Controlling Extra-Textual Attributes about Dialogue Participants: A Case Study of English-to-Polish Neural Machine Translation

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

Unlike English, morphologically rich languages can reveal characteristics of speakers or their conversational partners, such as gender and number, via pronouns, morphological endings of words and syntax. When translating from English to such languages, a machine translation model needs to opt for a certain interpretation of textual context, which may lead to serious translation errors if extra-textual information is unavailable. We investigate this challenge in the English-to-Polish language direction. We focus on the underresearched problem of utilising external metadata in automatic translation of TV dialogue, proposing a case study where a wide range of approaches for controlling attributes in translation is employed in a multi-attribute scenario. The best model achieves an improvement of $+5.81$ chrF++/$+6.03$ BLEU, with other models achieving competitive performance. We additionally contribute a novel attribute-annotated dataset of Polish TV dialogue and a morphological analysis script used to evaluate attribute control in models.

1 Introduction

In some languages, dialogue explicitly expresses certain information about the interlocutors: for example, while in English words describing the speaker "I" and the interlocutor "you" are ambiguous w.r.t. their gender, number and formality, languages such as Polish, German or Spanish will mark for one or more of these attributes. In industrial settings such as dubbing and speech translation, there is an abundance of available metadata about the interlocutors, such as their genders, that could be used to help resolve these ambiguities.

Field	Value
source	"Are you blind?"
spoken by (=speaker)	"Anne"
speaker's gender	"feminine"
spoken to (=interlocutor(s))	["Mark", "Colin"]
interlocutor(s)' gender	"masculine"
formality	"informal"

Table 1: A TV segment along with available metadata.

Table 1 shows an example of such a TV segment: the English sentence *'Are you blind?'*, should translate to Polish as *'Jesteście ślepi?'* as the addressee is a group of men and the setting is informal; however, when spoken e.g. formally to a mixed-gender group of people, the correct translation would read *'Są państwo ślepi?'*, using a different verb inflection and an honorific *państwo*. Since the contextual information required to resolve the ambiguity in this example does not belong to the text itself, traditional models do not use it. This yields hypotheses which introduce some assumptions about that context, typically reflecting biases present in the (often unbalanced) training data. To avoid this, a better solution is to resolve such ambiguities by using both the available metadata and the source text as translation input. Alternatively, when such information is unavailable, all possible contextual variants could be provided as output, passing the choice from the model to the user (Jacovi et al., 2021; Schioppa et al., 2021).

In the context of the gender of the speaker and interlocutor, prior research has explored two ways

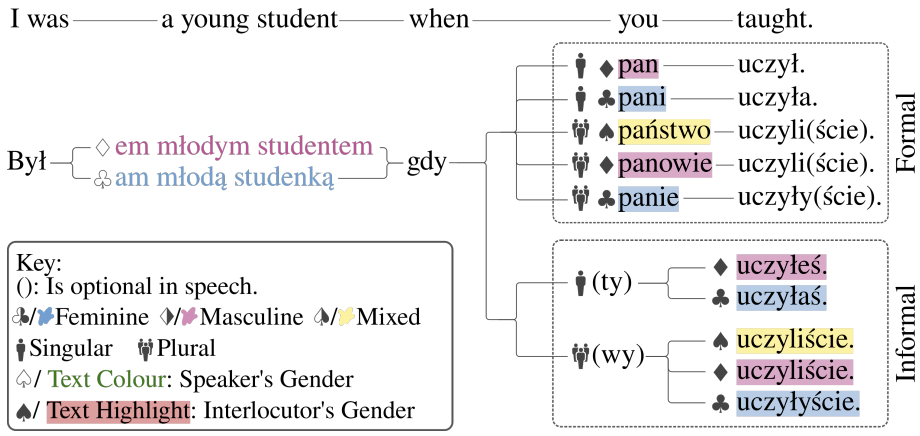

Figure 1: Example of an ambiguous English sentence with all plausible translations to Polish. There are a total of 18 equally plausible possible hypotheses based on the combination of contexts.

in which such information influences a text (Rabinovich et al., 2017; Vanmassenhove et al., 2018). Firstly, naturally occurring texts satisfy grammatical agreement between the gender of the speaker and interlocutor and the utterances which describe them. How this agreement is expressed in speech varies among different languages (Stahlberg et al., 2007). Polish is a *grammatical gender language*: every noun is assigned a gender, and grammatical forms must agree with that noun. In contrast, English is a *natural gender language*, with "no grammatical markings of sex" (Stahlberg et al., 2007, p. 165). Secondly, gender can be seen as a demographic factor that influences the way people express themselves (e.g. word choice). Hereinafter we refer to the former as *grammatical agreement* and the latter as *behavioural agreement*.

In this work, we seek to build machine translation (MT) models that satisfy grammatical agreement. Given an English sentence and a set of attributes (e.g. the gender of the speaker and number of interlocutors), an MT system must translate this sentence into Polish with a correct grammatical agreement to all attributes but introduce no markings of behavioural agreement.

