Personalized Machine Translation: Preserving Original Author Traits

Ella Rabinovich^{1,2}, Shachar Mirkin¹, Raj Nath Patel³, Lucia Specia⁴, Shuly Wintner²

¹IBM Research – Haifa, Israel
 ²Department of Computer Science, University of Haifa, Israel
 ³C-DAC Mumbai, India
 ⁴University of Sheffield, United Kingdom

EACL 2017, Valencia

Background – Personalized Machine Translation

- The language we produce reflects our personality
 - Demographics: gender, age, geography etc.
 - Personality: extraversion, agreeableness, openness, conscientiousness, neuroticism (the "Big Five")
- Authorial traits affect our perception of the content we face
 - -We may have a preference to a specific authorial style
- Personalized Machine Translation (PMT)
 - Preserving authorial traits in manual and machine translation (Mirkin et al., 2015)
 - Predicting user's translation preference (Mirkin and Meunier, 2015)

Background – Authorial Gender

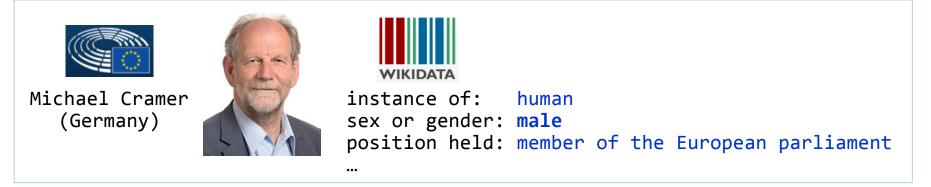
- Male and female speech differs, to an extent distinguishable by automatic classification (Koppel et al., 2002; Schler et al., 2006; Burger et al., 2011)
 - Male speakers use *nouns* and *numerals* more frequently
 - associated with the alleged "information emphasis"
 - Female prominent signals include verbs and pronouns
 - e.g., "we" as a marker of group identity

Research Questions

- Are the prominent authorial signals preserved through translation?
 Human (a translator involved) and machine translation
- Can machine-translation models be adapted to better preserve authorial traits?
- Are authorial traits in translated text retained from the source?
 Do they differ from those of the target language?
- We focus on SMT adaptation to better preserve authorial *gender* markers through automatic translation

Datasets

- Europarl proceedings of the European Parliament
 - -Automatically annotated¹ for speaker *gender and age* using:
 - Wikidata (manually curated dataset)



- Genderize.io (based on person's first name and country)
- Alchemy vision (image classification for gender)
- -Estimated accuracy of gender annotation in the dataset is 99.8%
 - Based on an evaluation against the Wikidata ground truth

¹ <u>http://cl.haifa.ac.il/projects/pmt/</u>

Datasets (cont.)

- TED talks transcripts
 - -English-French corpus of IWSLT 2014 Evaluation Campaign's MT track
 - Annotated for speaker gender (Mirkin et al., 2015)

gender / language pair	en-fr	fr-en	en-de	de-en
	Europarl			
# of sentences by M speakers	100K	67K	101K	88K
# of sentences by F speakers	44K	40K	61K	43K
additional (not annotated) data	1.7	7M	1.5	5M
	TED	<u> </u>		
# of sentences by M speakers	140K			
# of sentences by F speakers	43K			

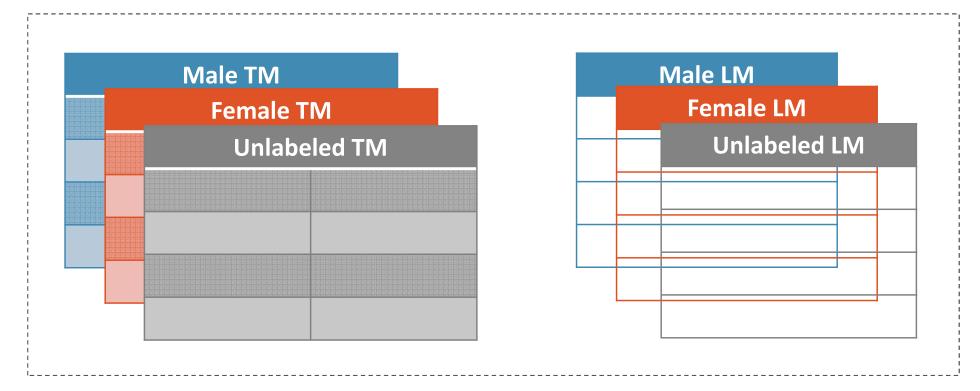
* the numbers refer to sentences originally uttered in the source language

