Procesamiento del Lenguaje Natural, Revista nº 70, marzo de 2023, pp. 199-212 recibido 16-12-2022 revisado 01-02-2023 aceptado 11-02-2023

Exploring politeness control in NMT: fine-tuned vs. multi-register models in Castilian Spanish

Estudio de la cortesía en traducción automática neuronal: modelos ajustados y modelos multirregistro para el castellano

Celia Soler Uguet[1] **Nora Aranberri**[2]
[1]University of the Basque Country (UPV/EHU)
[2]HiTZ Center - Ixa, University of the Basque Country (UPV/EHU)
csoler003@ikasle.ehu.eus
nora.aranberri@ehu.eus

Abstract: Nowadays neural machine translation can generate high quality translations with regard to grammatical accuracy and fluency. Therefore, it is time to broaden research efforts to consider aspects of language that go beyond the mentioned attributes to keep pushing the limits of the technology. In this work, we focus on politeness. Specifically, we adapt and explore, for Castilian Spanish, two different domain-adaptation approaches: fine-tuning and multilingual models. Results from automatic and manual evaluations seem to indicate that the latter might be a better solution to strike a quality balance between all registers (formal, informal, and neutral). Fine-tuning a baseline system for each specific register seems to suffer from a degree of catastrophic forgetting, which leads to a worse overall performance of the engines.
Keywords: neural machine translation, politeness, fine-tuning models, multi-register models.

Resumen: En la actualidad, la traducción automática neuronal es capaz de generar traducciones de alta calidad en lo que respecta a la precisión gramatical y la fluidez. Así, es hora de ampliar los objetivos de investigación y considerar aspectos de la lengua que van más allá de los atributos mencionados para seguir superando los límites de la tecnología. En este trabajo, nos centramos en la cortesía. En concreto, adaptamos y exploramos, para el castellano, dos enfoques diferentes de adaptación al dominio: modelos ajustados y modelos multilingües. Los resultados de las evaluaciones automáticas y manuales parecen indicar que el segundo podría ser mejor para lograr un equilibrio de calidad entre todos los registros (formal, informal y neutro). El ajuste de modelos parece sufrir de olvido catastrófico, lo que conduce a un peor rendimiento general de los motores.
Palabras clave: traducción automática neuronal, cortesía, modelos ajustados, modelos multirregistro.

1 *Introduction*

As Vanmassenhove, Shterionov, and Gwilliam (2021) suggest, now that Neural Machine translation (NMT) systems have reportedly reached a quality that is close to that of human translations, it is time to start paying attention to aspects of language that go beyond grammatical accuracy such as discourse phenomena. In this line, one such phenomenon is the level of politeness. Deviations from what is expected in its use can give rise to misunderstandings in communication, and although this might seem like a petty problem, it can become extremely critical for certain cultures and communicative situations (Haugh, 2005).

Now, what is politeness? Let us start by defining register. Register was described by Matthiessen and Halliday (1997) as the context of a situation in a speech act, which consists of three dimensions: field, mode and tenor. The field refers to the area in which

ISSN 1135-5948 DOI 10.26342/2023-70-16

the linguistic activity is operating (specialized vs. non-specialized discourse); the mode has to do with the means in which communication is taking place (written vs. oral); and the tenor denotes the relationship between the speakers (relatives vs. workmates) (Halliday, McIntosh, and Stevens, 1964). In this scenario, politeness presents itself as one of the aspects that comprise the tenor, described by Brown (2015, 1) as "a matter of taking into account the feelings of others as to how they should be interactionally treated, including behaving in a manner that demonstrates appropriate concern for the interactors' social status and their social relationship". Therefore, we can argue that politeness is one of the many aspects that NMT needs to address in order to adequately respond to a specific register.

In this work, we explore ways to control the level of politeness in NMT for an English to Spanish system. Specifically, we focus on Castilian, the Spanish variety spoken in Spain, where, as a general rule, different personal pronouns are used to address an interlocutor depending on the intended level of politeness: *tú* tends to be the form used in situations where interlocutors are (relatively) close, while *usted* tends to be the form used to show respect and distance. We explore two domain-adaption techniques, namely, a fine-tuning approach following research by Chu and Wang (2018) and a multi-register approach following Sennrich, Haddow, and Birch (2016a), by adapting their setups to address the new language.

Results seem to indicate that a multi-register system trained in three directions (formal, informal and neutral) using a mix of in-domain and out-of-domain data achieves better average scores when taking into account the three directions according to both automatic and human evaluations. This approach seems to slightly outperform the fine-tuning approach, which seems to suffer from catastrophic forgetting.

The remainder of this paper is organised as follows: Section 2 presents the related work on addressing politeness in NMT; Section 3 describes the experimental setup of our study; Section 4 reports the results obtained; and finally, Section 5 draws the main conclusions and presents some avenues for further analysis on the topic.

2 *Related work*

Domain-adaptation is a fairly researched area in MT, as general purpose systems usually perform poorly and systems geared towards specific domains are in high demand (Koehn and Knowles, 2017). One of the main approaches used in this area is the fine-tuning of a baseline system (Kell, 2018). In NMT, it involves leveraging out-of-domain corpora to improve in-domain translations (Kirkpatrick et al., 2017), and it has been implemented successfully in various works (Luong and Manning, 2015; Etchegoyhen et al., 2018).

