Personalized Machine Translation: Preserving Original Author Traits

Ella Rabinovich
Department of Computer Science
University of Haifa, Israel
& IBM Research - Haifa
ellarabi@gmail.com

Shachar Mirkin
IBM Research - Haifa
Mount Carmel, Haifa
31905, Israel
shacharm@il.ibm.com

Raj Nath Patel
C-DAC Mumbai
Gulmohar Cross Road No. 9, Juhu
Mumbai-400049, India
patelrajnath@gmail.com

Lucia Specia
Department of Computer Science
University of Sheffield, United Kingdom
l.specia@sheffield.ac.uk

Shuly Wintner
Department of Computer Science
University of Haifa, Israel
shuly@cs.haifa.ac.il

Abstract

The language that we produce reflects our personality, and various personal and demographic characteristics can be detected in natural language texts. We focus on one particular personal trait of the author, *gender*, and study how it is manifested in original texts and in translations. We show that author's gender has a powerful, clear signal in originals texts, but this signal is obfuscated in human and machine translation. We then propose simple domain-adaptation techniques that help retain the original gender traits in the translation, without harming the quality of the translation, thereby creating more personalized machine translation systems.

1 Introduction

Among many factors that mold the makeup of a text, gender and other authorial traits play a major role in our perception of the content we face. Many studies have shown that these traits can be identified by means of automatic classification methods. Classical examples include gender identification (Koppel et al., 2002), and authorship attribution and profiling (Seroussi et al., 2014). Most research, however, addressed texts in a single language, typically English.

We investigate a related but different question: we are interested in understanding what happens to personality and demographic textual markers during the translation process. It is generally agreed that good translation goes beyond transformation of the original content, by preserving more subtle and implicit characteristics inferred by author's personality, as well as era, geography, and various cultural and sociological aspects. In this work we explore whether translations preserve the stylistic characteristic of the author and, furthermore, whether the prominent signals of the source are retained in the target language.

As a first step, we focus on *gender* as a demographic trait (partially due to the absence of parallel data annotated for other traits). We evaluate the accuracy of automatic gender classification on original texts, on their manual translations and on their automatic translations generated through statistical machine translation (SMT). We show that while gender has a strong signal in originals, this signal is obfuscated in human and machine translation. Surprisingly, determining gender over manual translation is even harder than over SMT; this may be an artifact of the translation process itself or the human translators involved in it.

Mirkin et al. (2015) were the first to show that authorial gender signals tend to vanish through both manual and automatic translation, using a small TED talks dataset. We use their data and extend it with a version of Europarl that we annotated with age and gender (§3). Furthermore, we conduct experiments with two language pairs, in both directions (§4). We also adopt a different classification methodology based on the finding that the translation process itself has a stronger signal than the author's gender (§4.1).

We then move on to assessing gender traits in SMT (§5). Since SMT systems typically do not take personality or demographic information into account, we hypothesize that the author's style, affected by their personality, will fade. Furthermore, we propose simple domain-adaptation techniques that do consider gender information and can therefore better retain the original traits. We build "gender-aware" SMT systems, and show (§6) that they retain gender markers while preserving general translation quality. Our findings therefore suggest that SMT can be made much more personalized, leading to translations that are more faith-

Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers, pages 1074–1084,
Valencia, Spain, April 3-7, 2017. ©2017 Association for Computational Linguistics

ful to the style of the original texts.

Finally, we analyze the prominent features that reflect gender in originals and translations (§7). Our experiments reveal that gender markers differ greatly by language, and the specific source language has a significant impact on the features and classification accuracy of the translated text. In particular, gender traits of the original language overshadow those of the target language in both manual and automatic translation products.

The **main contributions** of this paper are thus: (i) a new parallel corpus annotated with gender and age information, (ii) an in-depth assessment of the projection of gender traits in manual and automatic translation, and (iii) experiments showing that gender-personalized SMT systems better project gender traits while maintaining translation quality.

2 Related work

While modeling of demographic traits has been proven beneficial in some NLP tasks such as sentiment analysis (Volkova et al., 2013) or topic classification (Hovy, 2015), very little attention has been paid to translation. We provide here a brief summary of research relevant to our work.

Machine translation (MT) Virtually no previous work in MT takes into account personal traits. State-of-the-art MT systems are built from examples of translations, where the general assumption is that the more data available to train models, the better, and a single model is usually produced. Exceptions to this assumption revolve around work on domain adaption, where systems are customized by using data that comes from a particular text domain (Hasler et al., 2014; Cuong and Sima'an, 2015); and work on data cleaning, where spurious data is removed from the training set to ensure the quality of the final models (Cui et al., 2013; Simard, 2014). Personal traits, sometimes well marked in the translation examples, are therefore not explicitly addressed. Learning from different, sometimes conflicting writing styles can hinder model performance and lead to translations that are unfaithful to the source text.

