Gender in Danger?
Evaluating Speech Translation Technology on the MuST-SHE Corpus

Luisa Bentivogli[1*]**, Beatrice Savoldi**[2*]**, Matteo Negri**[1]**,**
Mattia Antonino Di Gangi[1,2]**, Roldano Cattoni**[1]**, Marco Turchi**[1]

[1]Fondazione Bruno Kessler
[2]University of Trento
Trento, Italy
{bentivo,negri,digangi,cattoni,turchi}@fbk.eu
beatrice.savoldi@unitn.it

Abstract

Translating from languages without productive grammatical gender like English into gender-marked languages is a well-known difficulty for machines. This difficulty is also due to the fact that the training data on which models are built typically reflect the asymmetries of natural languages, gender bias included. Exclusively fed with *textual data*, machine translation is intrinsically constrained by the fact that the input sentence does not always contain clues about the gender identity of the referred human entities. But what happens with speech translation, where the input is an *audio* signal? Can audio provide additional information to reduce gender bias? We present the first thorough investigation of gender bias in speech translation, contributing with: *i)* the release of a benchmark useful for future studies, and *ii)* the comparison of different technologies (cascade and end-to-end) on two language directions (English-Italian/French).

1 Introduction

With the exponential popularity of deep learning approaches for a great range of natural language processing (NLP) tasks being integrated in our daily life, the need to address the issues of *gender fairness*[1] and *gender bias* has become a growing interdisciplinary concern. Present-day studies on a variety of NLP-related tasks, such as sentiment analysis (Kiritchenko and Mohammad, 2018) coreference resolution (Rudinger et al., 2018; Webster et al., 2018; Zhao et al., 2018), visual semantic-role labeling (Zhao et al., 2017) or language modeling (Lu

et al., 2019), attest the existence of a systemic bias that reproduces gender stereotypes discriminating women. In translation-related tasks, gender bias arises from the extent through which each language formally expresses the female or male gender of a referred human entity. Languages with a grammatical system of gender, such as Romance languages, rely on a copious set of morphological (inflection) and syntactic (gender agreement) devices applying to numerous parts of speech (Hockett, 1958). Differently, English is a natural gender language that only reflects distinction of sex via pronouns, inherently gendered words (*boy, girl*) and exceptionally with marked nouns (*actor, actress*). For all the other indistinct neutral words, the gender of the referred entity – if available – is inferred from contextual information present in the discourse, e.g. he/she is a *friend*.

Nascent inquiries on machine translation (MT) pointed out that machines tend to reproduce the linguistic asymmetries present in the real-world data they are trained on. In the case of gender inequality, this is made apparent by the attribution of occupational roles from gender-neutral linguistic forms into marked ones, where MT often wrongly chooses male-denoting (pro)nouns, e.g. identifying *scientist*, *engineer* or *doctor* as men (Prates et al., 2018; Escudé Font and Costa-jussà, 2019). Failing to pick the appropriate feminine form is both a technical and an ethical matter: gender-related errors affect the accuracy of MT systems but, more significantly, a biased system can dangerously perpetuate the under-/misrepresentation of a demographic group (Crawford, 2017).

Previous studies accounting for MT systems' strengths and weaknesses in the translation of gender shed light on the problem but, at the same time, have limitations. On one hand, the existing evaluations focused on gender bias were largely conducted on challenge datasets, which are controlled

---

artificial benchmarks that provide a limited perspective on the extent of the phenomenon and may force unreliable conclusions (Prates et al., 2018; Cho et al., 2019; Escudé Font and Costa-jussà, 2019; Stanovsky et al., 2019). On the other hand, the natural corpora built on conversational language that were used in few studies (Elaraby et al., 2018; Vanmassenhove et al., 2018) include only a restricted quantity of not isolated gender-expressing forms, thus not permitting either extensive or targeted evaluations. Moreover, no attempt has yet been made to assess if and how *speech* translation (ST) systems are affected by this particular problem. As such, whether ST technologies that leverage audio inputs can retrieve useful clues for translating gender in addition to contextual information present in the discourse, or supply for their lack, remains a largely unexplored question. In the light of above, the contributions of this paper are:

(1) We present the first systematic analysis aimed to assess ST performance on gender translation. To this aim, we compare the state-of-the-art cascaded approach with the emerging end-to-end paradigm, investigating their ability to properly handle different categories of gender phenomena.

(2) We publicly release MuST-SHE,[2] a multilingual, natural benchmark allowing for a fine-grained analysis of gender bias in MT and ST. MuST-SHE is a subset of the TED-based MuST-C corpus (Di Gangi et al., 2019a) and is available for English-French and English-Italian.[3] For each language pair, it comprises ∼1,000 (*audio*, *transcript*, *translation*) triplets annotated with qualitatively differentiated and balanced gender-related phenomena.

(3) We implement a new evaluation method that acknowledges and adapts previous related works to go beyond them and make BLEU scores informative about gender. It removes unrelated factors that may affect the overall performance of a system to soundly estimate gender bias.

On the two language pairs addressed, our comparative evaluation of cascade *vs.* end-to-end ST systems indicates that the latter are able to better exploit audio information to translate specific gender phenomena, for which the cascade systems require externally-injected information.

---

[2]MuST-SHE is released under a CC BY NC ND 4.0 International license, and is freely downloadable at `ict.fbk.eu/must-she`.

[3]The current release of the corpus also includes an English-Spanish section, which was completed in April 2020.