Personalized MT - Approach

- Gender-aware SMT models
 - -Personalization as a *domain-adaptation* task
 - Gender-specific model components (TM and LM)
 - Gender-specific tuning sets
- Baseline model disregarding the gender information
 - -A single TM and LM is built using male, female and unlabeled data
 - -Tuning is done using a random sample of sentences

Personalized MT Models

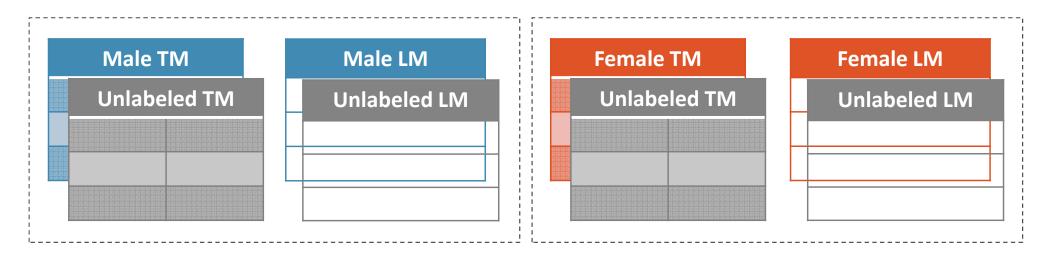
 MT-PERS1: a single system with 3 TMs and 3 LMs trained on male (M), female (F) and additional unlabeled data



- The model was tuned using the gender-specific tuning set
 - Resulting in 2 sub-models that differ in their tuning

Personalized MT Models (cont.)

 MT-PERS2: two separate systems, each one comprising gender-specific (M or F), as well as unlabeled TM and LM



Both models were tuned using the gender-specific tuning set

MT Evaluation Results (BLEU)

- Phrase-based SMT Moses (Koehn et al., 2007)
- Language modeling done using KenLM (Heafield, 2011)
 - 5-gram LMs with Kneser-Ney smoothing
- Tuning with MERT

	model / language-pair	en-fr	fr-en	en-de	de-en
arl	MT-baseline	38.65	37.65	21.95	26.37
Europar	MT-PERS1	38.42	37.16	21.65	26.35
Eur	MT-PERS2	38.34	37.16	21.80	26.21
	MT-baseline	33.25			
TED	MT-PERS1	33.19			
	MT-PERS2	33.16			

Personalized models do not harm MT quality

Preserving Gender Traits – Evaluation

- Binary (M vs F) classification of each model output
 Human- and machine-translation
- Features: frequencies of function words and POS-trigrams
 —Stylistic, content-independent features
- Classification units: random chunks of 1K tokens
 Inline with Schler et al., 2006 (classified blog posts)
 Gender classification at small units, e.g., sentence, is practically impossible
- Linear SVM classifier, 10-fold cross-validation evaluation

	language (-pair)	accuracy (%)		language (-pair)	accuracy (%)
	en O	77.3		en O	80.4
	fr O	81.4		en-fr HT	73.8
	fr-en HT	75.0	Ц Ц Ц	en-fr MT-baseline	70.7
L	fr-en MT-baseline	77.6	-	en-fr MT-PERS1	77.2
Europai	fr-en MT-PERS1	81.4		en-fr MT-PERS2	77.7
n	fr-en MT-PERS2	80.0			
	en-fr HT	56.5			
	en-fr MT-baseline	60.1			
	en-fr MT-PERS1	62.8			
	en-fr MT-PERS2	65.3			

	language (-pair)	accuracy (%)		language (-pair)	accuracy (%)
	en O	77.3		en O	80.4
	fr O	81.4		en-fr HT	73.8
	fr-en HT	75.0	ËD	en-fr MT-baseline	70.7
	fr-en MT-baseline	77.6	-	en-fr MT-PERS1	77.2
Europar	fr-en MT-PERS1	81.4		en-fr MT-PERS2	77.7
n	fr-en MT-PERS2	80.0			
	en-fr HT	56.5			
	en-fr MT-baseline	60.1			
	en-fr MT-PERS1	62.8			
	en-fr MT-PERS2	65.3			