Focusing on reader preferences, Mirkin and Meunier (2015) used a collaborative filtering approach from recommender systems, where a user's preferred translation is predicted based on the preferences of similar users. However, the user preferences in this case refer to the overall choice between MT systems of a specific reader, rather than a choice based on traits of the writer. Mirkin et al. (2015) motivated the need for personalization of MT models by showing that automatic translation does not preserve demographic and psychometric traits. They suggested treating the problem as a domain adaptation one, but did not provide experimental results of personalized MT models.

Gender classification A large body of research has been devoted to isolating distinguishing traits of male and female linguistic variations, both theoretically and empirically. Apart from content, male and female speech has been shown to exhibit stylistic and syntactic differences. Several studies demonstrated that literary texts and blog posts produced by male and female writers can be distinguished by means of automatic classification, using (content-independent) function words and n-grams of POS tags (Koppel et al., 2002; Schler et al., 2006; Burger et al., 2011).

Although the tendencies of *individual word* usage are a subject of controversy, distributions of *word categories* across male and female English speech is nearly consensual: pronouns and verbs are more frequent in female texts, while nouns and numerals are more typical to male productions. Newman et al. (2008) carried out a comprehensive empirical study corroborating these findings with large and diverse datasets.

However, little effort has been dedicated to investigating the variation of individual markers of demographic traits across different languages. Johannsen et al. (2015) conducted a large-scale study on linguistic variation over age and gender across multiple languages in a social media domain. They showed that gender differences captured by shallow syntactic features were preserved across languages, when examined by linguistic categories. However, they did not study the distribution of individual gender markers across domains and languages. Our work demonstrates that while marker categories are potentially preserved, individual words typical to male and female language vary across languages and, more prominently, across different domains.

Authorial traits in translationese A large body of previous research has established that translations constitute an autonomic *language variety*: a special dialect of the target language, often referred to as *translationese* (Gellerstam, 1986). Recent corpus-based investigations of translationese demonstrated that originals and translations are distinguishable by means of supervised and unsupervised classification (Baroni and Bernardini, 2006; Volansky et al., 2015; Rabinovich and Wintner, 2015). The identification of machine-translated text has also been proven an easy task (Arase and Zhou, 2013; Aharoni et al., 2014).

Previous work has investigated how gender artifacts are carried over into human translation in the context of social and gender studies, as well as cultural transfer (Simon, 2003; Von Flotow, 2010). Shlesinger et al. (2009) conducted a computational study exploring the implications of the translator's gender on the final product. They conclude that "the computer could not be trained to accurately predict the gender of the translator". Preservation of authorial style in literary translations was studied by Lynch (2014), identifying Russian authors of translated English literature, by using (shallow) stylistic and syntactic features. Forsyth and Lam (2014) investigated authorial discriminability in translations of French originals into English, inspecting two distinct human translations, as well as automatic translation of the same sources.

Our work, to the best of our knowledge, is the first to automatically identify speaker gender in manual, and more prominently, automatic translations over multiple domains and language-pairs, examining distribution of gender markers in source and target languages.

3 Europarl with demographic info

We created a resource[1] based on the parallel corpus of the European Parliament (Europarl) Proceedings (Koehn, 2005). More specifically, we utilize the extension of its *en-fr* and *en-de* parallel versions (Rabinovich et al., 2015), where each sentence-pair is annotated with speaker name, the original language the sentence was uttered in, and the date of the corresponding session protocol. To extend speaker information with demographic properties, we used the Europarl website's MEP information pages[2] and applied a procedure of gender and age identification, as further detailed in §3.1.

The final resource comprises *en-fr* and *en-de* parallel bilingual corpora where metadata of members of the European Parliament (MEPs) is enriched with their gender and age at the time of the corresponding session. The data is restricted to sentence-pairs originally produced in English, French, or German. Table 1 provides statistics on the two datasets. We also release the full list of 3,586 MEPs with their meta information.

	en-fr	*fr-en*	*en-de*	*de-en*
male	100K	67K	101K	88K
female	44K	40K	61K	43K
total	144K	107K	162K	131K

Table 1: Europarl corpora (EP) statistics (# of sentence-pairs); gender refers to an author of the source utterance.

3.1 Identification of MEP gender

Gender annotation was conducted using three different resources: Wikidata, Genderize and AlchemyVision, which we briefly describe below.

Wikidata (Vrandečić and Krötzsch, 2014) is a human-curated knowledge repository of structured data from Wikipedia and other Wikimedia projects. Wikidata provides an API[3] through which one can retrieve details about people in the repository, including place and date of birth, occupation, and gender. For MEPs found in the Wikidata, we first verified that the person holds (or held) a position of Member of the European Parliament and if so, retrieved the gender. Wikidata information is not complete: not all MEP names, positions or gender data is included. In total we obtained gender information for 2,618 MEPs (73% of the total 3,586), of which 1,882 (72%) are male and 736 female (28%).

Genderize[4] is an open resource containing over 2 million distinct names grouped by countries. It determines people's gender based on their first name and the country of origin. Provided with the first name and the country a MEP represents.[5] Genderize was able to predict the gender of 2,785 MEPs, the vast majority of them with a probability of 0.9 or higher. We filtered out the 55 lower-confidence entries, keeping 2,730 MEPs (76% of total), of which 2001 (73%) are male and 729 (27%) female.