We explore the agreement to one SPEAKER attribute: the gender of the speaker (SPGENDER), and three INTERLOCUTOR attributes: the gender(s) and number of interlocutor(s) (ILGENDER, ILNUMBER), as well as the desired FORMALITY of addressing the interlocutor(s). Figure 1 exemplifies the extent of ambiguity these attributes introduce in English-to-Polish translation.

The **main contributions** of our work are: (1) a novel English-Polish parallel corpus of TV dialogue annotated for SPGENDER, ILGENDER, ILNUMBER and FORMALITY; (2) a tool for analysing attributes expressed in Polish utterances; (3) the examination of a wide range of approaches to attribute control in NMT, showing that at least four of them can be reliably used for incorporating extra-linguistic information within English-to-Polish translation of dialogue.

The paper is structured as follows. Section 2 discusses previous work. Section 3 presents the problem definition, focusing on Polish as the target language. The creation of the parallel English-Polish corpus of dialogue utterances that mark subsets of the investigated attributes is presented in Section 4.1. How the MT models are trained to control the four extra-textual attributes is discussed in Section 4.3, whilst the results are presented in Section 4.2. Finally, we describe conclusions and potential directions for future work in Section 6.

2 Related Work

The state-of-the-art in MT is currently represented by neural MT (NMT) (Sutskever et al., 2014; Bahdanau et al., 2015) implemented via the Transformer architecture (Vaswani et al., 2017). Despite their unparalleled performance, these models are limited by ignoring the extra-textual context (e.g. speaker's gender). Consequently, much recent work aims to control NMT with various attributes. In particular, attention has been paid to tasks such as multilingual NMT (Johnson et al., 2017), by specifying the target language in the input; formality or politeness transfer (e.g. Sennrich et al. (2016)); controlling the gender of the speaker and/or interlocutor (Elaraby et al., 2018; Vanmassenhove et al., 2018; Moryossef et al., 2019); length and verbosity (Lakew et al., 2019; Lakew et

al., 2021); or constraining the vocabulary (Ailem et al., 2021).

Attribute control in NMT is most commonly facilitated with a *tagging* (or *side constraints*) approach, whereby a set of terms is added to the vocabulary, each embedding a certain type. These are trained alongside token embeddings and used in various ways during inference. Controlling multiple attributes with this approach has not been excessively studied (Schioppa et al., 2021), but can be facilitated by simply concatenating the tags (Takeno et al., 2017). However, for a set of equally important attributes, their ordering should not matter, but a tagging approach by design requires tags to be ordered in a specific way. Combining attributes by averaging their embeddings has also been explored in previous work (cf. Lample et al. (2019), Schioppa et al. (2021)), where authors incorporated the resulting vectors either into the input of the Encoder or the Decoder or directly into the model (Michel and Neubig, 2018; Schioppa et al., 2021).

Typically, attribute-controlling neural models are fully supervised, requiring annotated training data. Such annotations can be obtained directly, e.g. from metadata (Vanmassenhove et al., 2018); although most available corpora are unannotated. Sennrich et al. (2016) and Elaraby et al. (2018) automatically annotate the data using morphosyntactic parsers based on rules, validating agreement to the attribute in question in target-side sentences. To verify that the rules capture the attribute completely, a precision/recall score is computed against a manually labelled test set.

3 Problem Specification

Recognising the small number of studies within machine translation research on the English-to-Polish language direction, as well as our capacity (thanks to the available parsers and native speakers to validate their performance), we decide to focus the study on this language pair. Polish is a West Slavic language spoken by over 50M people over the world (Jassem, 2003). It uses an expanded version of the Latin alphabet and is characterised by a complex inflectional morphology (Feldstein, 2001). It is a grammatical gender language (Koniuszaniec and Błaszkowska, 2003) meaning all forms dependent on pronouns must agree to their gender and number. It uses a West Slavic system of honorifics *pani, pan, panie, panowie, państwo* (henceforth *Pan+*) (Stone, 1977). Being a null-subject language (Sigurðsson and Egerland, 2009), it does not require that pronouns signifying the speaker or the interlocutor are explicit, **unless** they belong to the *Pan+* group (Keown, 2003).

English lacks a grammatical gender or a system of honorifics, and the pronoun "you" is used for both plural and singular second person addressees. It is therefore ambiguous w.r.t. some expressions describing the speaker or the interlocutor, which we capture into four attributes, as follows (the attributes are summarised in Table 2).

SPEAKER attributes The gender of all forms dependent on the pronoun *ja* "I" must match the gender of the speaker SPGENDER $\in \{feminine, masculine\}$. This includes past and future verbal expressions (e.g. *byłam* 'I was$_{\text{fem}}$' vs. *byłem* 'I was$_{\text{masc}}$'), adjectives (e.g. *piękna* 'pretty$_{\text{fem}}$' vs. *piękny* 'pretty$_{\text{masc}}$') and nouns (e.g. *wariatka* 'lunatic$_{\text{fem}}$' vs. *wariat* 'lunatic$_{\text{masc}}$') that describe the speaker.

