Controlling Formality and Style of Machine Translation Output Using AutoML

Aditi Viswanathan$^{1[0000-0002-9373-8295]}$, Varden Wang$^{1[0000-0002-7091-9827]}$, and Antonina Kononova^{1}

Google
{aditiv,varden,akononova}@google.com

Abstract. An often overlooked difficulty of machine translation is producing a consistent formality (or register) in the target language. This is especially hard when the source language may have fewer levels of formality than the target language. We take a transfer learning approach using Google's AutoML Translate to train custom neural machine translation (NMT) models to consistently produce a specific formality. We experiment with formality levels for English to Spanish, English to French and English to Czech. This approach makes it possible to have better and more consistent in-context translation while still leveraging the strength of a general purpose machine translation system.

Keywords: Machine Translation · Domain Adaptation · Formality

1 Introduction

An important aspect of using machine translation (MT) in a business setting is maintaining a **consistent** tone and style–often called register [1] (or degree of politeness). Often, organizations want to translate text for a particular type of customer, situation or market where register is important. For example, translations for diplomatic communications differ in register from those for social media; or translations targeted at consumers are generally less formal than translations for business customers.

The style and tone of the MT output in these cases can be as important to businesses as meaning preservation and fluency. Furthermore, often the source text does not indicate the formality level of the target translation. English, for example, has fewer levels of formality [13] than many other languages, such as French or Korean–which has at least six levels of formality [11]. Consequently, there is inconsistency in using a generic MT system between these language pairs, where some sentences may be translated using formal grammatical markers while others get translated with informal grammatical markers. This leads to a loss of context across a series of sentences and is a significant impediment to the use of general purpose MT in business settings.

In this experiment, we do not seek to model the full range of variability that registers may cover. Such an experiment would be quite complicated and would require us to parse out many different sociolinguistic situations. Instead, we focus on the register associated with personal pronouns in the phenomena of **T-V distinctions** [2], which is an important factor in formality, especially for Romance languages [8] like French and Spanish. In the case of Spanish, the second person *tú* (”you”) is used when communicating with those familiar with the speaker while the more respectful form is the second-person *usted*.

In this paper we explore a simple and fast non-rules based technique for developing custom machine translation models that produce translations consistently in the desired style or formality/register using the Google Cloud AutoML Translation framework [14]. This can be seen as a special case of domain adaptation.

2 Related Work

Previous work on Formality-Sensitive Machine Translation [5] were developed using standard phrase-based MT architecture implemented as an n-best re-ranking system. Other explored techniques include style transfer after translation [9,6], but this technique requires already having an adequate translation available.

R. Sennrich et al. [10] proposed a method using *side constraints*–additional markers for input features such as politeness or formality. This approach relies on annotating politeness in the training set to obtain the politeness feature. Results are effective with English to German showing that translations constrained to be polite were in fact labelled polite or neutral 96% of the time and labelled informal or neutral when constrained to be informal 98% of the time. However, this method relies on passing in special tokens to mark politeness as part of the source input.

Similar approaches using domain adaptation [4] focus upon reflecting personal traits of the source speaker in the target translation. The aforementioned method proposes to learn speaker-specific parameters, which the authors cast as extreme domain adaptation. Our framing of the problem is orthogonal to this as we seek a consistent and uniform stylistic translation output regardless of any personal traits of the speaker or source input.

3 System and Training

Neural machine translation (NMT) is an approach to machine translation that uses a large artificial neural network to predict the likelihood of a sequence of words, typically modeling entire sentences in a single integrated model.

To create a general purpose formality model, we choose AutoML Translate[1], a Google Cloud AI product for customizing NMT engines for specific industries and domains. The AutoML Translation framework uses transfer learning and neural architecture search [14] to train new models on the basis of preexisting

[1] See `https://cloud.google.com/translate/automl/docs/` for official Google documentation.

NMT models. In particular, it leverages the Google NMT (GNMT) system– a sequence-to-sequence neural machine translation system consisting of a deep LSTM network [3,12] as the baseline model. This framework is well suited to creating domain-specific customized models from in-domain input datasets that can also generalize well to different tasks.