2 Background

Speech translation. The task of translating audio speech in one language into text in another language has been traditionally approached with cascade architectures combining automatic speech recognition (ASR) and MT components (Eck and Hori, 2005). The main advantage of this pipelined solution is that it can directly plug-in state-of-the-art technology for both components and exploit the wealth of training data available for the two tasks. The approach, however, has some drawbacks. One is error propagation: sub-optimal transcriptions by the ASR component have significant impact on the final output produced by the MT component. To cope with this issue, recent works focused on making MT models more robust to noisy input transcripts (Sperber et al., 2017, 2019; Di Gangi et al., 2019b).

A second issue, particularly relevant to this research, is the information loss when passing from audio to text representations. Even with perfect transcripts, subtle aspects that cannot be grasped from the text only (e.g. speaker's pitch as a clue of his/her gender) can only be reintroduced by injecting external knowledge to support the MT step (Elaraby et al., 2018). By avoiding intermediate text representations, direct end-to-end translation from audio to text (Bérard et al., 2016) can potentially cope with these limitations. However, due to the dearth of training corpora, it still underperforms with respect to the cascaded approach. Recent evaluation campaigns (Niehues et al., 2018, 2019) have shown that, although the gap is gradually closing (less than 2 BLEU points), cascade models still represent the state-of-the-art. In spite of the steady technological progress, little has so far been done to directly compare the two technologies on specific translation problems like the one addressed in this paper.

Measuring gender bias. Previous attempts to test the production of gender-aware automatic translations solely focused on MT, where a widespread approach involves the creation of challenge datasets focused on specific linguistic phenomena. Prates et al. (2018) and Cho et al. (2019) construct template sentences using occupational or sentiment words associated with a gender-neutral pronoun, to be translated into an English gender-specified one (*[x] is a professor: he/she is a professor*). Similarly, the Occupations Test (Escudé Font and Costa-jussà, 2019) and Wino MT (Stanovsky et al., 2019)

cast human entities into proto- or anti-stereotypical gender associations via coreference linking (e.g. the English sentence *"The janitor does not like **the baker** because <u>she</u>/<u>he</u> always messes up the kitchen"*, where *"the baker"* is to be translated into Spanish as **la panadera** or **el panadero** depending on the English pronoun). Although such simple constructions allow for targeted experiments, artificial data characterized by a qualitatively limited variety of phenomena generate constrained environments that may produce biased results. As far as studies on naturally occurring data are concerned, Vanmassenhove et al. (2018) estimate MT systems' performance in the realization of speaker's gender agreement on two male and female test sets containing first person singular pronouns. This strategy increases the chances to isolate speaker-dependent gendered expressions, but still, the employed BLEU metric does not pointedly grasp the effect of gender translation on the output, as the overall performance is also impacted by other factors. Analogously, Elaraby et al. (2018) design a set of agreement rules to automatically recover 300 gender-affected sentences in their corpus, but the evaluation relies on global BLEU scores computed on a bigger set (1,300 sentences) and does not consider male-female related differences. Moryossef et al. (2019) use a parser to detect morphological realizations of speakers' gender on a single female-speaker corpus that does not permit inter-gender comparisons.

In light of above, an ideal test set should consist of naturally occurring data exhibiting a diversified assortment of gender phenomena so to avoid forced predictions with over-controlled procedures. Also, a consistent amount of equally distributed feminine and masculine gender realizations need to be identified to disentangle the accuracy of gender translation from the overall model's performance. Accordingly, in §3 we present MuST-SHE, a multilingual test set designed for the investigation of gender bias in ST, which, as explained in §4, is used for a targeted gender-sensitive evaluation approach.

3   The MuST-SHE benchmark

We built MuST-SHE on naturally occurring data retrieved from MuST-C (Di Gangi et al., 2019a), the largest freely available multilingual corpus for ST, which comprises (*audio, transcript, translation*) triplets extracted from TED talks data. Besides being multilingual, MuST-C is characterized by high-quality speech and a variety of different speakers that adequately represent women, two aspects that determined its selection among other existing corpora (Post et al., 2013; Kocabiyikoglu et al., 2018; Sanabria et al., 2018). As such, MuST-SHE was compiled by targeting in the original dataset linguistic phenomena that entail a gender identification from English into Italian and French, two Romance languages that extensively express gender via feminine or masculine morphological markers on nouns, adjectives, verbs and other functional words (e.g. articles and demonstratives).

3.1   Categorization of gender phenomena

MuST-SHE is compiled with segments that require the translation of at least one English gender-neutral word into the corresponding masculine or feminine target word(s), where such formal distinction semantically conveys and conflates with an actual distinction of sex (Corbett, 1991). For instance, the English utterance "*a good teacher*" would either become in French "*un bon enseignant*" or "*une bonne enseignante*" for, respectively, a male or female referent. In spoken language data, the human entity that determines gender agreement is either the speaker him/herself (*I am a good teacher*) or another person the speaker is referring to (*he/she is a good teacher*). We classify our phenomena of interest in two categories based on where the necessary information to disambiguate gender can be recovered, namely (**Category 1**) from the audio signal, when gender-agreement only depends on the speaker's gender, which can be captured from intrinsic properties of the audio (*I am a teacher* uttered by a man/woman); (**Category 2**) from the utterance content, where contextual hints such as gender-exclusive words (*mom*), pronouns (*she, his*) and proper nouns (*Paul*) inform about the gender of the referent.