Alternatively, some recent research has proposed strategies to guide and control NMT output, for example, translation memory guided neural fuzzy repair (Bulte and Tezcan, 2019), domain control using side constraints such as tag-tokens and word features (Kobus, Crego, and Senellart, 2017), terminology constraints (Dinu et al., 2019), or constrained decoding (Post and Vilar, 2018).

However, to the best of our knowledge, register-related work, in general, and politeness, in particular, have received very little attention so far. In fact, we found that, to date, experiments have only been carried out with one main approach for the linguistic phenomenon at hand, namely, the application of a multilingual model, proposed by Sennrich, Haddow, and Birch (2016a) and later recreated by Feely, Hasler, and de Gispert (2019) to address politeness in German and Japanese, respectively.

In the following lines, we describe the two approaches, fine-tuning and multilingual models, in more detail.

2.1 Fine-tuning approach

Fine-tuning is considered model centric, or more precisely, training-objective centric according to the classification by Chu and Wang (2018). Here, an NMT system is trained on a resource-rich out-of-domain corpus until convergence, and then its parameters are fine-tuned on a resource-poor, in-domain corpus. A good number of positive results have been reported in the literature. For example, Luong and Manning (2015) adapted a baseline system to spoken language by further training an existing model based on formal texts (provided at WMT 2015) for 12 epochs using a smaller set of spoken text (provided at IWSLT 2015) in which after the first epoch, learning rates (initially set to 1.0) are halved every two epochs. They reported an improvement in BLEU of almost four points.

If we were to adapt this approach to tackle politeness, the baseline system could be trained with generic data, while data for specific politeness levels could be used to develop as many fine-tuned models as necessary. Yet, it is worth mentioning that, apart from the high maintenance requirements (Bapna, Arivazhagan, and Firat, 2019), one of the main drawbacks of these systems is what is called catastrophic forgetting. This is a phenomenon whereby a model that has been trained on task A and then retrained on task B forgets much of what it originally learned on task A (Kell, 2018). Yet, as Kell (2018) outlines, different approaches have been proposed for tackling this problem, such as combining multi-domain and fine-tuning methods or using regularization techniques such as elastic weight consolidation.

2.2 Multi-register approach

The *multi-register* approach was first introduced by Sennrich, Haddow, and Birch (2016a) and has then been used for other tasks such as multilingual NMT (Aharoni, Johnson, and Firat, 2019). This method uses the placement of tags in the training data to help the decoder at translation time. Instead of applying changes to a model architecture from a standard NMT system, it introduces an artificial token at the beginning of the input sentence to specify the required target language. Sennrich, Haddow, and Birch (2016a) performed English>German experiments on OpenSubtitles (Tiedemann, 2012), a parallel corpus of movie subtitles. They trained an attentional encoder-decoder NMT system using Groundhog[1] (Van Merriënboer et al., 2015) and used a joint BPE to represent the texts with a fixed vocabulary of subword units with size 90,000.

In their research the authors proved that it is possible to control the honorifics produced at test time by marking up the source side of the training data with a feature that encodes the use of honorifics on the target side. To automatically annotate politeness on a sentence level, they made use of rules based on the morphosyntactic annotation by ParZu (Sennrich, Volk, and Schneider, 2013), marking each instance as either being informal, formal or neutral (if none of the other two applied). Interestingly, to ensure that the engine learned to not overproduce honorifics when no side constraint was provided, they only marked a subset of the training instances with a politeness feature and set the probability that an instance was marked to 0.5.

They tested translations without side constraints (neutral) and with constrains (polite and informal), achieving 20.7, 17.9 and 20.2 BLEU points respectively. In another oracle experiment, they used the politeness label of the reference to determine the side constraint, which simulates a setting in which a user controls the desired politeness. In that case, BLEU was strongly affected by the choice in politeness: results showed an improvement of 3.2 BLEU points over the baseline.

3 Experimental setup

In this section we describe the steps taken to train our politeness-aware systems for the English-Castilian language combination. Firstly, we present the procedure followed to select, process and divide the data set according to the different levels of politeness. Secondly, we introduce the features of the fine-tuned and multilingual NMT models used for the experiment.

3.1 Data set

There is no bilingual data annotated according to its level of politeness for the English>Spanish language pair that can be used to train an NMT system. Therefore, our first task involved creating a set with those characteristics. We opted for the OpenSubtitles corpus (Tiedemann, 2012) and followed an automatic classification approach to divide it into the required subsets. OpenSubtitles consists of a parallel collection of user contributed subtitles of films and TV programs in various languages. The English-Spanish subset accounts for 46 million parallel segments. It must be noted that the alignment is not always correct but, most importantly for our experiment, the texts are not identified by diatopic varieties. This means that the bilingual corpus might contain instances from several dialects of the Spanish language, which use honorifics differently. In particular, in contrast to Castilian, in a number of Latin American countries the form

[1] github.com/sebastien-j/LV_groundhog

usted is used for familiar situations. Yet, we believed that the advantages of this corpus (mainly the orality, which results in the frequent use of second person pronouns) outweighed such disadvantage and turned it into an interesting case study.