3.1 Model Development

We create training, test and validation datasets by filtering parallel text for a set of seed words that are markers for informal and formal registers (T-V) in that language (See Table 1). These markers are chosen based on linguistic rules for the language, and are chosen such that they create sufficient (not necessary) conditions to determine the T-V register for that language. We create datasets based on these conditions[2] from Google's bilingual data (See Table 4 for details on the dataset sizes per model).

Table 1. Formality-specific markers are used to create datasets that are then used to train and evaluate custom models. (T) = Informal, (V) = Formal

Spanish (T)	Spanish (V)	French (T)	French (V)	Czech (V)
tú	él	tu	vous	vy
tu	ella	te	votre	vás
tus	Ud.	toi	vos	vám
ti	se	ton	vôtre	vámi
tuyo	usted	tes	vôtres	váš
tuyos	suyo	ta		vaše
tuya	suyos	tien		vašeho
tuyas	suya	tiens		vašemu
te	suyas	tienne		vaší
contigo		tiennes		vaším
				vašem

We then use these formality-specific datasets to train custom models that are biased towards the respective registers using AutoML Translate. We initially

[2] For example, in Spanish, if a target segment contains any of the words listed in Column 1 of Table 1, it is a sufficient condition to determine that its register is informal (T). However, determining that a target segment is of the formal register (V) is more challenging because some of the words that signify the formal register are also used to refer to the 3rd person (e.g. Spanish *suyo* can mean English formal *yours* or 3rd person *his*). To solve for this, we filter segment pairs where the target segment contains (V) markers **and** the source segment contains any English inclusion words like 2nd person pronouns (e.g. "you", "yours") **and** does **not** contain any English exclusion words like 3rd person pronouns (e.g. "her", "she", "them"). This combined rule is a sufficient condition to determine that the register of the target segment is formal (V).

use the generic GNMT model [12] as the base model and train a custom model on top of the base per formality register and language pair. We repeat this step multiple times, each time using the custom model trained in the previous step as the base model for the current step. The training data at each step remains the same–the intention with this approach is to force a strong bias on either the T or V form, while retaining the ability to generalize well (See Table 3 for example model outputs). We see significant incremental improvement in formality biasing with this iterative warm start approach. For French and Spanish, we observe that running the training 2-3 times performs the best in biasing towards a specific formality register while preserving meaning and fluency. We expect that further experimentation can help identify the optimal number of "warm re-starts" per register and language pair.

4 Evaluation

Setup In order to evaluate whether the translation models successfully produce the desired register, we ask human translators to develop translation references of differing formality registers (formal and informal) from the same source segments[3] (see example Table 2). We use 400 source segments that are drawn randomly from the WMT '11 and '12 Translation Task test sets.

The evaluation sets are then divided by formality level. The formal set will be used to evaluate the formal models and the informal set used to evaluate the informal models. The translated segments from each formality level–from both human translation and machine translation–is then sent through human evaluation to rate the quality and formality of the translations.

For automatic evaluation, we use larger evaluation sets of 10000 segments for each language and formality register. These evaluation sets are created using the same methodology used to create the training datasets (see Table 1).

Table 2. Use of formal singular 'you' vs informal singular 'you' with verb agreement.

English	Formal (V)	Informal (T)
Juan, how are you?	Juan, ¿cómo está **usted**?	Juan, ¿cómo está**s**?
Do you know where the house is?	¿Sabe **usted** dónde está la casa?	¿Sabe**s** **tú** dónde está la casa?

4.1 Automatic Evaluation

We use the standard automatic machine translation evaluation metric BLEU [7], with single references, to baseline the formality biased models relative to the

[3] Segments may consist of either a single sentence or multiple sentences.

Table 3. Example model outputs from Spanish custom models

English	Generic MT	**Formal Bias Model**	**Informal Bias Model**
However, you can get the second one for free.	Sin embargo, **puede** obtener el segundo de forma gratuita. (V)	Sin embargo, **usted puede** conseguir el segundo de forma gratuita.	Sin embargo, **te puedes** conseguir el segundo gratis.
You will just sleep better.	Sólo dormirá**s** mejor. (T)	**Usted** sólo dormirá mejor.	Sólo dormirá**s** mejor.