INTERLOCUTOR attributes All word forms dependent on the pronoun *ty/wy/Pan+* "you", including the pronoun itself, must match:

- the gender of the interlocutor (ILGENDER); this includes cases analogous to SPGENDER, extended to e.g. vocatives (e.g. *Ty wariatko/cie!* 'You lunatic$_{\text{fem/masc}}$!');
- the number of interlocutors (ILNUMBER); this includes verbs and pronouns in second person;
- the formality in addressing the interlocutor (FORMALITY)[1]; this entails using an inflection of the pronoun Pan+ consistent with ILGENDER and ILNUMBER where applicable, or using polite forms (e.g. *Proszę wejść.* 'Come in.').

Attribute	Abbreviation	Type
SPEAKER		
SPGENDER	<*sp:feminine*>	Feminine speaker
	<*sp:masculine*>	Masculine speaker
INTERLOCUTOR		
ILGENDER	<*il:feminine*>	Feminine interlocutor(s)
	<*il:masculine*>	Masculine interlocutor(s)
	<*il:mixed*>	Mixed-gender interlocutor(s)
ILNUMBER	<*singular*>	One interlocutor
	<*plural*>	Multiple interlocutors
FORMALITY	<*informal*>	Informal
	<*formal*>	Formal

Table 2: Attributes and types controlled in the experiment.

[1]While we define formality as binary, it can be more complex e.g. Japanese in Feely et al. (2019).

Throughout this paper, when discussing *gender* we refer solely to grammatical gender rendered in utterances. In the Polish language, the grammatical system of gender in first and second person is a dichotomy of masculine and feminine variants, lacking alternatives for people who identify as neither. We discuss potential solutions to this issue in directions for future work (§6).

4 Experimental Setup

4.1 Data Collection

We collect pre-training data from two corpora: the English-to-Polish part of OpenSubtitles18 (Lison and Tiedemann, 2016) and the Europarl (Koehn, 2005) corpus. The data quantities can be found in Table 3 (column "`pretrain`").

		`pretrain`	`finetune`	`amb_test`
train	#sents	10.8M	2.9M	–
	#tokens	82.1M	26M	–
dev	#sents	3K	3.5K	–
	#tokens	23.3K	48.7K	–
test	#sents	–	3.5K	1K
	#tokens	–	47.7K	10.3K

Table 3: Quantities of unique data used for: model pre-training (`pretrain`), model fine-tuning (`finetune`) and the test set for calculation of restricted impact (`amb_test`). Values are averaged for source and target text.

Corpus Extraction for Fine-tuning We extract the fine-tuning data directly from the pre-training corpus; each sample is paired with an annotation of up to four types of attributes. For that purpose we implement a set of morphosyntactic rules for the Polish SpaCy model (Tuora and Kobyliński, 2019) which uses the Morfeusz2 morphological analyser (Kieras and Wolinski, 2017).[2] Since attribute annotations vary at sentence level, we produce sentence-level annotations (instead of word- or scene-level). For both speaker and interlocutor gender attributes, the masculine gender makes up over 60% of the corpus. Altogether, a total of 34.33% of the corpus marks at least one of the attributes. Figure 2 shows how linguistic categories contributed to extracting each attribute.

Similarly to Elaraby et al. (2018) and Gonen and Webster (2020), we observe that certain nouns marked as describing the speaker or interlocutor have a fixed gender irrespective of that person's gender and are therefore inadequate determinants of their gender (e.g. *coward* "tchórz" is always masculine). We could not find a reliable (complete nor heuristic) method to resolve this other than creating a "stopwords" list of all inflexible nouns. The process is now performed in two steps: we first extract a list of sentences containing gender-marked words and then filter out those that were selected based on our "stopwords" list of inflexible nouns.

We extract 223.0K noun-dependent sentences with 9K unique lemmatised nouns in the first pass, build the "stopwords" list of 6.8K words and end up with 67.3K sentences.

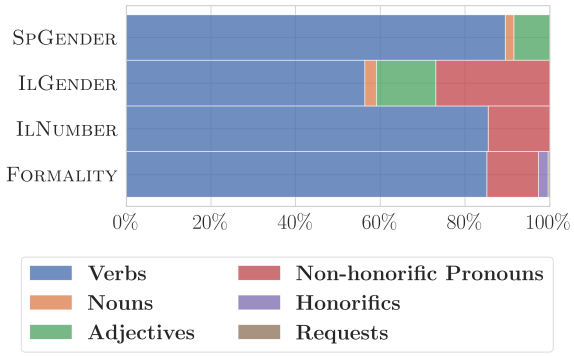

Figure 2: Contributions of each grammatical category to each attribute in the extracted corpus.

