy for all sets but (EN-DE, $\mathbb{I}$), and applying FORMALITYRERANK effectively brought all scores up to a mean accuracy of .999. Our pre-trained model for this track achieved lower BLEU scores than for the constrained track, which is explained by the test set coming from the same domain as the constrained training data.

3.5 Submitted model

Similarly to the constrained track, we submit two models to the unconstrained track: to the *full supervision* subtrack, we submit the INFERFULL model with FORMALITYRERANK ($k = 50$); for the *zero-shot* subtrack, we fine-tune an alternative version of that in which we skip the EN-{RU,IT} fine-tuning data, effectively making inference for these pairs zero shot.

4 Final Results

We report the final evaluation results in Table 8 (translation quality) and Table 9 (formality control). In the latter we also provide the performance of our baseline (pre-trained) model for reference.

Within the constrained track, we achieved near-ideal accuracy for the dominant formality for each language pair (between .961 and 1.000) with the supervised model. Scores for non-dominant formalities are weaker but still impressive for EN-{DE,ES} with an average of .880. Our best model for EN-{RU,IT} improved by .193 accuracy points over the baseline. The models submitted to the unconstrained track again achieved an impressive average accuracy of .992 for dominant formality; additionally, performance for non-dominant formality in EN-{DE,ES} improved significantly w.r.t. the constrained model, also averaging .992. This means that with enough training data our methods were capable of matching the performance on a minority class w.r.t. a majority class.

Finally, contrary to the constrained track, the *unconstrained-zero-shot* model achieved the best accuracy for zero-shot pairs, to an average of .659.

	MuST-C (BLEU)				IWSLT (Accuracy)				
					EN-DE		EN-ES		Mean
	EN-DE	EN-ES	EN-RU	EN-IT	$\mathbb{F}$	$\mathbb{I}$	$\mathbb{F}$	$\mathbb{I}$	
Pre-trained	28.9	39.5	18.5	29.3	.634	.366	.215	.785	.500
INFERFULL	**32.3**	**40.8**	**20.4**	**32.0**	.990	**1.000**	.952	.991	.983
+FORMALITYRERANK	**32.3**	**40.8**	**20.4**	**32.0**	**1.000**	**1.000**	**.995**	**1.000**	**.999**

Table 7: Results on the **development** sets for models built within the unconstrained track.

Model name	BLEU				COMET			
	EN-DE	EN-ES	EN-RU	EN-IT	EN-DE	EN-ES	EN-RU	EN-IT
constrained-supervised (1)	31.50	36.53	21.41	33.28	.4477	.6076	.3311	.5676
constrained-zero-shot (2)	31.25	36.65	21.43	33.15	.4368	.6108	.3298	.5525
unconstrained-supervised (3)	32.50	36.98	22.01	33.56	.4972	.6349	.3846	.5927
unconstrained-zero-shot (4)	32.47	36.83	21.45	33.12	.4851	.6209	.3565	.5623

Table 8: Translation quality results on the **test** sets for all submitted models. Numbers in brackets indicate number of model submitted.

Model name	EN-DE		EN-ES		EN-RU		EN-IT	
	$\mathbb{F}$	$\mathbb{I}$	$\mathbb{F}$	$\mathbb{I}$	$\mathbb{F}$	$\mathbb{I}$	$\mathbb{F}$	$\mathbb{I}$
constrained-pre-trained	.885	.115	.457	.543	.951	.049	.149	.851
constrained-supervised (1)	1.000	.886	.874	.980	.981	.234	.349	.961
constrained-zero-shot (2)	–	–	–	–	.981	.154	.294	.929
unconstrained-pre-trained	.745	.255	.323	.677	.964	.036	.052	.948
unconstrained-supervised (3)	1.000	1.000	.981	1.000	.992	.136	.188	.980
unconstrained-zero-shot (4)	–	–	–	–	.995	.142	.512	.986

Table 9: Accuracy results on the **test** data as measured by `scorer.py`.

5 Conclusions

Overall results suggest that it is easy for a pre-trained translation model to learn controlled expression of the dominant type within a dichotomous phenomenon while learning to render the less-expressed type is significantly harder, especially in a low-resource scenario. Our methods applied to the supervised language pairs (English-to-German, English-to-Spanish) worked near unfailingly, but using English as a pivot language to propagate formality information did not help achieve similar results for the zero-shot pairs.

We suspect that the significant accuracy gains from FORMALITYRERANKING may have been partially due to formality in the studied language pairs itself being expressed primarily via certain token words such as the honorific *Sie* in German creating a *pivot* effect (Fu et al., 2019). As such, it may be of interest for future research to study such methods applied to more complex phenomena, such as grammatical expression of gender.