Table 4. Comparison of BLEU scores across English to French and English to Spanish models

Language	Evaluation Set Register	Dataset Size (segment pairs)	**Formal Bias Model**	**Informal Bias Model**	Google Translate (GT)	Performance Gain over GT
French	Formal	33M	**62.362**	39.668	57.28	+5.082
	Informal	5.6M	29.972	**50.699**	40.844	+9.855
Spanish	Formal	33M	**54.394**	36.449	42.603	+11.791
	Informal	33M	37.165	**65.686**	46.476	+19.21
Czech	Formal	4K	**51.117**	-	33.75	+17.367
	Formal	2.3M	**56.323**	-	33.004	+23.319

base Google NMT provided by AutoML. BLEU has reasonably high correlation with human judgments of quality. It helps us understand how well the model is biasing towards a specific formality as matching markers in the model output should be reflected in the reference set. While BLEU should never be used as the only metric to assess translation quality, it provides a quick and useful measure for rapidly iterating and improving systems.

Results Table 4 exhibits our BLEU scores for French, Spanish and Czech formality models on Formal/Informal register evaluation datasets. As expected, the custom AutoML model with a formal bias does better than generic MT on evaluation sets that have a formal register; and the custom AutoML model with an informal bias does better on evaluation sets that have an informal register. The BLEU score performance differences are especially significant for English to Spanish and English to Czech models.

4.2 Human Evaluation

Human evaluation of translations tend to be a more reliable and authoritative method in measuring machine translation quality. So, in addition to using BLEU scores–which may conflate translation errors and formality mismatches–we ask bilingual human raters to rate the machine translation output on both the traditional measures of fluency and meaning, but also level of formality. For the languages we have chosen we ask raters to rate the formality as formal, informal, or neutral. The raters are bilingual native speakers of the target language.

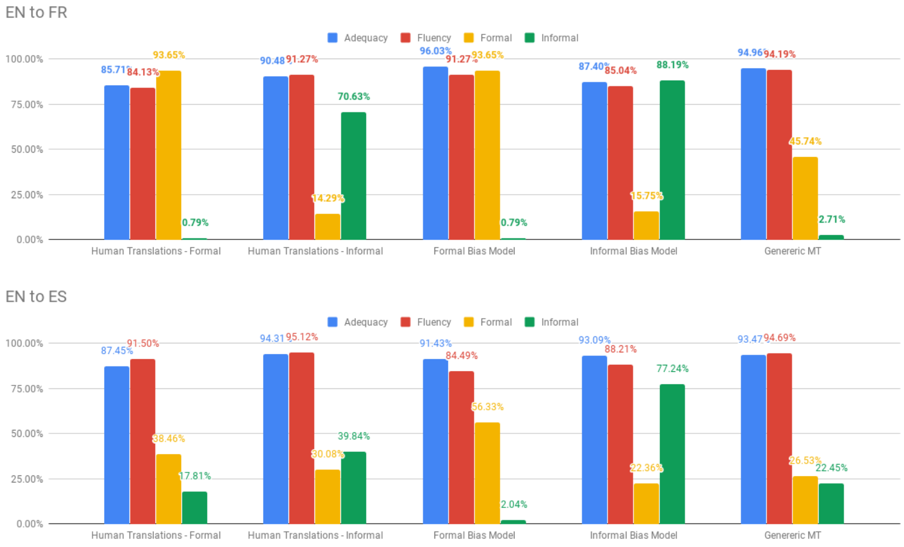

Fig. 1. Comparison of Adequacy, Fluency and Formality across models. Neutral formality ratings are not shown. In general, our models are able to bias the source text to the desired formality. Furthermore, our adequacy and fluency ratings are comparable to both human translations and Generic MT.

Results Adequacy is rated on a 4 point scale going from None, Little, Most, All. Fluency is also rated on a 4 point scale going from Nonsense, Poor, Good, Flawless. Formality was rated as Informal, Neutral, or Formal. In Fig. 1: Adequacy is shown as percentage of segments in the evaluation set receiving adequacy ratings of Most or All meaning preserved; Fluency is shown as a percentage of segments in the evaluation set receiving fluency ratings of Good or Flawless.

Interestingly, we suspect that the dramatic difference between the French and Spanish systems in the human evaluation results for formality in the Human Translations-Formal evaluation and the Formal Bias Model evaluation may have to do with the consistency and source of the training data we used. The data for our French models came primarily from parallel text aimed towards the variation of French in France. The data used for the Spanish translation models came from a larger variety of locales including different Latin American varieties of Spanish as well as Spanish from Spain. Therefore, we surmise that agreement on formality may be lower due to local differences on what is considered formal or informal.