Parser Rules We identify sentences marking for SpGender by finding tokens in first person singular and verifying that their head marks feminine or masculine gender. Formality is identified through the use of the inflected pronouns in the *Pan+* set (unless it is used as a title, e.g. in *'Ms Smith'*). Formal requests are selected by finding *proszę* (*'please'*) in the target sentence but not in the source. IlGender is trivially inferred in formal cases; for informal language, we match structures analogous to those for the SpGender and extend them to comparative phrases and vocatives. IlNumber follows from the plurality of second-person verbs as well as the use of the pronoun *ty* (*'you'*, singular) or *wy* (*'you'*, plural).

Parser Performance To measure the effectiveness of the parser, a native Polish speaker with expertise in NLP manually annotated a random sample of 1K sentence pairs from the training corpus for the provided attribute types. Given a sample, the annotator was instructed to identify a type from each attribute, and then highlight a part of the Polish sentence proving its occurrence. Preci-

[2]The code is available at `https://github.com/st-vincent1/grammatical_agreement_eamt/`.

Count			Context				Example	
train	dev	test	SPGENDER	ILGENDER	ILNUMBER	FORMALITY	*English*	*Polish*
419.9K	0.8K	0.8K	*sp:feminine*	∗	∗	∗	I'm an amateur.	Jestem amator**ką**.
743.6K	0.8K	0.8K	*sp:masculine*	∗	∗	∗	I'm all alone.	Jestem całkiem sa**m**.
9.3K	0.2K	0.2K	∗	*il:feminine*	*plural*	*informal*	You're smitten.	Jeste**ście** odurz**one**.
73.8K	0.2K	0.2K	∗	*il:masculine*	*plural*	*informal*	Have you met Pete?	Pozna**liście** Pete'a?
315.9K	0.2K	0.2K	∗	×	*plural*	*informal*	You need to leave.	Musi**cie** wyjść.
326.8K	0.2K	0.2K	∗	×	*singular*	*informal*	I got you something.	Przyniosłem **ci** coś.
273.0K	0.2K	0.2K	∗	*il:feminine*	*singular*	*informal*	Are you sick?	Jeste**ś** cho**ra**?
498.7K	0.2K	0.2K	∗	*il:masculine*	*singular*	*informal*	Understand?	Zrozumia**łeś**?
0.7K	0.1K	0.1K	∗	*il:feminine*	*plural*	*formal*	Please, let me explain.	Wyjaśnię **paniom**.
2.7K	0.2K	0.2K	∗	*il:masculine*	*plural*	*formal*	Aren't you?	**Panowie** nie są?
5.7K	0.2K	0.2K	∗	*il:mixed*	*plural*	*formal*	You are wrong.	My**lą** się **państwo**.
63.0K	0.2K	0.2K	∗	*il:feminine*	*singular*	*formal*	Martini for you?	Dla **pani** martini?
144.0K	0.2K	0.2K	∗	*il:masculine*	*singular*	*formal*	Let me have your coat.	Wezmę **pański** płaszcz.
33.5K	0.2K	0.2K	∗	×	×	*formal*	Go ahead.	**Proszę** kontynu**ować**.

Table 4: Training data quantities for all combinations of contexts with examples for each combination, with relevant grammatical expressions highlighted. Since SPEAKER and INTERLOCUTOR contexts are always independent, the counts include cases where they co-occur. ∗ = this attribute *may* occur in this place; × = this attribute is never expressed within this category.

sion and recall scores were measured between the judgements of the parser and the annotator. The parser (hereinafter *Detector*) scored near-perfectly (**99.82%** precision and **99.17%** recall averaged over all attributes) and proved suitable for the tasks of both extracting the corpus and evaluating attribute controlling. Beyond input errors leading to incorrect parsing, we observed two consistent cases of failure:

- when the interlocutor is addressed in plural but is in fact singular (in cases like "Go$_{\text{singular}}$ help her. Maybe you [two] will$_{\text{plural}}$ figure it out together." the addressee may be interpreted as *plural* instead of *singular* depending on the majority of grammatical matches for each type);
- some tag questions (e.g. "prawda?") or expressions (e.g. the words "kimś" ('someone$_{\text{instr.}}$'), "czymś" ('something$_{\text{instr.}}$')) are consistently incorrectly analysed for dependencies, which sometimes leads to triggering of incorrect rules.

Data Selection and Annotation Table 4 shows particular groups of contexts, their typical expression, and total count in the corpus.[3] Similarly to Sennrich et al. (2016), we mask the annotations of half the training samples every epoch at random and give half of the unannotated sentence pairs a random set of attributes. This helps preserve the translation quality of the model's outputs when insufficient context is given.

Our development and test sets are balanced across the 14 context groups (cf. table 4). We gather a total of 4K unique examples for each set. When evaluating each implemented approach, we provide two results: when *complete context* is given, or when an *isolated attribute* type is provided. Consider a complete-context test case within the ILNUMBER group of

<*il:feminine*>,<*plural*>,<*formal*> I like you.