[1] Available at `http://cl.haifa.ac.il/projects/pmt`

[2] `http://www.europarl.europa.eu/meps/en/`

[3] `https://www.mediawiki.org/wiki/Wikibase/API`

[4] `https://genderize.io/`

[5] We assume that the country MEPs represent is highly correlated, if not strictly identical, to their country of origin.

AlchemyVision The European Parliament website maintains a page for every MEP, including personal photos. We classified MEP personal images using AlchemyVision,[6] a publicly available image recognition service. In total, we retrieved the gender of $2,236$ MEPs using AlchemyVision. Similarly to Genderize, we filtered out all predictions with a confidence score below 0.9, thus obtaining the gender of $2,138$ MEPs (60% of total), of which $1,528$ are male and 610 female (71% and 29%, respectively).

3.2 Resource evaluation and statistics

Even though Wikidata was created manually, to verify its correctness, we manually annotated the gender of 100 randomly selected MEPs with available Wikidata gender information; we found the metadata perfectly accurate. We therefore rely on Wikidata as a gold-standard against which we can assess the accuracy of the two other resources. Table 2 presents the accuracy and coverage of each resource based on this methodology.

resource	Wikidata	Genderize	Alchemy
coverage	73.0	76.1	59.6
accuracy	100.0	99.6	99.1

Table 2: Gender prediction performance (%).

Given information obtained from the three resources, we assign each MEP with a single gender prediction in the following way: whenever it is found in Wikidata ($2,618$ MEPs), the gender is determined by this resource. Otherwise, if both Genderize and AlchemyVision produced agreed-upon gender information (336 out of 338 cases), we set gender according to this prediction; the same applies to the case where only one of Genderize or AlchemyVision provided a prediction (346 and 178, respectively). We ended up with gender annotation for a total of $3,478$ out of $3,586$ members. The remaining 108 MEPs (92 male, 16 female) were annotated manually, a rather labor-intensive annotation in this case.

In total, the resource includes 947 (26%) female and $2,639$ (74%) male MEPs. Based on the above accuracy estimations, and assuming that manual annotation is correct, the overall accuracy of gender information in this resource is 99.88%.

Utilizing the information on session dates and MEPs dates of birth available in the metadata, we also annotated each sentence-pair with the age of the MEP at the time the sentence was uttered. To summarize, we release the following resources: (i) meta information for $3,586$ MEPs, as described above, (ii) bilingual parallel *en-fr* and *en-de* corpora, where each sentence-pair metadata is enriched with speaker MEPID, gender and age.

4 Experimental setup

We evaluate the extent to which gender traits are preserved in translation by evaluating the accuracy of gender classification of original and translated texts. The rationale is that the more prominent gender markers are in the text, the easier it is to classify the gender of its author.

4.1 Translationese vs. gender traits

Since we use the accuracy of gender identification as our evaluation metric, we isolate the dimension of gender in our data: the classification experiments are carried out separately on original, human translated text, as well as on each one of the MT products. Human, and more prominently, machine translations constitute distinct and distinguishable language variation, characterized by unique feature distributions (§2). We posit that in both human and machine translation products, the differences between original texts and translations overshadow the differences in gender. We corroborate this assumption by analysing a sample data distribution by two dimensions: (i) translation status and (ii) gender. Figure 1 presents the results for the English Europarl corpus. Both charts display data distributions of the same four classes: original (O) and translated (T) English[7] by male (M) and female (F) speakers (OM, OF, TM, TF). For the sake of visualization, the dimension of function words feature vectors was reduced to 2, using principal component analysis (Jolliffe, 2002). The left graph depicts color-separation by gender (male vs. female), while the right one by translation status (original vs. translated). Evidently, the linguistic variable of translationese stands out against the weaker signal of gender.

4.2 Datasets

In addition to the Europarl corpus annotated for gender (§3), we experimented with a corpus of

[6] https://www.ibm.com/smarterplanet/us/en/ibmwatson/developercloud/alchemy-vision.html

[7] This experiment refers to English translated from French; other language-pairs exhibited similar trends.

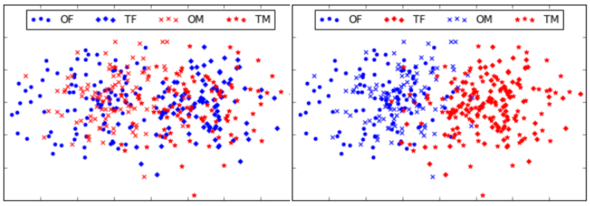

Figure 1: English EP data distributions across two dimensions: gender (left) and trans. status (right).

TED talks (transcripts and translations): a collection of texts from a completely different genre, where demographic traits may manifest differently. Testing the potential benefits of personalized SMT models on these two very diverse datasets allows us to examine the robustness of our approach. We used the TED gender-annotated data from Mirkin et al. (2015).[8] This corpus contains annotation of the speaker's gender included in the English-French corpus of the IWSLT 2014 Evaluation Campaign's MT track (Cettolo et al., 2012). We annotated 68 additional talks from the development and test sets of IWSLT 2014, 2015 and 2016. Using the full set, we split the TED parallel corpora by gender to obtain sub-corpora of 140K and 43K sentence pairs for male and female speakers, respectively.