Let us remember that one way of marking politeness in Spanish is by using different honorifics.[2] In Castilian, the personal pronoun *tú* tends to be the form used in situations where interlocutors are (relatively) close, while *usted* tends to be the form used to show respect and distance. Given that this is similar to how German works, we adapted the classification approach by Sennrich, Haddow, and Birch (2016b) to allocate segments into three register subsets: informal, formal and neutral (cases with no second person pronouns or verbs).

This involved a two-step exercise. Firstly, we searched for occurrences of lexical forms that belong to the paradigms of *tú* and *usted* using regex.[3] However, Spanish is a predominantly pro-drop language, that is, pronouns can be omitted if their information can be inferred pragmatically or grammatically. If we only use sentences with overt pronouns to train or fine-tune the formal and informal engines, the language produced could sound quite unnatural, since such engines might over-generate pronouns. To counterbalance this behaviour, we also identified grammatical forms, in particular, verbs, in segments with no overt lexical forms.

Remember that, in Castilian Spanish, the informal pronoun *tú* requires the verb to be conjugated with the mark for the second person singular, while the formal pronoun *usted* requires the verb to be conjugated with the mark for the third person singular. As a result, using Spacy[4], if the Spanish sentence contained a verb conjugated in the second person, we classified it as *informal*; if it contained a verb conjugated in the third person, we classified it as *formal*, and if there was no verb or there was a verb conjugated using a different person, we classified it as *neutral*. The challenge here lies in that the third person forms are ambiguous: they can belong to either *usted* or to the regular third person pronouns *él*, *ella*, *ellos* or *ellas*. The same happened with other lexical forms such as possessives *su*, *suyo*, *suya*, etc. Because Spacy could not disambiguate these cases efficiently, to classify these correctly, we searched for *you*, *your* or *yours* in the parallel source segment to identify second-person cases (Sennrich, Haddow, and Birch, 2016b).

We checked the accuracy of our approach by analysing a random set of 100 instances from each of the subsets (50 extracted using the regex approach, and 50 extracted by parsing). The majority of the segments was correctly classified, with an accuracy of 99%, 76% and 93% for the informal, formal and neutral subsets, respectively.

During a qualitative analysis of the results, we observed that to a large extent, the incorrect instances were due to errors in the disambiguation of third person verbs, the misalignment of the English *you*, originally misaligned source and target segments and segments of dubious quality. Solving the first two cases would require implementing a more complex disambiguation process and were not modified. After all, we expected that the amount of false positives in the formal corpus would not hurt the performance of our engines to a great extent, and if so, it could also shed some light on our study when comparing the different engines. However, for the problem with segments of dubious quality, we filtered our data using Marian's scorer[5] (Junczys-Dowmunt et al., 2018) following the advice of Bane and Zaretskaya (2021). The scorer calculates negative log likelihood of a segment with respect to a model. We used the Helsinki−NLP EN>ES model[6] Tiedemann and Thottingal (2020) and filtered our data with a threshold of -6.5, which reduced the data sets around 20% (see Table 1 for the distribution of register classes of the corpus).[7]

Not surprisingly, the number of segments

[2] In his study on registers, Briz (2010) gives a definition of what he denotes as the prototype of colloquial and formal registers. Among their characteristics, he mentions the use of an informal or a formal tone, and refers to politeness as one of the several features that conform register.

[3] Informal lexical forms: *tú, tu, tus, contigo, tuyo, tuyos, tuya, tuyas, ti, te, vosotros, vosotras, vuestro, vuestros, vuestras vuestros*; formal lexical forms: *usted, ustedes, le, les, su, sus, se, suyo, suyos, suya, suyas.*

[4] https://spacy.io

[5] https://marian-nmt.github.io

[6] https://github.com/Helsinki-NLP/Opus-MT

[7] The politeness-specific corpus is open-source and can be freely downloaded from github.com/c-soler-u/exploring_politeness_control

formal subset	1,821,381
informal subset	4,453,708
neutral subset	3,670,602

Table 1: Distribution of the corpus segments across register subsets after full processing.

allocated to each subset is different. Note, however, that a randomly selected even part of each subset was used for training, thus eliminating such unbalances.

3.2 NMT systems

We explored two domain-adaptation approaches to manage politeness in NMT: a fine-tuning approach (FTA) and a multilingual –or multi-register– approach (MRA). We used the Fairseq toolkit[8] (Ott et al., 2019) to train the NMT systems for both approaches. For tokenization and byte-per-encoding (BPE) segmentation, we used Moses[9] and Subword-NMT[10] (Sennrich, Haddow, and Birch, 2016b).

Fine-tuning approach

For the FTA, we first trained a baseline model using 3 million segments containing a balanced mix of formal, informal and neutral subsets (e.g. 1 million segments of each distribution). We trained a joint BPE vocabulary of size 32,000 and applied it to the training data. We used separate vocabularies created with Fairseq and trained a system based on the Transformer architecture (Vaswani et al., 2017) using Adam as an optimizer, a learning rate of 5e-4, dropout of 0.3, label-smoothing of 0.1 and 50 epochs. Our engine was trained with an early-stopping of 5 validation runs.