The input for the isolated attribute is as follows:

<*plural*> I like you.

that is, we omit all types but those belonging to the examined attribute. For the *complete context* case we provide the full input. To evaluate each individual type (e.g. <*il:feminine*> or <*formal*>), in the isolated attribute case we gather all development/test cases which match the selected type, with a total count of minimum 200 examples (for <*il:mixed*>) up to 1200 (for <*plural*>).

4.2 Model Settings

We use the Transformer architecture (Vaswani et al., 2017) implemented in PyTorch (Paszke et al., 2019). Similarly to Lakew et al. (2021), we test a range of model alterations.

We split them into two categories: Types as Tags (TAG*) and Embedded Types (EMB*). We scale each approach that was originally proposed as a way of controlling a single attribute to a multi-attribute scenario: for TAG*, we supply multiple tags in a random order, and for EMB* we average the embeddings (see Table 5).

[3]Note that ILGENDER, ILNUMBER, FORMALITY are co-dependent, since they all concern the same entity (the interlocutor), and thus different combinations of their types lead to different grammatical expressions.

Approach	Multi-attribute solution	Embedding size	Input space occupied
Types as Tags			
TAGENC▲ (Sennrich et al., 2016)	++	$n_{types} * d_{model}$	n_{types}
TAGDEC (Takeno et al., 2017)			$n_{types} + 1$
TAGENCDEC▲ (Lakew et al., 2021)			$2 * n_{types} + 1$
Embedded Types			
EMBPWSUM (Lakew et al., 2021)	$\frac{\sum types}{n_{types}}$	$n_{types} * d_{model}$	0
EMBADD (Schioppa et al., 2021)			0
EMBENC (Ours)			1
EMBSOS (Lample et al., 2019)			0
EMBENCSOS (Ours)			1
OUTBIAS▲ (Michel and Neubig, 2018)	$\frac{\sum types}{n_{types}}$	$n_{types} * len_{vocab}$	0

Table 5: Comparison of examined approaches. ++ = concatenation. ▲ = Approach originally proposed for single-attribute control and extended by us.

Types as Tags We encode each type of each attribute as a special vocabulary token (e.g. $<singular>$, cf. Table 2). During fine-tuning, these *tags* are concatenated to the source or target[4] sentences and trained like other tokens. We use three settings:

- TAGENC: appending the tags to the source sentence (Sennrich et al., 2016).
- TAGDEC: prepending the tag to the target sentence (Takeno et al., 2017).
- TAGENCDEC: applying tags to both sentences (Niu and Carpuat, 2020).

Average Embedding As an alternative to sequential tagging, embedded types T can be averaged and supplied as a single vector $\overline{E(T)}$ (Lample et al., 2019). We test five settings:

- EMBPWSUM: adding $\overline{E(T)}$ position-wise to each input token (Lakew et al., 2021).
- EMBADD: adding $\overline{E(T)}$ position-wise to Encoder outputs (Schioppa et al., 2021).
- EMBENC: concatenating $\overline{E(T)}$ to the input (cf. Dai et al. (2019), but in our approach the embedding is not trained adversarially).
- EMBSOS: replace the start-of-sequence ($<sos>$) token in the Decoder input with $\overline{E(T)}$ (Lample et al., 2019).
- EMBENCSOS: as an additional setting, we test combining EMBENC and EMBSOS.

As a special case, we test OUTBIAS: adding a type embedding as a bias on the final layer of the Decoder (Michel and Neubig, 2018). We omit the *black-box injection* method of Moryossef et al. (2019) due to its inapplicability to ILGENDER in plural and to FORMALITY. Our baseline is the pre-trained model without attribute information.

4.3 Training Details

We preprocess the corpus with Moses tools for detokenisation and normalising punctuation[5], and by running a short set of custom rules. We train a joint sub-word segmentation model of 16K tokens with SentencePiece (Kudo and Richardson, 2018) and encode both sides of the corpus. We follow the standard training regimen for a 6-layer Transformer (Vaswani et al., 2017) with an input length limit of 100 tokens; this model has just over 52.3M trainable parameters. All training is done on a single 32GB GPU. As the decoding algorithm, we use beam search with a beam size of 5. We pre-train the model until a patience criterion of the chrF++ (Popović, 2017) validation score not increasing for 5 consecutive validation steps (which occur every $3/4$th epoch). This happens around the 24th epoch, or after 66 hours of training.

Each of the nine architectural upgrades is a copy of the pre-trained model expanded with the relevant component and fine-tuned. The fine-tuning process exposes the model to the fine-tuning corpus in 10 epochs; performance is validated every half epoch. We select the best checkpoint based on the highest chrF++ score on the development set.

4.4 Evaluation

We consider the following criteria in evaluation:

1. **Translation Quality.** Attribute-controlled

[4]During inference, we supply tags by forcibly decoding the relevant type tokens, followed by a $<null>$ token, before the main decoding step commences.