3.2   Dataset creation and annotation

To gain a better insight into MuST-C linguistic data and capture the features of gender, we initially conducted a qualitative cross-lingual analysis on 2,500 parallel sentences randomly sampled from the corpus. The analysis led to the design of an automatic approach aimed to quantitatively and qualitatively maximize the extraction of an assorted variety of gender-marked phenomena belonging to categories 1 and 2. Regular expressions were employed to transform gender-agreement rules into search patterns to be applied to MuST-C. Our queries were

Form		Category 1: *Gender info in audio*	Speaker
Fem.	SRC	I was **born** and **brought up** in Mumbai.	Female
	C-REF$_{It}$	Sono **nata** e **cresciuta** a Mumbai.	
	W-REF$_{It}$	Sono **nato** e **cresciuto** a Mumbai.	
	SRC	I was **born** and brought up in Mumbai.	
	C-REF$_{Fr}$	Je suis **née** et j'ai grandi à Mumbai.	
	W-REF$_{Fr}$	Je suis **né** et j'ai grandi à Mumbai.	
Masc.	SRC	I **myself** was **one** of them, and this is what I talk about at the HALT events.	Male
	C-REF$_{It}$	Io **stesso** ero **uno** di loro, e parlo di questo agli eventi HALT.	
	W-REF$_{It}$	Io **stessa** ero **una** di loro, e parlo di questo agli eventi HALT.	
	SRC	I myself was **one** of them, and this is what I talk about at the HALT events.	
	C-REF$_{Fr}$	Moi-même, j'ai été l'**un** d'eux, et voilà de quoi je parle aux événements d'HALT.	
	W-REF$_{Fr}$	Moi-même, j'ai été l'**une** d'eux, et voilà de quoi je parle aux événements d'HALT.	
		Category 2: *Gender info in utterance content*	
Fem.	SRC	She'd get together with two **of her dearest friends**, **these older** <u>women</u>...	Male
	C-REF$_{It}$	Tornava per incontrare un paio **delle sue** più **care amiche**, **queste** signore **anziane**...	
	W-REF$_{It}$	Tornava per incontrare un paio **dei suoi** più **cari amici**, **questi** signore **anziani**...	
	SRC	She'd get together with two of her **dearest friends**, these **older** <u>women</u>...	
	C-REF$_{Fr}$	Elle se réunissait avec deux de ses **amies** les plus **chères**, ces femmes plus **âgées**...	
	W-REF$_{Fr}$	Elle se réunissait avec deux de ses **amis** les plus **chers**, ces femmes plus **âgés**...	
Masc.	SRC	<u>Dean</u> Kamen, **one of the** great DIY **innovators**. <u>His</u> technology...	Female
	C-REF$_{It}$	Dean Kamen, **uno dei** più grandi **innovatori** del fai-da-te. La sua tecnologia...	
	W-REF$_{It}$	Dean Kamen, **una delle** più grandi **innovatrici** del fai-da-te. La sua tecnologia...	
	SRC	<u>Dean</u> Kamen, **one** of the **great** DIY **innovators**. <u>His</u> technology...	
	C-REF$_{Fr}$	Dean Kamen, l'**un** des **grands innovateurs** autonomes. Sa technologie...	
	W-REF$_{Fr}$	Dean Kamen, l'**une** des **grandes innovatrices** autonomes. Sa technologie...	

Table 1: MuST-SHE annotated segments organized per category. For each example in En-It and En-Fr, the Correct Reference Translation (C-REF) shows the realization of target gender-marked forms (Masc/Fem) corresponding to English gender-neutral words in the source (SRC). In the Wrong Reference Translation (W-REF), Italian and French gender-marked words are swapped to their opposite gender form. The last column of the table provides information about the speaker's gender (Male/Female).

designed and adapted to the targeted language pairs, categories, and masculine/feminine forms. To specifically match a differentiated range of gender-marked lexical items, we also compiled two series of 50 human-referring adjectives in French and Italian, as well as a list with more than 1,000 English occupation nouns obtained from the US Department of Labour Statistics[4] (Prates et al., 2018).

For each language direction, the pool of sentence pairs retrieved from MuST-C was manually checked in order to: *i)* remove noise and keep only pairs containing at least one gender phenomenon, *ii)* include all En-It/En-Fr corresponding pairs to create a common multilingual subset, and *iii)* select the remaining pairs ensuring a balanced distribution of categories, feminine/masculine forms, and female/male speakers. Once the textual part of MuST-SHE was created, all the corresponding au-

dio segments were manually checked in order to correct possible misalignments.

The resulting dataset was then manually enriched with different types of information that allow for fine-grained evaluations. Annotations include: category, masculine/feminine form, speaker's gender, and all the gender-marked expressions in the reference translation. Finally, in order to perform a sound evaluation able to discriminate gender-related issues from other non-related factors that may affect systems' performance, for each correct reference translation (C-REF) we created an almost identical "wrong" alternative (W-REF) in which all the gender-marked words are swapped to their opposite form (details in §4). Some examples extracted from MuST-SHE are presented in Table 1.

To ensure data quality, the whole dataset was created and annotated by an expert linguist with a background in translation studies, who produced strict and comprehensive guidelines based on the

preliminary manual analysis of a sample of MuST-C data (2,500 segments). Then, a second linguist independently re-annotated each MuST-SHE segment with the corresponding category and produced an additional "wrong" reference. Being the annotation per category a straightforward task, it resulted in no disagreement for Category 1 and around 0.03% for Category 2. Such few cases were removed from the dataset, which thus contains only segments in complete agreement. Disagreements were more common in the "wrong" references, since the task requires producing subtle variations that can be hard to spot. Disagreements, amounting to around 11%, were all oversights and thus reconciled.