We then used 700,000 segments from the formal subset and 700,000 segments from the informal subset to fine-tune the baseline system towards these two directions using the last training epoch (see Table 2 for final segment configuration).

For the fine-tuned systems, we reused the BPE code from the baseline engine, but following Subword-NMT best practices (Sennrich, Haddow, and Birch, 2016b), we extracted the vocabulary for each register and passed it along when applying the BPE with a vocabulary threshold of 50 so that the script would only produce symbols which also appeared in the vocabulary. According to the authors, learning BPE on the concatenation of the involved languages increases the consistency of segmentation, and reduces the problem of inserting/deleting characters when copying/transliterating names. Moreover, applying a vocabulary to this would prevent words from being segmented in a way that was seen only in the other language (or register in our case). We used the parameters of the baseline system for the fine-tuned systems, which are trained for 10 epochs with early stopping of 2 validation runs reusing the separate vocabularies that were created for the baseline.

Multi-register approach

For the MRA approach we trained two engines. The first followed the work by Sennrich, Haddow, and Birch (2016a) where a portion of segments from the other registers was added to each subset to avoid excessive bias towards the trained register (MRA-noise). The second was treated as a multilingual system where the three different registers replaced the usual languages (MRA-nonoise), which allowed us to check if the bias was effectively meaningful for our task. To signal the politeness on the target language, the authors prepend a token to each segment. However, for our research, we made use of Fairseq's implementation to train a multilingual system, which dealt with this process automatically.

We trained the MRAnonoise engine using 1.5 million segments from each register subset amounting to a total of 4.5 million segments (see Table 3). We trained a joint BPE code using the three directions and applied it as for the FTA systems, using separate vocabularies. The English vocabulary was trained using the English source data from all three subsets, while the vocabulary for each respective direction was extracted from their particular training-data. We used the Transformer architecture for multilingual translation from Fairseq and applied the same parameters as the previous model but with shared encoder-embeddings: Adam optimizer, learning rate of 5e-4, label-smoothing of 0.1 and dropout of 0.3. We trained the model for 50 epochs with early stopping of 5.

Starting from the data sets that were used to train the MRAnonoise engine (each set containing 1.5 million parallel segments as shown in Table 3), we trained the MRA-

[8] https://github.com/pytorch/fairseq

[9] https://github.com/moses-smt/mosesdecoder

[10] https://github.com/rsennrich/subword-nmt

Baseline system	**Fine-tuned systems**		
Training set	**Training set**	**Development set**	**Test set**
3,000,000	696,000	2,000	2,000

Table 2: Number of bilingual segments used for the FTA systems.

Politeness level	**Training set**	**Development set**	**Test set**
informal	1,498,600	700	700
formal	1,498,600	700	700
neutral	1,498,600	700	700
Total	4,495,800	2,100	2,100

Table 3: Number of bilingual segments used for the MRAnonoise system.

	Informal direction	**Formal direction**	**Neutral direction**
informal segments	750,000	0	750,000
formal segments	0	1,000,000	750,000
neutral segments	325,000	75,000	750,000
Total segments	1,075,000	1,075,000	2,250,000

Table 4: Number of bilingual segments used for the MRAnoise system.

noise by redistributing portions of the sentences following Sennrich, Haddow, and Birch (2016a) where, in order to reduce bias, the probability of an instance pertaining to either the formal or informal subset is marked to 0.5 (note that we did not re-marked it for each epoch of training) (see Table 4 for data size).

As it can be observed in Table 4, around half of the informal and formal training sets were used for their respective registers, while the other half were added to the neutral register. We also set aside 0.70 million segments from the neutral training set and divided them between the informal and formal sets (0.35 million each). However, to compensate for the higher level of noise in the formal set (see Section 3.1), we reduced this amount in its training data. The BPE code and vocabularies, and the training was carried out as following the same steps used in the MRAnonoise engine.

4 Results

In this section we report the results from the evaluation of each approach. We start by providing the score for a number of automatic metrics to test the overall quality of the systems (Section 4.1). Then, we describe the process and insights gathered from human assessments (Section 4.2).

When generating the translations that are used for testing, we use the last checkpoint from each engine with a beam search of 5 and batch size of 128.

4.1 Automatic evaluation

We carried out a two-fold automatic analysis, that is, we used a specific test set for each register of each engine (e.g. 9 specific test sets in total), as well as a common test set to all the engines. The first intends to test each system on a subset of the specific data distribution collected for their development (set aside prior to training), while the second aims at testing the relative performance of the engines. In order to compile the common set and find a balance between the varying data distributions of the engines, we extracted 200 segments from each of the following specific test sets: 600 segments from the FTA test set (200 from each the informal, the formal and the baseline test sets), and 600 from the MRA test set (200 from each the informal, the formal and the neutral test sets). Therefore, the final common test set contains 1,200 segments.

We obtained the automatic metric scores using MT-Telescope (Rei et al., 2021a) and report results for COMETINHO (Rei et al., 2021b), sacreBLEU (Post, 2018) and chr-F (Popović, 2015). For the common test set, we also perform significance testing using t-tests with bootstrap re-sampling (Koehn, 2004) with default parameters (re-samples of 0.5 and 300 iterations).