[5]https://github.com/alvations/sacremoses

	isolated attribute			*complete context*			
Model	chrF++↑	BLEU↑	*Agree*↑ (%)	chrF++↑	BLEU↑	*Agree*↑ (%)	AMBID↑
Baseline	46.60	23.13	74.35	46.60	23.13	74.35	–
TAGENC	**48.95**	**25.52**	99.03	**52.41**	**29.16**	**99.39**	**95.87**
TAGDEC	48.65	**25.40**	99.21	50.83	27.65	96.84	93.15
TAGENCDEC	48.28	25.26	99.35	51.01	28.15	**99.26**	82.66
EMBPWSUM	46.03	22.37	100	51.90	28.69	97.90	88.67
EMBADD	47.45	23.61	**99.96**	51.77	28.56	98.24	87.76
EMBENC	47.72	24.39	83.42	52.23	**28.98**	**99.30**	**95.58**
EMBSOS	48.28	24.90	**99.91**	**52.38**	**29.09**	98.47	92.07
EMBENCSOS	48.60	25.08	**99.87**	51.94	28.77	98.55	92.37
OUTBIAS	48.59	24.98	96.71	49.32	26.11	86.25	94.05

Table 6: Translation performance of all models; "*isolated attribute*" means that only one (the investigated) attribute was revealed to the model. The highlighted scores include the best one in the column and all statistically equivalent results according to a bootstrap resampling method ($p < 0.05$).

translations should be of quality no worse than translations of the non-specialised model.

2. **Grammatical Agreement.** Attribute-controlled hypotheses should completely agree to the specified type where necessary.
3. **Restricted Impact.** Grammatical agreement should only affect words that explicitly render the attributes. Therefore, if no attribute is to be expressed in the hypotheses, then they should be no different from baseline hypotheses.

We evaluate translation quality with chrF++ (Popović, 2017)[6] and BLEU (Papineni et al., 2002). Grammatical agreement is quantified with the help of the *Detector*. For every attribute, we calculate how many hypotheses agree to the correct type t and to the incorrect type $\hat{t}$. Let hyp_t be a hypothesis translated using type t as context, and $agree(hyp, t)$ denote that the *Detector* has found evidence of type t expressed in hyp. We express the total agreement score as:

$$Agree = \frac{agree(hyp_t, t)}{agree(hyp_t, t) + agree(hyp_t, \hat{t})}$$

Finally, we quantify restricted impact with a custom metric, which measures that attribute-independent sentences do not carry any attribute-reliant artifacts; we define this metric, AMBID, as:

$$\text{chrF++}(\text{NMT}(src_a, A), \text{NMT}(src_a, \widehat{A}))$$

where A is a set of attribute types and $\widehat{A}$ is the reverse set.[7] We use an attribute-ambivalent test set of a 1K sentences to calculate this score (Table 3, column "`amb_test`").

[6]For clarity, we normalise chrF++ scores to a $[0, 100]$ range.

[7]For the type triplet ILGENDER we assume that *il:masculine* = *il:feminine*, *il:mixed* = *il:feminine*, *il:feminine* = *il:masculine*.

5 Results

We report quantitative results in Table 6.

Grammatical Agreement The *Agree* column in Table 6 shows the agreement scores given by the *Detector*. In the isolated attribute scenario, all methods but OUTBIAS and EMBENC achieve near-perfect agreement scores. The agreement scores in the *complete context* scenario remain high for other models except TAGDEC, and pick up for EMBENC, suggesting that controlling several attributes generally has no negative impact on individual attributes.

Translation Quality Attribute-controlling models achieve significant gains over baseline for both the isolated attribute and complete context scenarios, and the gains are consistently higher in the latter, suggesting that exposing the models to more context yields better translations. TAGENC achieves the highest improvement over the baseline in terms of chrF++/BLEU for complete context ($+5.81$ chrF++/$+6.03$ BLEU). The gains in translation quality are correlated with agreement scores, except for EMBPWSUM, for which the isolated attribute scenario leads to a near-perfect agreement but low quality scores. Further investigation shows that this model learned to overproduce context-sensitive words when given a context of only a subset of types (e.g. translating "you" as "I" to introduce SPGENDER marking), leading to high agreement scores but degradation in quality. This highlights the importance of pairing an accuracy measure with a translation quality metric.