3.3 Dataset statistics

MuST-SHE comprises 2,136 (*audio*, *transcript*, *translation*) triplets (1,062 for En-It and 1,074 for En-Fr) uttered by 273 different speakers. A common subset of 696 instances allows for comparative evaluations across the two language directions. As shown by the statistics in Table 2, the corpus presents a balanced distribution across *i)* masculine and feminine forms, and *ii)* gender phenomena per category. Female and male speakers (558/513 for En-It, 577/498 for En-Fr) are substantially balanced. The gender of the speaker and of the referred entity in the utterance is the same in Category 1 (where the speakers talk about themselves), while it differs in about 50% of the segments in Category 2 (where they refer to other entities).

MuST-SHE differs from standard test sets, as it is precisely designed to: *i)* equally distribute gender references as well as speakers, and *ii)* allow for a sound and focused evaluation on the accuracy of gender translation. As such, it satisfies the parameters to be qualified as a GBET, *Gender Bias Evaluation Testset* (Sun et al., 2019), and represents the very first of its kind for ST and MT created on natural data.

4 Experimental Setting

4.1 Evaluation Method

MT evaluation metrics like BLEU (Papineni et al., 2002) or TER (Snover et al., 2006) provide a global score about translation "quality" as a whole. Used as-is, their holistic nature hinders the precise evaluation of systems' performance on an individual phenomenon as gender translation, since the variations of BLEU score are only a coarse and indi-

	En-It			**En-Fr**		
	Fem	Masc	*Tot.*	Fem	Masc	*Tot.*
Cat. *1*	278	282	*560*	316	296	*612*
Cat. *2*	238	264	*502*	226	236	*462*
Tot.	*516*	*546*		*542*	*532*	
Total	**1,062** (1,940)			**1,074** (2,010)		

Table 2: MuST-SHE statistics. En-It and En-Fr number of segments split into feminine and masculine gender phenomena and category. In parentheses the total number of gender-marked words annotated in the references.

rect indicator of better/worse overall performance (Callison-Burch et al., 2006). This represents a limitation of recent related works, which over-rely on the results of a BLEU-based quantitative analysis. For instance, the BLEU gains obtained by prepending gender tags or other artificial antecedents to the input source, as in Vanmassenhove et al. (2018) and Moryossef et al. (2019), cannot be assuredly ascribed to a better control of gender features. To overcome this problem, Moryossef et al. (2019) complement their discussion with a qualitative syntactic analysis, which implies the availability of a parser for the target language and a higher complexity of the whole evaluation protocol. Instead, our aim is to keep using BLEU[5] and make the resulting scores informative about systems' ability to produce the correct gender forms.

To this aim, for each reference c in the corpus we create a "wrong" one that is identical to c, except for the morphological signals that convey gender agreement. In particular, for each gender-neutral English word in the source utterance (e.g. *"one"*, *"great"* and *"innovators"* in the 4th example of Table 1), the correct translation (containing the French words with masculine inflection *"un"*, *"grands"* and *"innovateurs"*) is swapped into its opposite gender form (containing feminine-marked words *"une"*, *"grandes"* and *"innovatrices"*). The result is a new set of references that, compared to the correct ones, are "wrong" only with respect to the formal expression of gender.

The underlying idea is that, as the two reference sets differ only for the swapped gendered forms, results' differences for the same set of hypotheses produced by a given system can measure its capability to handle gender phenomena. In partic-

---

[5]Still the *de facto* standard in MT evaluation in spite of constant research efforts towards metrics that better correlate with human judgements.

ular, we argue that higher values on the wrong set can signal a potentially gender-biased behaviour. In all the cases where the required gender realization is *feminine*, significantly higher BLEU results computed on the wrong set would signal a bias towards producing masculine forms, and vice versa. Although this idea recalls the gender-swapping approach used in previous NLP studies on gender bias (Sun et al., 2019; Lu et al., 2019; Kiritchenko and Mohammad, 2018; Zhao et al., 2018; Cao and Daumé III, 2019), in such works it is only applied to pronouns; here we extend it to any gender-marked part of speech.

In addition to the quantitative BLEU-based evaluation[6], we also perform a fine-grained qualitative analysis of systems' accuracy in producing the target gender-marked words. We compute accuracy as the proportion of gender-marked words in the references that are correctly translated by the system. An upper bound of one match for each gender-marked word is applied in order not to reward over-generated terms. Besides global accuracy, we also compute scores on both the correct and the wrong reference sets, as well as per category.

It's worth remarking that the BLEU-based and the accuracy-based evaluations are complementary: the former aims to shed light on system's translation performance with respect to gender phenomena; the latter, which is more discriminative, aims to point to the actual words through which gender is realized. Compared to the standard BLEU-based evaluation with correct references only, we expect that the possible differences suggested by its extension with gender swapping will be reflected and amplified by sharper accuracy differences.

4.2 ST Systems

In our experiments, we compare an `End2End` system with two cascade systems (`Cascade` and `Cascade+tag`), whose architectures are described below.

Our **End2End** system uses the S-transformer architecture, which has proved to work reasonably well for this task (Di Gangi et al., 2019c). It is an encoder-decoder architecture that modifies the Transformer architecture (Vaswani et al., 2017) in two aspects. First, the audio input – in the form of sequences of 40 MFCCs (Davis and Mermelstein, 1980) – is processed by a stack of 2D CNNs (LeCun

et al., 1998), each followed by batch normalization (Ioffe and Szegedy, 2015) and ReLU nonlinearity. Second, the output of the CNNs is processed by 2D self-attention networks to provide a larger context to each element. The output of the 2D attention is then summed with the positional encoding and fed to transformer encoder layers. In the second part, a distance penalty is added to the non-normalized probabilities in the encoder self-attention networks in order to bias the computation towards the local context. To improve translation quality, the `End2End` systems are trained on the MuST-C and Librispeech (Kocabiyikoglu et al., 2018) corpora using SpecAugment (Park et al., 2019). Since Librispeech is a corpus for ASR, we augmented it by automatically translating the original English transcripts into both target languages. Translations are performed at character level, using the MT systems integrated in the cascade model.