The sizes of the datasets used for training, tuning and testing of SMT models are shown in Table 3. Relatively large test sets are used for evaluation of the MT results for the sake of reliable per-outcome gender classification (§4.1).

Although the size of the training/tuning/test sets in either direction for any language-pair is the same, their content is different. We use data in both translation directions (i.e., *en-fr* and *fr-en*, or *en-de* and *de-en*) for both SMT experiments. Out of these data, 2K and 15K sentence-pairs (for each gender) are held out for tuning and test, respectively, where they comply with the translation direction. That is, for *en-fr* experiments, tuning and test sets are sampled from the *en-fr* direction only and vice-versa. The additional bilingual data (ADD) for training the models comes from the gender-unannotated portion of Europarl (all but the gender-annotated sub-corpus detailed in §3) for the EP experiments, and from combining TED's male and female data for the experiments with TED.

4.3 Classification setting

All datasets were split by sentence, filtering out sentence alignments other than one-to-one. For POS tagging, we employed the Stanford implementation[9] with its models for English, French and German. We divided all datasets into chunks of approximately 1,000 tokens, respecting sentence boundaries, and normalized the values of lexical features by the actual number of tokens in each chunk. For classification, we used Platt's sequential minimal optimization algorithm (Keerthi et al., 2001) to train support vector machine classifiers with the default linear kernel (Hall et al., 2009). In all experiments we used (the maximal) equal amount of data from each category (M and F), specifically, 370 chunks for each gender.

Aiming to abstract away from content and capture instead stylistic and syntactic characteristics, we used as our feature set the combination of function words (FW)[10] and (the top-1,000 most frequent) POS-trigrams. We employ 10-fold cross-validation for evaluation of classification accuracy.

4.4 SMT setting

We trained phrase-based SMT models with Moses (Koehn et al., 2007), an open source SMT system. KenLM (Heafield, 2011) was used for language modeling. We trained 5-gram language models with Kneser-Ney smoothing (Chen and Goodman, 1996). The models were tuned using Minimum Error Rate Tuning (MERT) (Och, 2003). Our preprocessing included cleaning (removal of empty, long and misaligned sentences), tokenization and punctuation normalization. The Stanford tokenizer (Manning et al., 2014) was used for tokenization and standard Moses scripts were used for other preprocessing tasks. We used BLEU (Papineni et al., 2002) to evaluate MT quality against one reference translation.

5 Personalized SMT models

In order to investigate and improve gender traits transfer in MT, we devise and experiment with gender-aware SMT models. We demonstrate that despite their simplicity, these models lead to better preservation of gender traits, while not harming the general quality of the translations.

[8] Downloaded from `http://cm.xrce.xerox.com/`.

[9] `http://nlp.stanford.edu/software/tagger.shtml`

[10] We used the lists of function words available at `https://code.google.com/archive/p/stop-words`.

		training			tuning		test	
dataset	language-pair	M	F	ADD	M	F	M	F
EP	*en-fr* & *fr-en*	144K	65K	1.71M	2K	2K	15K	15K
	en-de & *de-en*	170K	86K	1.50M	2K	2K	15K	15K
TED	*en-fr*	117K	21K	138K	2K	2K	20K	20K

Table 3: MT datasets split for train, tuning and test, after cleaning.

We treat the task of personalizing SMT models as a domain adaptation task, where the *domain* is the gender. We applied two common techniques: (i) gender-specific model components (phrase table and language model (LM)) and (ii) gender-specific tuning sets. These personalized configurations are further compared to a baseline model where gender information is disregarded, as described below. In all cases, we use a single reordering table built from the entire training set.

Baseline The baseline (*MT-B*) system was trained using the complete parallel corpus available for a language-pair. The training set contained both gender-specific and unannotated data, but no distinction was made between them. A single translation model and a single LM were built, and the model was tuned using a random sample of 2K sentence-pairs from the mixed data dedicated for tuning, preserving, therefore, the gender distribution of the underlying dataset.

Personalized models These models use three datasets: male, female, and additional in-domain bilingual data. Two configurations were devised: *MT-P1*, a model with three phrase tables and three LMs trained on the three datasets; and *MT-P2*, where for each gender a phrase table and a language model were built using only the gender-specific data, as well as a general phrase table and LM. In both configurations, each of the two genderized model variants was tuned using the gender-specific tuning set. In order to evaluate the translation quality of a personalized model, we separately translated the male and female source segments, merged the outputs and evaluated the merged result.

6 Results

Recall that we use the accuracy of gender classification as a measure of the strength of gender markers in texts. We assessed this accuracy below on originals and (human and machine) translations. First, however, we establish that the quality of SMT is not harmed with our personalized models.