To investigate how successful the models are at modelling each context group individually, we report the mean chrF++ scores obtained for each

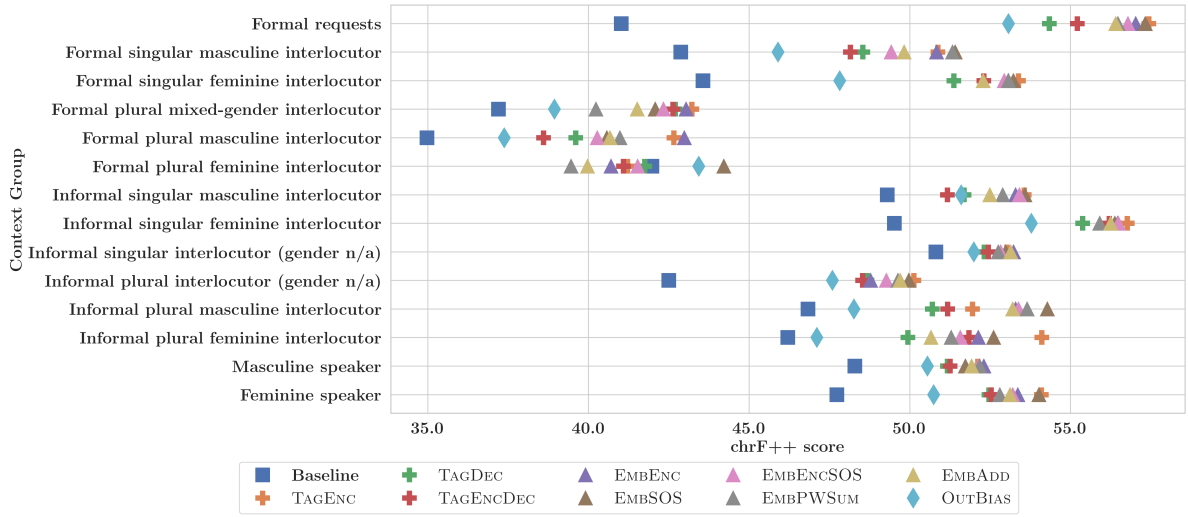

Figure 3: Translation quality (chrF++) for each contextual group.

group's test set (Figure 3). All contextual models bring significant improvements over the baseline except in the *Formal plural feminine interlocutor group*, for which there was little training data (cf. Table 4); improvements are consistently greater for feminine than masculine groups. No single model performs consistently better than others than others, but TagDec, EmbPWSum and OutBias fall behind on most groups. Finally, we observe no significant gain generally from including information in both the Encoder and the Decoder.

Restricted Impact The AmbID scores shown in Table 6 reveal that TagEnc and EmbEnc introduce the least variation in attribute-ambivalent utterances, suggesting that adding contextual information to the Encoder input only helps limit creation of unwanted artifacts. The distance of only 4.13 chrF++ points to the ideal score of 100 for the highest-scoring model suggests good separation of grammatical and behavioural agreement. Some separation-specific modelling may further improve this score, but it was outside the scope of this work.

General Discussion The results suggest that TagEnc is the most reliable approach to the presented problem, followed by EmbSOS and EmbEnc. Notably, we find other methods dubbed as superior to TagEnc in previous work (EmbAdd, TagDec and TagEncDec) to underperform in our case.

6 Conclusions and Future Work

In this work, we have highlighted the problem of grammatical agreement in translation of TV dialogue in the English-to-Polish language direction. We have created and described a dataset annotated for four speaker and interlocutor attributes that directly influence grammar in dialogue: speaker's gender, interlocutor's gender and number and formality relations between them. We have presented a selection of models capable of controlling these attributes in translation, yielding a performance gain of up to +5.81chrF++/+6.03BLEU over the baseline (non-controlling) model. Finally, we have produced a tool that produces an accuracy score for agreement to each type.

Considering all criteria of evaluation, we have identified TagEnc as the best performing approach, with EmbEnc, and EmbSOS also achieving competitive performance. TagEnc may be more attractive in scenarios where interventions in the model architecture are impossible as it can be implemented via data preprocessing alone, but the other two have a more scalable design (cf. §2). Finally, contrary to some previous work, we found no advantages stemming from including the contextual information in the Decoder as well as the Encoder.

Future Work NMT research should strive to move beyond seeing gender as a dichotomous phenomenon (Savoldi et al., 2021). Within this paper we did not consider the scenarios with non-binary interlocutors due to i) lack of available data and ii) lack of consensus regarding non-binary gender expression in the Polish language (Misiek, 2020). However, our work can be applied to non-binary expression once data and more studies are avail-

able. Furthermore, the influence in NMT of other extra-textual attributes (e.g. multimodal ones, like spatial information, or emergent ones, such as personal attributes) is yet to be explored. It remains an open question whether such attributes should all be considered individually, or whether there is a way of identifying and/or using them implicitly.

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

Ailem, Melissa, Jingshu Liu, and Raheel Qader. 2021. Encouraging neural machine translation to satisfy terminology constraints. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1450–1455, Online, August. Association for Computational Linguistics.

Bahdanau, Dzmitry, Kyung Hyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*, pages 1–15.

Dai, Ning, Jianze Liang, Xipeng Qiu, and Xuanjing Huang. 2019. Style transformer: Unpaired text style transfer without disentangled latent representation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5997–6007, Florence, Italy, July. Association for Computational Linguistics.