Our **Cascade** systems share the same core (ASR, MT) technology. The ASR component is based on the KALDI toolkit (Povey et al., 2011), featuring a time-delay neural network and lattice-free maximum mutual information discriminative sequence-training (Povey et al., 2016). The audio data for acoustic modeling include the clean portion of LibriSpeech (Panayotov et al., 2015) (∼460h) and a variable subset of the MuST-C training set (∼450h), from which 40 MFCCs per time frame were extracted; a MaxEnt language model (Alumäe and Kurimo, 2010) is estimated from the corresponding transcripts (∼7M words). The MT component is based on the Transformer architecture, with parameters similar to those used in the original paper. The training data are collected from the OPUS repository,[7] resulting in 70M pairs for En-It and 120M for En-Fr. For each language pair, the MT system is first trained on the OPUS data and then fine-tuned on MuST-C training data (∼250K pairs) – from which the MuST-SHE segments are removed. Byte pair encoding (BPE) (Sennrich et al., 2015) is applied to obtain 50K sub-word units. To mitigate error propagation and make the MT system more robust to ASR errors, similarly to (Di Gangi et al., 2019b) we tune it on a dataset derived from MuST-C, which includes both human and automatic transcripts. The training set, consisting of (*audio*, *transcript*) pairs, is split in two equally-sized parts: the first one is used to adapt the ASR system to the TED talk language, while

---

[6]We also computed TER scores and the results are fully in line with the reported BLEU scores.

[7]http://opus.nlpl.eu

	Systems	All			Feminine			Masculine		
		Correct	Wrong	Diff	Correct	Wrong	Diff	Correct	Wrong	Diff
En-It	End2End	21.5	19.7	1.8	20.2	19.3	0.9	22.7	20.0	2.7
	Cascade	24.1	22.4	1.8	22.8	21.9	0.8	25.5	22.8	2.7
	Cascade+Tag	23.8	20.9	2.9	23.0	20.4	2.6	24.5	21.3	3.2
En-Fr	End2End	27.9	25.8	2.1	26.3	25.0	1.3	29.5	26.4	3.1
	Cascade	32.2	30.1	2.1	30.4	29.4	1.0	33.8	30.8	3.0
	Cascade+Tag	32.2	28.6	3.6	31.6	28.0	3.6	32.7	29.2	3.5

Table 3: BLEU scores for En-It and En-Fr on MuST-SHE. Results are provided for the whole dataset (All) as well as split according to feminine and masculine word forms. Results are calculated for both the *Correct* and *Wrong* datasets, and their difference is provided (Diff).

the second part is transcribed by the tuned ASR system. The human transcripts of the first half and the automatic transcripts of the second half are concatenated and used together with their reference translations to fine-tune the MT system. This process makes the MT system aware of possible ASR errors and results in more than 2 BLEU points improvement on the MuST-C test set.

We also train an enhanced version of the Cascade system. Similarly to Vanmassenhove et al. (2018), it is informed about speaker's gender by pre-pending a gender token (<toM> or <toF>) to each source transcript. The gender token is obtained by manually assigning the correct gender label to each speaker in MuST-C. This externally-injected knowledge allows the Cascade+Tag system to mimic end-to-end technology by leveraging gender information during translation.

To check the overall quality of our systems, we compared them with published results on MuST-C test data. Our End2End systems (En-It: 21.5, En-Fr: 31.0) outperform all the models proposed in Di Gangi et al. (2019c), which were trained only on MuST-C (En-It: end2end 16.8, cascade 18.9; En-Fr: end2end 26.9, cascade 27.9). Our Cascade (En-It: 27.4 En-Fr: 35.5) also outperforms the system described in Indurthi et al. (2019) (En-Fr: 33.7). Our results are in line with the findings of IWSLT 2019 (Niehues et al., 2019) showing that the cascade approach still outperforms the direct one, although with a gradually closing gap.

5 Results and Discussion

BLEU. Table 3 presents translation results in terms of BLEU score on the MuST-SHE dataset. Looking at overall translation quality (*All/Correct* column), the results on both language pairs show that the highest performance is achieved by cascade architectures, which are better than End2End by 2.6 points for En-It and 4.3 for En-Fr. We do not observe a statistically significant difference between Cascade and Cascade+Tag, suggesting that the injection of gender information into Cascade+Tag does not have visible effects in terms of translation quality, even on a focused dataset like MuST-SHE where each segment contains at least one gender realization. Our results thus seem to be in contrast with previous works implementing the same injection approach (Vanmassenhove et al., 2018; Elaraby et al., 2018).

However, looking at the scores' gap between the *Correct* and the *Wrong* datasets (*All/Diff* column), it becomes evident that the standard evaluation based on BLEU calculated on a single correct reference hides specific relevant aspects in translation. In fact, despite the lower overall BLEU scores, for both language pairs End2End performs on par with Cascade as far as gender phenomena are concerned (1.8 on En-It and 2.1 on En-Fr). Also, the largest *All/Diff* value achieved by the enhanced Cascade+Tag supports the results obtained in previous studies (Vanmassenhove et al., 2018; Elaraby et al., 2018), confirming the importance of applying gender-swapping in BLEU-based evaluations focused on gender translation.