MT evaluation We trained a baseline (*MT-B*) and two personalized models (*MT-P1* and *MT-P2*) for each language pair as detailed in §5. The BLEU scores of *en-fr* and *fr-en* personalized models were 38.42, 38.34 and 37.16, 37.16, with the baseline models scoring 38.65 and 37.35, respectively. Similarly, for experiments with *en-de* and *de-en* and the TED data, the baseline scores (21.95, 26.37 and 33.25) were only marginally higher than those of the personalized models (21.65, 21.80; 26.35, 26.21; and 33.19, 33.16), with differences ranging from 0.02 to 0.3. Neither *MT-P1* nor *MT-P2* was consistently better than the other. We conclude, therefore, that all MT systems are comparable in terms of general quality.

Classification accuracy Tables 4 and 5 present the results of gender classification accuracy in original (O), human- (HT) and machine-translated texts in the EP corpus. Female texts are distinguishable from their male counterparts with 77.3% and 77.1% accuracy for English originals, in line with accuracies reported in the literature (Koppel et al., 2002). Classification of original French and German texts reach 81.4% (Table 4) and 76.1% (Table 5), respectively.

	precision		recall		acc.
dataset	M	F	M	F	
en O	77.7	76.9	76.5	78.1	77.3
fr O	80.9	81.9	82.2	80.5	81.4
fr-en HT	75.6	74.4	73.8	76.2	75.0
fr-en MT-B	77.0	78.2	78.6	76.5	77.6
fr-en MT-P1	82.0	80.7	80.3	82.4	**81.4**
fr-en MT-P2	79.1	81.0	81.6	78.4	80.0
en-fr HT	56.6	56.4	55.7	57.3	56.5
en-fr MT-B	60.2	60.1	60.0	60.3	60.1
en-fr MT-P1	62.7	63.0	63.5	62.2	62.8
en-fr MT-P2	65.2	65.3	65.4	65.1	**65.3**

Table 4: EP *en-fr*, *fr-en* classification scores (%).

Evidently, gender traits are significantly obfuscated by both manual and non-personalized ma-

	precision		recall		acc.
dataset	M	F	M	F	
en O	77.5	76.7	76.5	77.7	77.1
de O	76.4	75.7	75.4	76.8	76.1
de-en HT	68.6	67.9	67.3	69.2	68.2
de-en MT-B	69.3	69.9	70.3	68.9	69.6
de-en MT-P1	77.4	75.9	75.1	78.1	**76.6**
de-en MT-P2	76.2	75.7	75.4	76.5	75.9
en-de HT	59.8	59.7	59.5	60.0	59.7
en-de MT-B	63.8	64.0	64.3	63.5	63.9
en-de MT-P1	69.6	69.4	69.2	69.7	**69.5**
en-de MT-P2	66.7	67.7	68.6	65.7	67.2

Table 5: EP *en-de*, *de-en* classification scores (%).

chine translation. The relatively low accuracy for human translation can be (partially) explained by the extensive editing procedure applied on Europarl proceedings prior to publishing (Cucchi, 2012), as well as the potential "fingerprints" of (male or female) human translators left on the final product.

Both *MT-P1* and *MT-P2* models yield translations that better preserve gender traits, compared to their manual and gender-agnostic automatic counterparts: accuracy improvements vary between 3.8 for *fr-en* translations to 7.0 percent points for *de-en*[11] (*MT-P1* vs *MT-B* in both cases). Per-class precision and recall scores do not exhibit significant differences, despite the unbalanced amount of per-gender data used for training the MT models.

Gender classification results in the TED dataset are presented in Table 6. The classification accuracy of English originals is 80.4%. While, similarly to Europarl, the gender signal is generally weakened in human translations[12] and baseline MT, overall accuracies are in most cases higher than in Europarl across all models. We attribute this difference to the more emotional and personal nature of TED speeches, compared with the formal language of the EP proceedings. Both personalized SMT models significantly outperform their baseline counterpart, as well as the manual translation, yielding 77.2% and 77.7% accuracy for *MT-P1* and *MT-P2*, respectively.

	precision		recall		acc.
dataset	M	F	M	F	
en O	81.2	79.7	79.2	81.6	80.4
en-fr HT	74.0	73.5	73.2	74.3	73.8
en-fr MT-B	71.3	70.1	69.2	72.2	70.7
en-fr MT-P1	77.5	76.8	76.5	77.8	77.2
en-fr MT-P2	78.2	77.2	76.8	78.6	**77.7**

Table 6: TED *en-fr* classification scores (%).

7 Analysis

Analysis of gender markers To analyze the extent to which personal traits are preserved in translations, we extract the set of most discriminative FWs in various texts by employing the InfoGain feature selection procedure (Gray, 1990). Gender markers vary across *original* languages (with few exceptions); in EP, the most discriminating English features are *also, very, perhaps, as, its, others, you*. The French list includes *on, vous, dire, afin, doivent, doit, aussi, avait, voilà, je*, while the German list consists of *wir, man, wirklich, sollten, von, für, dass, allen, ob*. The list of discriminative markers in the TED English dataset contains mainly personal pronouns: *she, her, I, you, my, our, me, and, who, it*.