For the engine-specific test sets, results show solid +30 BLEU points for all direc-

System	sacreBLEU	COMETINHO	chr-F
FTA baseline	35.3	38.9	56.8
FTA informal	39.7	47.5	58.7
FTA formal	35.0	37.3	56.9
MRAnonoise neutral	36.8	38.6	58.0
MRAnonoise informal	40.3	46.6	59.5
MRAnonoise formal	38.4	42.2	59.3
MRAnoise neutral	30.3	25.5	53.0
MRAnoise informal	32.8	30.0	54.1
MRAnoise formal	31.8	27.8	55.1

Table 5: Automatic metric scores for all systems on the specific test sets.

System	sacreBLEU	COMETINHO	chr-F
FTA baseline	35.4**	36.3**	56.7**
FTA informal	30.5	28.3	52.7
FTA formal	30.7	27.3	53.1
FTA average	32.2	30.6	54.2
MRAnonoise neutral	30.1	23.6	52.8
MRAnonoise informal	32.3	30.8**	55.0**
MRAnonoise formal	33.7**	30.8**	55.5**
MRAnonoise average	32.0	28.4	54.4
MRAnoise neutral	**36.5***	**38.1***	**57.5***
MRAnoise informal	34.1*†	35.1*†	55.8*†
MRAnoise formal	32.8†	30.1†	55.2†
MRAnoise average	**34.8**	**34.4**	**56.3**

Table 6: Automatic metric scores for all systems on the 1,200 segment common test set. Best results are highlighted in bold. Statistically significant results are also marked: * for comparisons between the MRAnonoise and MRAnoise engines per direction, † for MRAnoise and FTA, and ** for MRAnonoise and FTA.

tions (see Table 5). In general, the informal directions achieve the overall highest scores for each approach, while the formal directions tend to achieve better scores than their respective baseline/neutral directions (except for the FTA engine, where the baseline outperforms the formal direction). Even when this seems to emerge as a trend, note that further analysis is required for precise conclusions, as these particular test sets are not directly comparable.

Comparisons across systems based on the common test set (see Table 6) show that the baseline/neutral engines achieve some of the best scores even when they obtained the worst scores in their specific test sets. This strengthens the idea that the informal engines might be in general over-fitted to their training data, while the baseline/neutral models might be better suited to respond to other data.

If we turn to the MRA engines, we see that MRAnoise achieves significantly better results than its MRAnonoise counterpart for the neutral and informal registers, and also significantly better results for the informal and the formal registers than the FTA engine. When comparing the MRAnonoise and the FTA engines, the informal and formal registers of the former significantly outperform the latter, yet, not the baseline. This might imply that, when fine-tuning a baseline to the different registers, there is a bigger drop in performance. This is not the case when training a multi-register model with noise added to each register.

In Table 6, we also present the average performance of each engine (averaging the scores from the baseline/neutral, formal and informal registers). As it can be seen, the directions from the MRAnoise engine achieve the best average scores for all metrics, with a difference of more than 2 points for each metric over the second best engine (FTA). The MRAnonoise engine presents the lowest scores.

4.2 Human evaluation

Automatic metrics are dependent on the reference segments and their original quality. Therefore, in order to have an assessment of the quality from a human perspective, we also performed a set of human evaluations.

For these assessments, we created a test suite *ad-hoc*, from now on LINGtest[11], which contains 50 segments divided into two categories: those with overt second person forms in the source (YOU_FORMS), intended to cover the different forms that *tú* and *usted* can take in Spanish (*you*, *your*, *yours*), and those with no overt forms or verbs (NO_FORMS). This will allow us to check how the systems perform when faced with overt and non-overt cases.

We translated the 50 segments from the LINGtest using each of the 9 directions of the three approaches trained, which amounted to 450 unique translations.

In order to create the sets for evaluators, we allocated 50 segments to each set in a way that all the sets included translations from all engines while no source segment was repeated, and we could collect responses for all 450 translations. Given the subjective nature of the evaluation, we collected three assessments per translation. A total of 30 volunteer evaluators (native or near-native speakers of Spanish with varying expertise in NLP) were asked to score the translations of the LINGtest according to accuracy and fluency on a 5-point scale. Additionally, they were given the opportunity to comment on any aspect they considered relevant. It is important to note that, to avoid bias, they were not aware of the focus of the assessment (politeness) nor that they were evaluating output from different engines.

To obtain the final human results, we averaged the scores for each translation given by each evaluator. For the general system-level score, we averaged the previous segment-scores again. The average inter-annotator agreement of our research was 0.25 (calculated using Fleiss' Kappa).

Results for quality assessment show that all engines achieve adequacy and fluency scores above 4 points, which in our measuring scale means all engines tend to preserve most of the meaning of the original sentence and have good fluency, although they are not flawless (see Table 7). Contrary to automatic metrics, human assessments seem to indicate that the FTA baseline achieves the best adequacy and overall scores, and is the second best for fluency.

Interestingly, for adequacy, we observe that, when compared to the formal and baseline/neutral registers within the same engine, all the informal directions achieve worse results except for MRAnoise. This might indicate that the MRAnoise informal direction indeed benefited from the addition of sentences belonging to the neutral and formal subsets. In fact, average scores for each approach show that MRAnoise achieves the best overall score.