The fact that the *All/Diff* values are always positive indicates that all the systems perform better on the *Correct* dataset (i.e. they generate the correct gender-marked words more often than the wrong ones). However, examining the results at the level of masculine/feminine word forms, we notice that *Diff* values are higher on the *Masculine* subset (where the required gender realization is masculine) than in the *Feminine* one (where the required gender realization is feminine). As discussed in §4.1, this signals a bias of the systems towards producing masculine forms. The only exception is the En-Fr Cascade+Tag, where the *Diff* values remain stable across the two subsets (3.6 and 3.5). This absence of bias towards the masculine forms

	Systems	All			Feminine			Masculine		
		Correct	Wrong	Diff	Correct	Wrong	Diff	Correct	Wrong	Diff
En-It	End2End	43.3	16.4	26.9	34.2	24.0	10.2	51.3	9.6	41.7
	Cascade	41.1	17.5	23.6	33.7	24.5	9.2	47.6	11.2	36.4
	Cascade+Tag	48.0	10.4	37.6	44.7	14.0	30.7	51.0	7.2	43.8
En-Fr	End2End	46.0	19.0	27.0	35.8	25.0	13.8	55.3	13.8	41.5
	Cascade	49.6	20.5	29.1	39.6	26.2	13.4	58.7	15.2	43.5
	Cascade+Tag	57.2	11.3	45.9	53.8	11.8	42.0	60.3	10.7	49.6

Table 4: Accuracy scores for En-It and En-Fr on MuST-SHE. Results are provided for the whole dataset (All) as well as split according to feminine and masculine word forms. Results are calculated for both the *Correct* and *Wrong* datasets, and their difference is provided (Diff).

	En-it							En-Fr					
	Feminine			Masculine				Feminine			Masculine		
	End2End							End2End					
	Corr.	Wrong	Diff.	Corr.	Wrong	Diff.		Corr.	Wrong	Diff.	Corr.	Wrong	Diff.
Cat. 1	26.7	27.2	-0.5	46.3	6.8	39.5		25.4	29.5	-4.1	48.0	7.7	40.3
Cat. 2	40.6	20.5	20.1	53.9	10.9	43.0		45.0	20.3	24.7	60.0	17.6	42.4
	Cascade							Cascade					
Cat. 1	15.9	34.5	-18.6	40.0	12.0	28.0		20.4	37.5	-17.1	49.1	13.0	36.1
Cat. 2	48.9	15.7	33.2	51.2	10.8	40.4		56.3	15.6	40.7	64.9	16.7	48.2
	Cascade+Tag							Cascade+Tag					
Cat. 1	43.0	10.4	32.6	48.5	2.9	45.6		53.7	7.0	46.7	55.4	4.3	51.1
Cat. 2	46.9	15.5	31.4	51.7	9.6	42.1		54.2	14.8	39.4	62.8	15.5	47.3

Table 5: Accuracy scores for En-It and En-Fr split according to MuST-SHE categories (Cat 1: information in audio, Cat 2: information in utterance content). For each category, results are further split into masculine/feminine forms. Results are calculated for both the *Correct* and *Wrong* datasets, and their difference is provided (Diff).

is in line with the *All/Diff* results indicating that this system is the best one in translating gender.

Although our gender-swapping methodology allows us to measure differences across systems that cannot be observed with standard BLEU evaluations, the results obtained so far may still conceal further interesting differences. This can depend on the fact that BLEU works at the corpus level and the small proportion of gender-marked words in MuST-SHE ($\sim$2,000 out of $\sim$ 30,000 total words, avg. 1.8 per sentence) can have limited influence on global measurements. To dig into these aspects, our final analysis relies on accuracy, which is exclusively focused on gender-marked words.

Accuracy. The results shown in Table 4 are not only consistent with the BLEU ones, but also highlight differences that were previously indistinguishable. While the *All/Diff* BLEU results for End2End and Cascade were identical on both languages, the *All/Diff* accuracy scores show that, although End2End performs better than Cascade for En-It, it performs worse for En-Fr. Also, with regards to Cascade+Tag, we can see that the *Diff* value is higher on the *Masculine* subset, thus showing that also this system is affected by gender bias, although to a lesser extent.

We now focus on systems' results on the two

categories represented in MuST-SHE: Category 1, where the information necessary to disambiguate gender can be recovered from the audio (speaker talking about him/herself) and Category 2, where such information occurs in the utterance content (speaker talking about someone else). Results are shown in Table 5.

As for **Category 1**, *Diff* values show that Cascade performance is the worst on both languages. This is due to the fact that its MT component cannot access the speaker's gender information necessary for a correct translation. This weakness becomes particularly evident in the *Feminine* class, where the higher values on the *Wrong* datasets (leading to negative values in columns *Feminine/Diff*) highlight a strong bias towards producing masculine forms. Although still negative for the Feminine class, the much better *Diff* values obtained by End2End show its ability to leverage audio features to correctly translate gender. However, the gap with respect to Cascade+Tag – by far the best system in Cat. 1 – is still large. On one side, End2End might benefit from better audio representations. Indeed, as shown in Kabil et al. (2018), the MFCC features used by state-of-the-art models are not the most appropriate for gender recognition. On the other side, Cascade+Tag

does not only take advantage of huge amounts of data to train its basic components, but it is also an oracle supported by the artificial injection of correct information about speakers' gender.