Finally, results for the EN-{RU,IT} language pairs may not have been as good as expected because we used the inferred data from the constrained track to build the relative frequency model, but the inferred data turned out to be not as high quality as we expected. Future work may investigate a robust solution to this problem of propagating formality via a source (pivot) language to extract training data for other language pairs.

Code used for our implementation can be accessed at https://github.com/st-vincent1/iwslt_formality_slt_cdt_uos/.

Acknowledgements

This work was supported by the Centre for Doctoral Training in Speech and Language Technologies (SLT) and their Applications funded by UK Research and Innovation [grant number EP/S023062/1].

References

Antonios Anastasopoulos, Luisa Bentivogli, Marcely Z Boito, Ondřej Bojar, Roldano Cattoni, Anna Currey, Georgiana Dinu, Kevin Duh, Maha Elbayad, Marcello Federico, Christian Federmann, Hongyu Gong, Roman Grundkiewicz, Barry Haddow, Benjamin Hsu, Dávid Javorský, Věra Kloudová, Surafel M Lakew, Xutai Ma, Prashant Mathur, Paul McNamee, Kenton Murray, Maria N\uadejde, Satoshi Nakamura, Matteo Negri, Jan Niehues, Xing Niu, Juan Pino, Elizabeth Salesky, Jiatong Shi, Sebastian Stüker, Katsuhito Sudoh, Marco Turchi, Yogesh Virkar, Alex Waibel, Changhan Wang, and Shinji Watanabe. 2022. FINDINGS OF THE IWSLT 2022 EVALUATION CAMPAIGN. In *Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT 2022)*, Dublin, Ireland. 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.

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.

Loic Barrault, Ondrej Bojar, Fethi Bougares, Rajen Chatterjee, Marta R. Costa-jussa, Christian Federmann, Mark Fishel, Alexander Fraser, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Paco Guzman, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Tom Kocmi, Andre Martins, Makoto Morishita, and Christof Monz, editors. 2021. *Proceedings of the Sixth Conference on Machine Translation*. Association for Computational Linguistics, Online.

Mattia A. Di Gangi, Roldano Cattoni, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2019. MuST-C: a Multilingual Speech Translation Corpus. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2012–2017, Minneapolis, Minnesota. Association for Computational Linguistics.

Yao Fu, Hao Zhou, Jiaze Chen, and Lei Li. 2019. Rethinking text attribute transfer: A lexical analysis. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 24–33, Tokyo, Japan. 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.

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2017. Google's Multilingual Neural Machine Translation System: Enabling Zero-Shot Translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Philipp Koehn. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. In *Conference Proceedings: the tenth Machine Translation Summit*, pages 79–86, Phuket, Thailand. AAMT, AAMT.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic. Association for Computational Linguistics.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Surafel Melaku Lakew, Mattia Di Gangi, and Marcello Federico. 2019. Controlling the Output Length of Neural Machine Translation. *arXiv*.

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*, Seattle, USA. Association for Computational Linguistics.

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

pages 48–53, Minneapolis, Minnesota. Association for Computational Linguistics.

Gema Ramírez-Sánchez, Jaume Zaragoza-Bernabeu, Marta Bañón, and Sergio Ortiz Rojas. 2020. Bifixer and bicleaner: two open-source tools to clean your parallel data. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 291–298, Lisboa, Portugal. European Association for Machine Translation.

Roberts Rozis and Raivis Skadiņš. 2017. Tilde MODEL - multilingual open data for EU languages. In *Proceedings of the 21st Nordic Conference on Computational Linguistics*, pages 263–265, Gothenburg, Sweden. Association for Computational Linguistics.

Holger Schwenk, Vishrav Chaudhary, Shuo Sun, Hongyu Gong, and Francisco Guzmán. 2021. 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.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Controlling politeness in neural machine translation via side constraints. *2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL HLT 2016 - Proceedings of the Conference*, pages 35–40.

Jason R. Smith, Herve Saint-Amand, Magdalena Plamada, Philipp Koehn, Chris Callison-Burch, and Adam Lopez. 2013. Dirt cheap web-scale parallel text from the Common Crawl. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1374–1383, Sofia, Bulgaria. Association for Computational Linguistics.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218, Istanbul, Turkey. European Language Resources Association (ELRA).

Eva Vanmassenhove, Christian Hardmeier, and Andy Way. 2018. Getting gender right in neural machine translation. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, EMNLP 2018*, pages 3003–3008.

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).