To check whether the performance of the engines degrades with certain types of linguistic phenomena in particular, we next took a more detailed look into the scores given to the different types of segments (YOU_FORMS and NO_FORMS). In Table 8, we present the overall scores (calculated as the mean of adequacy and fluency) for each engine and register, as well as the difference in the performance between the YOU_FORMS and the NO_FORMS segments.

The results show a difference in behaviour according to the type of segment. The engines that were trained with more strictly filtered data show better performance in the YOU_FORMS segments, while their performance decreases with the NO_FORMS segments to some extent. However, the directions trained with data belonging to the different register subsets achieve worse performance in the YOU_FORMS segments but do not experience such a sharp decrease in quality with the NO_FORMS segments. This calls for further experiments to establish the optimal proportion of register segment types at training-time.

On the other hand, while MRAnonoise achieves some of the best results for the informal and formal registers in the YOU_FORMS segments, its neutral register lags behind, which suggests that this engine did not benefit from being trained with only segments extracted from the neutral subset.

We carried out a final analysis to focus on the specific handling of the honorifics by the engines. We reviewed all the translations in the LINGtest and annotated (1) whether the systems overgenerated honorifics for seg-

[11] Can we found in Appendix A.

System	Adequacy	Fluency	Overall
FTA baseline	**4.51**	4.45	**4.48**
FTA informal	4.05	4.32	4.18
FTA formal	4.18	4.14	4.16
FTA average	4.25	4.30	4.28
MRAnonoise neutral	4.21	4.16	4.18
MRAnonoise informal	4.13	4.43	4.28
MRAnonoise formal	4.47	4.35	4.41
MRAnonoise average	4.27	4.31	4.29
MRAnoise neutral	4.39	4.37	4.38
MRAnoise informal	4.36	4.42	4.39
MRAnoise formal	4.34	**4.47**†	4.35
MRAnoise average	**4.36**	**4.42**	**4.37**

Table 7: Average human assessment scores for adequacy and fluency on the LINGtest. Best scores are in bold. † marks statistically significant differences when comparing the MRAnoise and the FTA approach.

System	YOU_FORMS	NO_FORMS	DIFFERENCE
FTA baseline	4.38	**4.59**†	+0.21
FTA informal	4.41	3.79	-0.62
FTA formal	4.26	4.08	-0.18
MRAnonoise neutral	4.16	4.23	+0.7
MRAnonoise informal	**4.6**	3.87	-0.73
MRAnonoise formal	4.45	4.47	+0.02
MRAnoise neutral	4.51*	4.21	-0.3
MRAnoise informal	4.44	4.33*†	-0.11
MRAnoise formal	4.38	4.28	-0.1

Table 8: Average human assessment scores for adequacy and fluency on the LINGtest per segment type. Best scores are in bold. Statistically significant results are also marked: * for comparisons between the MRAnonoise and MRAnoise engines per register, † for MRAnoise and FTA, and ** for MRAnonoise and FTA.

System	POLITENESS ACCURACY	HALLUCINATIONS
FTA informal	96.7%	20%
FTA formal	90%	15%
MRAnonoise informal	100%	50%
MRAnonoise formal	96.7%	0%
MRAnoise informal	90.3%	5%
MRAnoise formal	90.3%	5%

Table 9: Politeness test of segments with second person forms in the source.

ments with no overt second person forms or no verbs in the source segments (NO_FORMS segments); and (2) whether they actually produced the correct formal and informal forms as intended (YOU_FORMS). *Politeness accuracy* is calculated as the number of times the informal and formal engines outputted the right register divided by the total number of YOU_FORMS instances (30), while *Halluciations* is calculated as the total number of segments with overgenerated honorifics divided by the total number of NO_FORMS instances (20). Scores for *Politeness accuracy* were not calculated for the neutral and baseline systems, since they were not intended to handle any particular register and, due to their training data, did not overgenerate honorifics in the NO_FORMS segments.

Results show that honorifics are very accurately handled in all engines and registers, with over 90% of the instances correctly generated (see Table 10). MRAnonoise is the best approach, at par with the FTA for the informal register. However, we observe different tendencies with regards hallucinations: overall MRAnoise is the best performer, with the more consistent low level at 5%. MRA-nonoise is able to avoid all overgeneration for the formal register, but reaches a 50% high for the informal register. Meanwhile, the proportions for the FTA engine remain between 15% and 20%.

5 Conclusions and future work

In this work, we studied ways to control politeness in NMT for Castilian Spanish. Our first contribution to this topic was the creation of politeness-specific sets for the new language pair –based on the approach used by Sennrich, Haddow, and Birch (2016a) for German–. By adapting their methodology, we classified Spanish segments from the OpenSubtitles corpus into three levels of politeness (formal, informal and neutral) with an average accuracy on a population sample of over 90%.

We then used the separate subsets to explore two main domain-adaptation techniques to address politeness in English>Castilian Spanish NMT: fine-tuning and multilingual models.

Automatic evaluation results seem to show that, overall for our case, the multi-register approach with noise might be better suited than the fine-tuning approach when a balance between accuracy at choosing the honorific and performance in the different registers is the key. However, whether these results are due to the domain adaptation technique used for training or to the incorporation of noise into the training data should be further studied.