In **Category 2**, where having direct access to the audio is not an advantage since gender information is present in the textual transcript, results show a different scenario. While scores on the *Masculine* class are not conclusive across languages, on the *Feminine* class `End2End` always shows the worst performance. This can be explained by the fact that, being trained on a small fraction of the data used by the cascade systems, `End2End` is intrinsically weaker and more prone to gender mistranslations. Also, it is noticeable that `Cascade+Tag` is slightly worse than `Cascade`, although the MT components are trained on the same amount of data. This is due to the dataset design choice (see §3.3) to include ∼50% of segments where the speaker's gender does not agree with the gender of the phenomenon to translate. This feature makes MuST-SHE particularly challenging for systems like `End2End` and `Cascade+Tag` since, in these specific cases, speaker's gender information (extracted from the source audio or artificially injected) is not relevant and can introduce noise.

All in all, translating gender is still an issue in ST and current technologies are affected by gender bias to variable extent. Through the analysis made possible by MuST-SHE, we have been able to pinpoint their specific strengths and weaknesses and pave the way for more informed future studies.

6 Conclusion

If, like human beings, "machine learning is what it eats", the different "diet" of MT and ST models can help them to develop different skills. One is the proper treatment of gender, a problem when translating from languages without productive grammatical gender into gender-marked ones. With respect to this problem, by eating parallel texts during training, MT performance is bounded by the statistical patterns learned from written material. By eating (*audio*, text) pairs, ST has a potential advantage: the possibility to infer speakers' gender from input audio signals. We investigated for the first time the importance of this information in ST, analysing the behaviour of cascade (the state of the art in the field) and end-to-end ST technology (the emerging approach). To this aim, we created MuST-SHE, a benchmark annotated with different types of gender-related phenomena in two language directions. Our evaluation shows that, in spite of lower overall performance, the direct approach can actually exploit audio information to better handle speaker-dependent gender phenomena. These are out of reach for cascade solutions, unless the MT step is supplied with external (not always accessible) knowledge about the speaker. Back to our title: if, in ST, gender is still in danger, we encourage our community to start its rescue from MuST-SHE and the findings discussed in this paper.

References

Tanel Alumäe and Mikko Kurimo. 2010. Efficient Estimation of Maximum Entropy Language Models with n-gram Features: an SRILM Extension. In *Proceedings of INTERSPEECH 2010*, pages 1820–1823, Chiba, Japan.

Alexandre Bérard, Olivier Pietquin, Laurent Besacier, and Christophe Servan. 2016. Listen and Translate: A Proof of Concept for End-to-End Speech-to-Text Translation. In *Proceedings of the NIPS Workshop on end-to-end learning for speech and audio processing*, Barcelona, Spain. NIPS.

Chris Callison-Burch, Miles Osborne, and Philipp Koehn. 2006. Re-evaluating the Role of BLEU in Machine Translation Research. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics*, Trento, Italy. Association for Computational Linguistics.

Yang Trista Cao and Hal Daumé III. 2019. Toward Gender-Inclusive Coreference Resolution. *arXiv preprint arXiv:1910.13913*.

Won Ik Cho, Ji Won Kim, Seok Min Kim, and Nam Soo Kim. 2019. On Measuring Gender Bias in Translation of Gender-neutral Pronouns. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 173–181, Florence, Italy. Association for Computational Linguistics.

Greville G. Corbett. 1991. *Gender*. Cambridge University Press, Cambridge, UK.

Kate Crawford. 2017. The Trouble with Bias. In *Conference on Neural Information Processing Systems (NIPS) – Keynote*, Long Beach, California.

Steven Davis and Paul Mermelstein. 1980. Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 28(4):357–366.

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

Mattia A. Di Gangi, Robert Enyedi, Alessandra Brusadin, and Marcello Federico. 2019b. Robust Neural Machine Translation for Clean and Noisy Speech Transcripts. In *Proceedings of the 16th International Workshop on Spoken Language Translation (IWSLT)*, Hong Kong.

Mattia A. Di Gangi, Matteo Negri, and Marco Turchi. 2019c. Adapting Transformer to End-to-end Spoken Language Translation. In *Proceedings of INTERSPEECH 2019*, Graz, Austria.

Matthias Eck and Chiori Hori. 2005. Overview of the IWSLT 2005 Evaluation Campaign. In *Proceedings of the 2nd International Workshop on Spoken Language Translation (IWSLT)*, Pittsburgh, Pennsylvania.

Mostafa Elaraby, Ahmed Y. Tawfik, Mahmoud Khaled, Hany Hassan, and Aly Osama. 2018. Gender Aware Spoken Language Translation Applied to English-Arabic. In *Proceedings of the 2nd International Conference on Natural Language and Speech Processing (ICNLSP)*, pages 1–6, Algiers, Algeria.

Joel Escudé Font and Marta R. Costa-jussà. 2019. Equalizing gender bias in neural machine translation with word embeddings techniques. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 147–154, Florence, Italy. Association for Computational Linguistics.

Charles F. Hockett. 1958. *A Course in Modern Linguistics*. Macmillan, New York, US.

Sathish Indurthi, Houjeung Han, Nikhil Kumar Lakumarapu, Beomseok Lee, Insoo Chung, Sangha Kim, and Chanwoo Kim. 2019. Data Efficient Direct Speech-to-Text Translation with Modality Agnostic Meta-Learning. *arXiv preprint arXiv:1911.04283*.

Sergey Ioffe and Christian Szegedy. 2015. Batch normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. In *Proceedings of the 32nd International Conference on Machine Learning*, pages 448–456, Lille, France.