Figure 2 (top) presents weights assigned to various gender markers by the InfoGain attribute evaluator in originals and translations. Gender markers are carried over to (both manual and machine) translations to an extent that overshadows the original markers of the target language. In particular, the markers observed in translated English mirror their original French counterparts, in the same marker role: *I* (M) in English translations reflecting the original French *je* (M), *say* (M) reflecting *dire* (M), *must* (F) translated from *doit* (F) and *doivent* (F); the latter contradicting the original English *must* which characterizes M speech. The original English prominent gender markers (e.g., *also, very*) almost completely lose their discriminative power in translations. A similar phenomenon is exhibited by English translations from German, as depicted in Figure 2 (bottom): the German *wir (we)*, *für (for)* and *ob (whether)* are preserved in (both manual and machine) English translations, in the same marker role.

We conclude that (i) gender traits in translation are weakened, compared to their originals. Furthermore, (ii) translations tend to embrace gender tendencies of the original language, thus resulting

[11] All differences between *MT-P1* and *MT-P2* and baseline models are statistically significant.

[12] TED talks are subtitled, rather than transcribed, undergoing some editing and rephrasing.

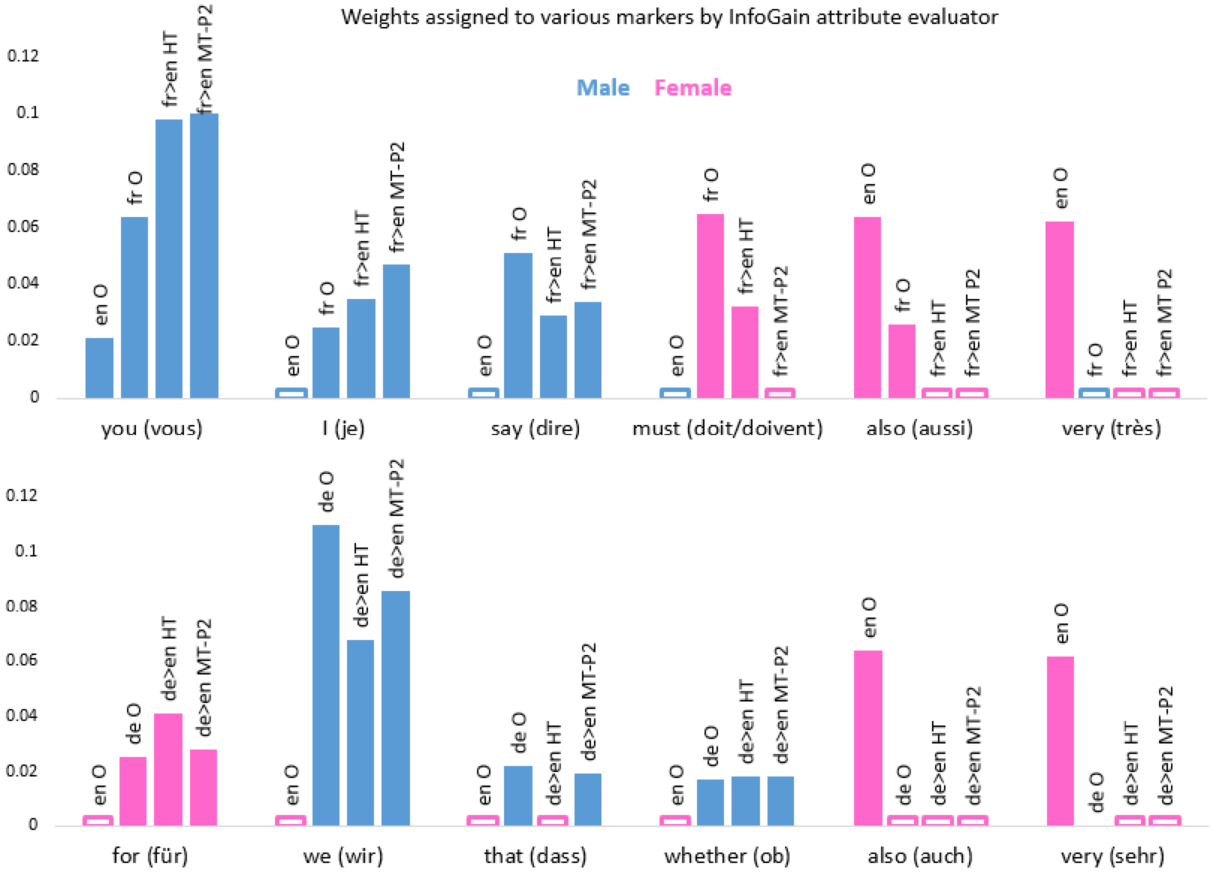

Figure 2: Persistence of *en* and *fr* markers in *fr-en* translations (top); *en* and *de* markers in *de-en* translations (bottom). The transparent bars refer to (weak) F/M markers, assigned weight<0.01 by InfoGain.

in a *hybrid* outcome, where male and female traits are affected both by markers of the source and (to a much lesser extent) the target language.