We extended the evaluation of the results with multiple human evaluations, which help to understand the handling of the registers more in detail. According to the adequacy and fluency judgements, the ranking of the engines varies slightly. Adequacy and overall quality seem to be better achieved by the baseline system trained as part of the fine-tuning approach. The best overall fluency is achieved by the multi-register formal engine trained with noise.

It is interesting to note that several annotators reported concern about their assessment, stating that they were not too sure about how to evaluate politeness-related issues. We take these statements not as a weakness of the evaluation but rather as a clear sign that politeness is a relevant feature to establish the appropriateness -and quality- of a translation. Therefore, as politeness can be a factor that can direct the assessment, we suggest that evaluations, whether register-related or general- may benefit from including specific guidelines as to how to treat register-related issues.

Additionally, the specific politeness-related analyses showed that the engines did not always perform consistently for the different types of segments that display (or omit) register-related elements. In any case, we observed that the accuracy of honorifics was above 90% for all engines and registers. In terms of hallucinations, multi-register models performed better –except for the informal direction trained with no noise– while the FTA models seem to have suffered from some degree of catastrophic forgetting, which can lead to a worse overall performance of those models in segments with no second person forms and no verbs in the source when compared to the MRAnoise system.

In this line, in future work, we aim to explore the use of mixed fine-tuning (as proposed by Chu, Dabre, and Kurohashi (2017)) in the quest for palliating catastrophic forgetting in fine-tuned systems. Additionally, driven by the growing interest in the research and development of register-aware NLP technologies, we also intend to work with other features that configure politeness for Spanish beyond the use of honorifics and to provide an enlarged and refined version of the register-annotated corpus created for this work with the aim of contributing to the community with a high-quality resource to be used in other NLP applications beyond MT.

Acknowledgements

The research leading to this work was partially funded by the *SignOn* RIA ICT-57-2020-101017255 project, the *TANDO* ELKARTEK 20/49 project and the IT1570-22 project (Basque Government). We would also like to thank the volunteers who freely participated in the human evaluation process. Finally, we would like to express our gratitude to Fred Bane for laying the foundations of this research.

Bibliography

Aharoni, R., M. Johnson, and O. Firat. 2019. Massively multilingual neural machine translation. 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 3874–3884, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Bane, F. and A. Zaretskaya. 2021. Selecting the best data filtering method for NMT training. In *Proceedings of Machine Translation Summit XVIII: Users and Providers Track*, pages 89–97, Virtual, August. Association for Machine Translation in the Americas.

Bapna, A., N. Arivazhagan, and O. Firat. 2019. Simple, scalable adaptation for neural machine translation. *arXiv preprint arXiv:1909.08478.*

Briz, A. 2010. Lo coloquial y lo formal, el eje de la variedad lingüística. *De moneda nunca usada: Estudios dedicados a José Mª Enguita Utrilla*, 125:133.

Brown, P. 2015. Politeness and language. *The International Encyclopedia of the Social and Behavioural Sciences (IESBS),(2nd ed.)*, pages 326–330.

Bulte, B. and A. Tezcan. 2019. Neural fuzzy repair: Integrating fuzzy matches into neural machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1800–1809, Florence, Italy, July. Association for Computational Linguistics.

Chu, C., R. Dabre, and S. Kurohashi. 2017. An empirical comparison of domain adaptation methods for neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 385–391, Vancouver, Canada, July. Association for Computational Linguistics.

Chu, C. and R. Wang. 2018. A survey of domain adaptation for neural machine translation. In *Proceedings of the 27th International Conference on Computational Linguistics*, page 1304–1319, Santa Fe, New Mexico, USA, August. Association for Computational Linguistics.

Dinu, G., P. Mathur, M. Federico, and Y. Al-Onaizan. 2019. Training neural machine translation to apply terminology constraints. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3063–3068, Florence, Italy, July. Association for Computational Linguistics.

Etchegoyhen, T., E. Martínez Garcia, A. Azpeitia, G. Labaka, I. Alegria, I. Cortes Etxabe, A. Jauregi Carrera, I. Ellakuria Santos, M. Martin, and E. Calonge. 2018. Neural machine translation of Basque. In *Proceedings of the 21st Annual Conference of the European Association for Machine Translation*, pages 139–148, Alacant, Spain, May. European Association for Machine Translation.

Feely, W., E. Hasler, and A. 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.

Halliday, M., A. McIntosh, and P. Stevens. 1964. *The language Science and Language Teaching.* London. Longman.

Haugh, M. 2005. The importance of "place" in japanese politeness: Implications for cross-cultural and intercultural analyses. *Japanese Politeness: Implications for Cross-Cultural and Intercultural Analyses*, 2(1):41–68.

Junczys-Dowmunt, M., R. Grundkiewicz, T. Dwojak, H. Hoang, K. Heafield, T. Neckermann, F. Seide, U. Germann, A. F. Aji, N. Bogoychev, A. F. T. Martins, and A. Birch. 2018. Marian: Fast neural machine translation in C++. In *Proceedings of ACL 2018, System Demonstrations*, pages 116–121, Melbourne, Australia, July. Association for Computational Linguistics.

Kell, G. 2018. *Overcoming catastrophic forgetting in neural machine translation.* Ph.D. thesis, MPhil dissertation, University of Cambridge.