Elaraby, Mostafa, Ahmed Y. Tawfik, Mahmoud Khaled, Hany Hassan, and Aly Osama. 2018. Gender aware spoken language translation applied to English-Arabic. *2nd International Conference on Natural Language and Speech Processing, ICNLSP 2018*, pages 1–6.

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

Feldstein, Ron F. 2001. *A Concise Polish Grammar*. Slavic and East European Language Research Center (SEELRC), Duke University, 2001.

Gonen, Hila and Kellie Webster. 2020. Automatically identifying gender issues in machine translation using perturbations. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1991–1995, Online, November. Association for Computational Linguistics.

Jacovi, Alon, Ana Marasović, Tim Miller, and Yoav Goldberg. 2021. Formalizing Trust in Artificial Intelligence: Prerequisites, Causes and Goals of Human Trust in AI. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 624–635, New York, NY, USA. Association for Computing Machinery.

Jassem, Wiktor. 2003. Polish. *Journal of the International Phonetic Association*, 33(1):103–107.

Johnson, Melvin, 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.

Keown, Anne. 2003. Motivations for Polish pronouns of address. *Glossos*, 4(4).

Kieras, Witold and Marcin Wolinski. 2017. Morfeusz 2–analizator i generator fleksyjny dla jezyka polskiego. *Jezyk Polski*, 97(1):75–83.

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

Koniuszaniec, G and Hanka Błaszkowska. 2003. Language and gender in Polish. *Gender across Languages*, 3:259–285.

Kudo, Taku 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, November. Association for Computational Linguistics.

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

Lakew, Surafel M., Marcello Federico, Yue Wang, Cuong Hoang, Yogesh Virkar, Roberto Barra-Chicote, and Robert Enyedi. 2021. Machine translation verbosity control for automatic dubbing. *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, 2021-June:7538–7542.

Lample, Guillaume, Sandeep Subramanian, Eric Michael Smith, Ludovic Denoyer, Marc'Aurelio Ranzato, and Y. Lan Boureau. 2019. Multiple-attribute text rewriting. *7th International Conference on Learning Representations, ICLR 2019*, pages 1–20.

Lison, Pierre and Jörg Tiedemann. 2016. OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 923–929, Portorož, Slovenia, May. European Language Resources Association (ELRA).

Michel, Paul and Graham Neubig. 2018. Extreme adaptation for personalized neural machine translation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 312–318, Melbourne, Australia, July. Association for Computational Linguistics.

Misiek, Szymon. 2020. Misgendered in Translation?: Genderqueerness in Polish Translations of English-language Television Series. *Anglica. An International Journal of English Studies*, pages 165–185.

Moryossef, Amit, Roee Aharoni, and Yoav Goldberg. 2019. Filling gender & number gaps in neural machine translation with black-box context injection. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 49–54, Florence, Italy, August. Association for Computational Linguistics.

Niu, Xing and Marine Carpuat. 2020. Controlling neural machine translation formality with synthetic supervision. *AAAI 2020 - 34th AAAI Conference on Artificial Intelligence*, 2(1):8568–8575.

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, July. Association for Computational Linguistics.

Paszke, Adam, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. PyTorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems*, 32(NeurIPS).

Popović, Maja. 2017. chrF++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation*, pages 612–618, Copenhagen, Denmark, September. Association for Computational Linguistics.

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

Savoldi, Beatrice, Marco Gaido, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2021. Gender Bias in Machine Translation. *Transaction of the Association for Computational Lignuistics (TACL)*.

Schioppa, Andrea, David Vilar, Artem Sokolov, and Katja Filippova. 2021. Controlling Machine Translation for Multiple Attributes with Additive Interventions. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6676–6696, Online and Punta Cana, Dominican Republic, 11. Association for Computational Linguistics.

Sennrich, Rico, 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.

Sigurðsson, Halldór Ármann and Verner Egerland. 2009. Impersonal null-subjects in Icelandic and elsewhere*. *Studia Linguistica*, 63(1):158–185.

Stahlberg, Dagmar, F Braun, L Irmen, and Sabine Sczesny. 2007. Representation of the sexes in language. *Social Communication*, pages 163–187.

Stone, Gerald. 1977. Address in the Slavonic Languages. *The Slavonic and East European Review*, 55(4):491–505.

Sutskever, Ilya, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. *Advances in Neural Information Processing Systems*, 4(January):3104–3112.

Takeno, Shunsuke, Masaaki Nagata, and Kazuhide Yamamoto. 2017. Controlling Target Features in Neural Machine Translation via Prefix Constraints. *Afnlp*, pages 55–63.

Tuora, Ryszard and Łukasz Kobyliński. 2019. Integrating Polish Language Tools and Resources in Spacy. In *Proceedings of PP-RAI 2019 Conference*, pages 210–214.

Vanmassenhove, Eva, 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.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*, pages 5999–6009.