Human evaluation for Czech was only performed for the 2.3M Formal Czech model. Adequacy and fluency were 89% and 87% respectively, with 98.3% of the segments rated as formal.

5 Conclusion

In this paper, we use a domain adaptation technique to bias a model to produce translations according to a desired formality or register while still maintaining a high level of fluency and meaning. After proper training, translations with unintended formality levels have been almost eliminated from our models. Additionally, the Czech models indicate that by leveraging transfer learning from the base model, it is possible to develop a formal model by tuning with a dataset of fewer than 5 thousand sentences. Our evaluation shows the effectiveness of this technique in producing consistent in-context translations with a specific formality register, without a significant loss in translation quality.

5.1 Further Experiments

We would like to extend our technique to other languages. Languages like Korean are said to have at least six levels of formality. It would be interesting to see how well this technique captures the differentiation between them. In a few cases, our models produce translations with mixed formalities. Reducing or detecting such errors is also an interesting basis for future work on this technique.

Lastly, we want to expand this technique beyond just T-V distinctions. Based on some experiments we've run on French to English parallel text from 12 Shakespearean comedies (See Table 5 for example output), it is possible to use this technique to create domain-adapted custom models that reflect a personality or language style.

Table 5. Example model outputs from an experiment on French to English Shakespearean data

French	Generic MT	Custom Shakespearean Model
Qu'est-ce que tu fais?	What are you doing?	What art thou doing?
Oui! C'est toi que je veux dire.	Yes! It's you I want to say.	Aye! I mean thee.
Comment on est aujourd'hui?	How are we today?	How now?

References

1. Biber, D., Finegan, E.: Sociolinguistic perspectives on register. Oxford University Press on Demand (1994)
2. Brown, R., Gilman, A., et al.: The pronouns of power and solidarity. Bobbs-Merrill (1960)
3. Chen, M.X., Firat, O., Bapna, A., Johnson, M., Macherey, W., Foster, G., Jones, L., Parmar, N., Schuster, M., Chen, Z., et al.: The best of both worlds: Combining recent advances in neural machine translation. arXiv preprint arXiv:1804.09849 (2018)
4. Michel, P., Neubig, G.: Extreme adaptation for personalized neural machine translation. arXiv preprint arXiv:1805.01817 (2018)
5. Niu, X., Martindale, M., Carpuat, M.: 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. pp. 2814–2819 (2017)
6. Niu, X., Rao, S., Carpuat, M.: Multi-task neural models for translating between styles within and across languages. arXiv preprint arXiv:1806.04357 (2018)
7. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: A method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics. pp. 311–318. ACL '02, Association for Computational Linguistics, Stroudsburg, PA, USA (2002). https://doi.org/10.3115/1073083.1073135, `https://doi.org/10.3115/1073083.1073135`
8. Posner, R.: The romance languages. Cambridge University Press (1996)
9. Rao, S., Tetreault, J.: Dear sir or madam, may i introduce the gyafc dataset: Corpus, benchmarks and metrics for formality style transfer (2018)
10. Sennrich, R., Haddow, B., Birch, A.: 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. pp. 35–40 (2016)
11. Sohn, H.M.: The Korean Language. Cambridge University Press (2001)
12. Wu, Y., Schuster, M., Chen, Z., Le, Q.V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., Klingner, J., Shah, A., Johnson, M., Liu, X., Łukasz Kaiser, Gouws, S., Kato, Y., Kudo, T., Kazawa, H., Stevens, K., Kurian, G., Patil, N., Wang, W., Young, C., Smith, J., Riesa, J., Rudnick, A., Vinyals, O., Corrado, G., Hughes, M., Dean, J.: Google's neural machine translation system: Bridging the gap between human and machine translation. CoRR **abs/1609.08144** (2016), `http://arxiv.org/abs/1609.08144`
13. Xiao, Z., McEnery, A.: Two approaches to genre analysis: Three genres in modern american english. Journal of English Linguistics **33**(1), 62–82 (2005)
14. Zoph, B., Le, Q.V.: Neural architecture search with reinforcement learning. arXiv preprint arXiv:1611.01578 (2016)