Selen Hande Kabil, Hannah Muckenhirn, and Mathew Magimai-Doss. 2018. On Learning to Identify Genders from Raw Speech Signal Using CNNs. In *Proceedings of INTERSPEECH 2018*, pages 287–291, Hyderabad, India.

Svetlana Kiritchenko and Saif Mohammad. 2018. Examining Gender and Race Bias in Two Hundred Sentiment Analysis Systems. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 43–53, New Orleans, Louisiana.

Ali Can Kocabiyikoglu, Laurent Besacier, and Olivier Kraif. 2018. Augmenting Librispeech with French Translations: A Multimodal Corpus for Direct Speech Translation Evaluation. In *Proceedings of LREC 2018*, Miyazaki, Japan.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. 1998. Gradient-based Learning Applied to Document Recognition. *Proceedings of the IEEE*, 86(11):2278–2324.

Kaiji Lu, Piotr Mardziel, Fangjing Wu, Preetam Amancharla, and Anupam Datta. 2019. Gender Bias in Neural Natural Language Processing. *arXiv preprint arXiv:1807.11714v2*.

Amit Moryossef, 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. Association for Computational Linguistics.

Jan Niehues, Roldano Cattoni, Sebastian Stüker, Mauro Cettolo, Marco Turchi, and Marcello Federico. 2018. The IWSLT 2018 Evaluation Campaign. In *Proceedings of the 15th International Workshop on Spoken Language Translation (IWSLT)*, Bruges, Belgium.

Jan Niehues, Roldano Cattoni, Sebastian Stüker, Matteo Negri, Marco Turchi, Thanh-Le ha, Elizabeth Salesky, Ramon Sanabria, Loic Barrault, Lucia Specia, and Marcello Federico. 2019. The IWSLT 2019 Evaluation Campaign. In *Proceedings of the 16th International Workshop on Spoken Language Translation (IWSLT)*, Hong Kong, Hong Kong.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: an ASR Corpus Based on Public Domain Audio Books. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5206–5210, Brisbane, Australia.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of ACL 2002*, pages 311–318, Philadelphia, Pennsylvania. Association for Computational Linguistics.

Daniel S. Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D. Cubuk, and Quoc V. Le. 2019. Specaugment: A Simple Data Augmentation Method for Automatic Speech Recognition. *arXiv preprint arXiv:1904.08779*.

Matt Post, Gaurav Kumar, Adam Lopez, Damianos Karakos, Chris Callison-Burch, and Sanjeev Khudanpur. 2013. Improved Speech-to-Text Translation with the Fisher and Callhome Spanish–English Speech Translation Corpus. In *Proceedings of the 10th International Workshop on Spoken Language Translation (IWSLT)*, Heidelberg, Germany.

Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, Jan Silovsky, Georg Stemmer, and Karel Vesely. 2011. The Kaldi Speech Recognition Toolkit. In *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pages 1–4, Big Island, Hawaii.

Daniel Povey, Vijayaditya Peddinti, Daniel Galvez, Pegah Ghahremani, Vimal Manohar, Xingyu Na, Yiming Wang, and Sanjeev Khudanpur. 2016. Purely Sequence-Trained Neural Networks for ASR Based on Lattice-Free MMI. In *Proceedings of INTER-SPEECH 2016*, pages 2751–2755, San Francisco, California.

Marcelo O. R. Prates, Pedro H. C. Avelar, and Luís C. Lamb. 2018. Assessing Gender Bias in Machine Translation: a Case Study with Google Translate. *Neural Computing and Applications*, pages 1–19.

Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 2018. Gender bias in coreference resolution. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 8–14, New Orleans, Louisiana. Association for Computational Linguistics.

Ramon Sanabria, Ozan Caglayan, Shruti Palaskar, Desmond Elliott, Loïc Barrault, Lucia Specia, and Florian Metze. 2018. How2: A Large-scale Dataset For Multimodal Language Understanding. In *Proceedings of the Workshop on Visually Grounded Interaction and Language (ViGIL)*, Montreal, Canada. NIPS.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015. Neural Machine Translation of Rare Words with Subword Units. *arXiv preprint arXiv:1508.07909*.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A Study of Translation Edit Rate with Targeted Human Annotation. In *Proceedings of AMTA*, pages 223–231, Cambridge, Massachusetts.

Matthias Sperber, Graham Neubig, Ngoc-Quan Pham, and Alex Waibel. 2019. Self-attentional models for lattice inputs. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1185–1197, Florence, Italy. Association for Computational Linguistics.

Matthias Sperber, Jan Niehues, and Alex Waibel. 2017. Toward Robust Neural Machine Translation for Noisy Input Sequences. In *Proceedings of the 14th International Workshop on Spoken Language Translation (IWSLT)*, Tokyo, Japan.

Gabriel Stanovsky, Noah A. Smith, and Luke Zettlemoyer. 2019. Evaluating gender bias in machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1679–1684, Florence, Italy. Association for Computational Linguistics.

Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. 2019. Mitigating Gender Bias in Natural Language Processing: Literature Review. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1630–1640, Florence, Italy. Association for Computational Linguistics.

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

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Proceedings of NIPS 2017*, pages 5998–6008, Long Beach, California. NIPS.

Kellie Webster, Marta Recasens, Vera Axelrod, and Jason Baldridge. 2018. Mind the GAP: A balanced corpus of gendered ambiguous pronouns. *Transactions of the Association for Computational Linguistics*, 6:605–617.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2017. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2979–2989, Copenhagen, Denmark. Association for Computational Linguistics.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender bias in coreference resolution: Evaluation and debiasing methods. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 15–20, New Orleans, Louisiana. Association for Computational Linguistics.