Capturing the "personalization" effect Both manual- and all machine-translations of Europarl are tested on a strictly identical set of sentences; therefore, the performance gap introduced by personalized SMT models can be captured by a subset of sentences misclassified by the baseline model, but classified correctly when applying a more personalized approach. The inspection of differences in these translations can shed some light on the underlying nature of our personalized models. Table 7 (top) shows manual, baseline, and personalized machine translations of examples of French and German sentences. The translation of the French word "vraiment" (in a *male* utterance) varies in English as "really" or "exactly", where the former is more frequent in female English texts, and the latter is a male marker. The choice of a *male* English marker over its *female* equivalent by the gender-aware SMT model demonstrates the effect of personalization as proposed in this paper. The translations of the German *female* sentence into English, as presented in Table 7 (bottom), further highlight this phenomenon by choosing the English *female* marker *think* in its personalized translation over the more neutral *consider* and *believe* in the manual and baseline versions, respectively.

8 Conclusions

We presented preliminary results of employing personalized SMT models for better preservation of gender traits in automatic translation. This work leaves much room for further research and practical activities. Authors' personal traits are utilized by recommendation systems, conversational agents and other personalized applications. While resources annotated for personality traits mainly exist for English (and recently, for a small set of additional languages), they are scarce or missing from most other languages. Employing MT models that are sensitive to authors' personal traits can

fr O	*... on a corrigé la traduction du mot qui a été traduit en français par “propriété” qui n’est pas* ***vraiment*** *la même chose qu’ “appropriation”.*
fr-en HT	*... it had been translated into French using the word for “property”, which is not* ***really*** *the same thing as “ownership”.*
fr-en MT-B	*... it was corrected the translation of the word which has been translated into French as “ownership”, which is not* ***really*** *the same as “ownership”.*
fr-en MT-P1	*... it has corrected the translation of the word which has been translated into French as “ownership”, which is not* ***exactly*** *the same as “ownership”.*
de O	*Entsprechend* ***halte ich es auch für notwendig****, daß die Kennzeichnung möglichst schnell und verpflichtend eingeführt wird, und zwar für Rinder und für Rindfleisch .*
de-en HT	*Accordingly,* ***I consider it essential*** *that both the identification of cattle and the labelling of beef be introduced as quickly as possible on a compulsory basis.*
de-en MT-B	*Similarly,* ***I believe that it is necessary****, as quickly as possible and that compulsory labelling will be introduced, and for bovine animals and for beef and veal.*
de-en MT-P1	*Accordingly,* ***I also think it is essential*** *that the labelling and become mandatory as quickly as possible, and for bovine animals and for beef.*

Table 7: Translation of *fr* (M) and *de* (F) sentences into English manually, and by different MT models.

facilitate user modeling in other languages as well as augment English data with translated content.

Our future plans include experimenting with more sophisticated MT models, and with additional demographic traits, domains and language-pairs.

Acknowledgments

This research was partly supported by the H2020 QT21 project (645452, Lucia Specia). We are grateful to Sergiu Nisioi for sharing the initial collection of properties of Members of the European Parliament. We also thank our anonymous reviewers for their constructive feedback.

References

Roee Aharoni, Moshe Koppel, and Yoav Goldberg. 2014. Automatic detection of machine translated text and translation quality estimation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014, June 22-27, 2014, Baltimore, MD, USA, Volume 2: Short Papers*, pages 289–295.

Yuki Arase and Ming Zhou. 2013. Machine translation detection from monolingual web-text. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, ACL 2013, 4-9 August 2013, Sofia, Bulgaria, Volume 1: Long Papers*, pages 1597–1607.

Marco Baroni and Silvia Bernardini. 2006. A new approach to the study of Translationese: Machine-learning the difference between original and translated text. *Literary and Linguistic Computing*, 21(3):259–274, September.

John D. Burger, John Henderson, George Kim, and Guido Zarrella. 2011. Discriminating gender on twitter. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP ’11, pages 1301–1309, Stroudsburg, PA, USA. Association for Computational Linguistics.

Mauro Cettolo, Christian Girardi, and Marcello Federico. 2012. Wit^3: Web inventory of transcribed and translated talks. In *Proceedings of the 16th Conference of the European Association for Machine Translation (EAMT)*, pages 261–268, Trento, Italy, May.

Stanley F. Chen and Joshua Goodman. 1996. An empirical study of smoothing techniques for language modeling. In *Proceedings of ACL*.

Costanza Cucchi. 2012. Dialogic features in EU non-native parliamentary debates. *Review of the Air Force Academy*, 11(3):5–14.

Lei Cui, Dongdong Zhang, Shujie Liu, Mu Li, and Ming Zhou. 2013. Bilingual data cleaning for smt using graph-based random walk. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 340–345, Sofia, Bulgaria.

Hoang Cuong and Khalil Sima’an. 2015. Latent domain word alignment for heterogeneous corpora. In *Proceedings of the Annual Meeting of the North American Chapter of the Association for Computational Linguistics*, pages 398–408, Denver, USA.