Kirkpatrick, J., R. Pascanu, N. Rabinowitz, J. Veness, G. Desjardins, A. A. Rusu, K. Milan, J. Quan, T. Ramalho,

A. Grabska-Barwinska, et al. 2017. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526.

Kobus, C., J. Crego, and J. Senellart. 2017. Domain control for neural machine translation. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 372–378, Varna, Bulgaria, September. INCOMA Ltd.

Koehn, P. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 388–395, Barcelona, Spain, July. Association for Computational Linguistics.

Koehn, P. and R. Knowles. 2017. Six challenges for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 28–39, Vancouver, August. Association for Computational Linguistics.

Luong, M.-T. and C. Manning. 2015. Stanford neural machine translation systems for spoken language domains. In *Proceedings of the 12th International Workshop on Spoken Language Translation: Evaluation Campaign*, pages 76–79, Da Nang, Vietnam, December 3-4.

Matthiessen, C. and M. Halliday. 1997. *Systemic functional grammar*. Amsterdam and London: Benjamins & Whurr.

Ott, M., S. Edunov, A. Baevski, A. Fan, S. Gross, N. Ng, D. Grangier, and M. 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, June. Association for Computational Linguistics.

Popović, M. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal, September. Association for Computational Linguistics.

Post, M. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium, October. Association for Computational Linguistics.

Post, M. and D. Vilar. 2018. Fast lexically constrained decoding with dynamic beam allocation for neural machine translation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1314–1324, New Orleans, Louisiana, June. Association for Computational Linguistics.

Rei, R., A. C. Farinha, C. Stewart, L. Coheur, and A. Lavie. 2021a. MT-Telescope: An interactive platform for contrastive evaluation of MT systems. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 73–80, Online, August. Association for Computational Linguistics.

Rei, R., A. C. Farinha, C. Zerva, D. van Stigt, C. Stewart, P. Ramos, T. Glushkova, A. F. T. Martins, and A. Lavie. 2021b. Are references really needed? unbabel-IST 2021 submission for the metrics shared task. In *Proceedings of the Sixth Conference on Machine Translation*, pages 1030–1040, Online, November. Association for Computational Linguistics.

Sennrich, R., B. Haddow, and A. Birch. 2016a. 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, June. Association for Computational Linguistics.

Sennrich, R., B. Haddow, and A. Birch. 2016b. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany, August. Association for Computational Linguistics.

Sennrich, R., M. Volk, and G. Schneider. 2013. Exploiting synergies between open resources for German dependency parsing, POS-tagging, and morphological analysis. In *Proceedings of the International Conference Recent Advances in Natural Language Processing RANLP 2013*, pages 601–609, Hissar, Bulgaria, September. INCOMA Ltd. Shoumen, BULGARIA.

Tiedemann, J. 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, May. European Language Resources Association (ELRA).

Tiedemann, J. and S. Thottingal. 2020. OPUS-MT – building open translation services for the world. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 479–480, Lisboa, Portugal, November. European Association for Machine Translation.

Van Merriënboer, B., D. Bahdanau, V. Dumoulin, D. Serdyuk, D. Warde-Farley, J. Chorowski, and Y. Bengio. 2015. Blocks and fuel: Frameworks for deep learning. *arXiv preprint arXiv:1506.00619*.

Vanmassenhove, E., D. Shterionov, and M. Gwilliam. 2021. Machine translationese: Effects of algorithmic bias on linguistic complexity in machine translation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2203–2213, Online, April. Association for Computational Linguistics.

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. 2017. Attention is all you need. In *31st Conference on Neural Information Processing Systems (NIPS 2017)*, pages 1–11, Long Beach, CA, USA, December.

A Appendix: LINGtest

YOU_FORMS	NO_FORMS
- You should go to the doctor if you are feeling sick. - What did you do yesterday? - We are available via Whatsapp to solve any questions you may have during the purchase - It was you who started the fight. - Who did it? Was it you? - Yesterday, we went out for a couple of drinks downtown. What about you guys? - Is it you, Tom? - You need to be the one that picks up the parcel. - Can you check your agenda and let me know when you are free? - How was your experience with us? - Did you break your arm? - I believe that T-shirt was yours. - Let's take my car, not yours. - Please enter your address. - Where do you wish to receive your items? - Your purchase is almost done! - How was your experience with us? - Come with us, please! - Contact us at XXXXX. - Call me when you get home. - Click on the item you wish to purchase. - Look at this. - Can I come with you? - We have all these new items for you! - No, thank you - Please, do not hesitate to contact us and ask for a refund. - I made all this for you. - We would love to go to the cinema with you tonight. - Did she come with you? - I was waiting for you guys forever!	- Nonsense! - Why not? - How cool! - Seriously? - Postal code - Next item - Hey, there! - Welcome! - Where? There? - Customized delivery services - We are delighted to be here today. - I am really happy to be here today. - They were suppose to come today. - We enjoyed it so much! - Personally, I think that is not true. - He was such a nice person. - She moved to Madrid to attend University. - Offering customized delivery services since 1996. - They asked me whether I wanted a refund. - Let's go together.

Table 10: Test suite for human evaluation.