Richard S Forsyth and Phoenix WY Lam. 2014. Found in translation: To what extent is authorial discriminability preserved by translators? *Literary and Linguistic Computing*, 29(2):199–217.

Martin Gellerstam. 1986. Translationese in Swedish novels translated from English. In Lars Wollin and Hans Lindquist, editors, *Translation Studies in Scandinavia*, pages 88–95. CWK Gleerup, Lund.

Robert M Gray. 1990. Entropy and information. In *Entropy and Information Theory*, pages 21–55. Springer.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. 2009. The WEKA data mining software: an update. *SIGKDD Explorations*, 11(1):10–18.

Eva Hasler, Barry Haddow, and Philipp Koehn. 2014. Dynamic topic adaptation for smt using distributional profiles. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 445–456, Baltimore, USA.

Kenneth Heafield. 2011. KenLM: faster and smaller language model queries. In *Proceedings of the EMNLP 2011 Sixth Workshop on Statistical Machine Translation*, pages 187–197, Edinburgh, Scotland, United Kingdom, July.

Dirk Hovy. 2015. Demographic factors improve classification performance. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, pages 752–762, Beijing, China.

Anders Johannsen, Dirk Hovy, and Anders Søgaard. 2015. Cross-lingual syntactic variation over age and gender. In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning*, pages 103–112, Beijing, China, July. Association for Computational Linguistics.

Ian Jolliffe. 2002. *Principal component analysis*. Wiley Online Library.

S.S. Keerthi, S.K. Shevade, C. Bhattacharyya, and K.R.K. Murthy. 2001. Improvements to platt's smo algorithm for svm classifier design. *Neural Computation*, 13(3):637–649.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, ACL '07, pages 177–180, Stroudsburg, PA, USA. Association for Computational Linguistics.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. MT Summit.

Moshe Koppel, Shlomo Argamon, and Anat Rachel Shimoni. 2002. Automatically categorizing written texts by author gender. *LLC*, 17(4):401–412.

Gerard Lynch. 2014. A supervised learning approach towards profiling the preservation of authorial style in literary translations. In *COLING*, pages 376–386.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60, Baltimore, Maryland, June. Association for Computational Linguistics.

Shachar Mirkin and Jean-Luc Meunier. 2015. Personalized machine translation: Predicting translational preferences. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2019–2025, Lisbon, Portugal, September. Association for Computational Linguistics.

Shachar Mirkin, Scott Nowson, Caroline Brun, and Julien Perez. 2015. Motivating personality-aware machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1102–1108, Lisbon, Portugal, September. Association for Computational Linguistics.

Matthew L. Newman, Carla J. Groom, Lori D. Handelman, and James W. Pennebaker. 2008. Gender differences in language use: An analysis of 14,000 text samples. *Discourse Processes*, 45(3):211–236.

Franz Josef Och. 2003. Minimum error rate training in statistical machine translation. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1 (ACL 2003)*, ACL '03, pages 160–167.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of ACL*, Philadelphia, Pennsylvania, USA.

Ella Rabinovich and Shuly Wintner. 2015. Unsupervised identification of translationese. *Transactions of the Association for Computational Linguistics*, 3:419–432.

Ella Rabinovich, Shuly Wintner, and Ofek Luis Lewinsohn. 2015. The haifa corpus of translationese. *arXiv preprint arXiv:1509.03611*.

Jonathan Schler, Moshe Koppel, Shlomo Argamon, and James W. Pennebaker. 2006. Effects of age and gender on blogging. In *Computational Approaches to Analyzing Weblogs, Papers from the 2006 AAAI Spring Symposium, Technical Report SS-06-03, Stanford, California, USA, March 27-29, 2006*, pages 199–205.

Yanir Seroussi, Ingrid Zukerman, and Fabian Bohnert. 2014. Authorship attribution with topic models. *Comput. Linguist.*, 40(2):269–310.

Miriam Shlesinger, Moshe Koppel, Noam Ordan, and Brenda Malkiel. 2009. Markers of translator gender: do they really matter? *Internet. Disponível em http://u. cs. biu. ac. il/˜ koppel/Publications. ht ml (consultado em 31de março de 2011).*

Michel Simard. 2014. Clean data for training statistical mt: the case of mt contamination. In *Proceedings of the Eleventh Conference of the Association for Machine Translation in the Americas*, pages 69–82, Vancouver, BC, Canada.

Sherry Simon. 2003. *Gender in translation*. Routledge.

Vered Volansky, Noam Ordan, and Shuly Wintner. 2015. On the features of translationese. *Digital Scholarship in the Humanities*, 30(1):98–118, April.

Svitlana Volkova, Theresa Wilson, and David Yarowsky. 2013. Exploring demographic language variations to improve multilingual sentiment analysis in social media. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1815–1827, Seattle, Washington, USA.

Luise Von Flotow. 2010. Gender in translation. *Handbook of Translation Studies*, 1:129–133.

Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: A free collaborative knowledgebase. *Commun. ACM*, 57(10):78–85, September.