	language (-pair)	accuracy (%)		language (-pair)	accuracy (%)
	en O	77.3		en O	80.4
	fr O	81.4		en-fr HT	73.8
	fr-en HT	75.0	ED.	en-fr MT-baseline	70.7
L	fr-en MT-baseline	77.6	-	en-fr MT-PERS1	77.2
do	fr-en MT-PERS1	81.4		en-fr MT-PERS2	77.7
Europarl	fr-en MT-PERS2	80.0			
	en-fr HT	56.5	I		
	en-fr MT-baseline	60.1			
	en-fr MT-PERS1	62.8			
	en-fr MT-PERS2	65.3			

	language (-pair)	accuracy (%)		language (-pair)	accuracy (%)
	en O	77.3		en O	80.4
	fr O	81.4		en-fr HT	73.8
	fr-en HT	75.0	ЦЦ	en-fr MT-baseline	70.7
	fr-en MT-baseline	77.6		en-fr MT-PERS1	77.2
do	fr-en MT-PERS1	81.4		en-fr MT-PERS2	77.7
Europarl	fr-en MT-PERS2	80.0			
	en-fr HT	56.5			
	en-fr MT-baseline	60.1			
	en-fr MT-PERS1	62.8			
	en-fr MT-PERS2	65.3			

	language (-pair)	accuracy (%)		language (-pair)	accuracy (%)
	en O	77.3		en O	80.4
	fr O	81.4		en-fr HT	73.8
	fr-en HT	75.0	ЦЦ	en-fr MT-baseline	70.7
L	fr-en MT-baseline	77.6	-	en-fr MT-PERS1	77.2
Europai	fr-en MT-PERS1	81.4		en-fr MT-PERS2	77.7
n	fr-en MT-PERS2	80.0			
	en-fr HT	56.5			
	en-fr MT-baseline	60.1			
	en-fr MT-PERS1	62.8			
	en-fr MT-PERS2	65.3			

Binary classification using function words and top-1000 POS-trigrams

	language (-pair)	accuracy (%)		language (-pair)	accuracy (%)
	en O	77.3		en O	80.4
	fr O	81.4		en-fr HT	73.8
	fr-en HT	75.0	ED	en-fr MT-baseline	70.7
L	fr-en MT-baseline	77.6	-	en-fr MT-PERS1	77.2
do	fr-en MT-PERS1	81.4		en-fr MT-PERS2	77.7
Europar	fr-en MT-PERS2	80.0			
	en-fr HT	56.5			
	en-fr MT-baseline	60.1			
	en-fr MT-PERS1	62.8			
	en-fr MT-PERS2	65.3			

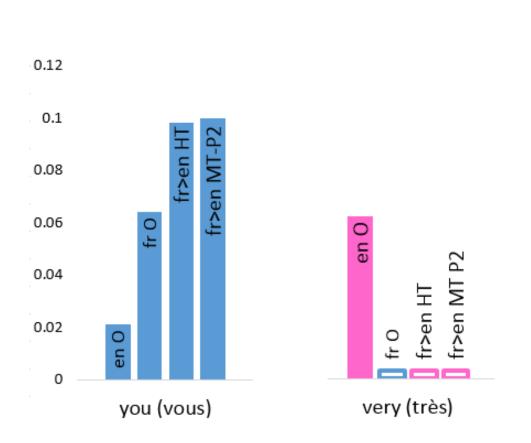
* similar results obtained for en-de and de-en translations

Analysis – Gender Markers

- Are gender markers of the original language preserved in translation?
- Distribution of individual gender markers varies between languages
 - English: "must" is a *male* marker
 - -French: "doit" and "doivent" are more frequent in *female* speech
 - English: "we" exhibits nearly equal frequencies in *male and female* texts
 German: "wir" is a prominent *female* marker
- Translations tend to embrace gender tendencies of the original language
 - Resulting in a hybrid outcome where M and F traits are affected both by markers of the *source* and (to a much lesser extent) the *target* language

Analysis (cont.)

Weights assigned to various gender marker by InfoGain attribute evaluator



Male Female

Summary

- Author gender is strongly marked in original texts
- This signal is obfuscated in human and machine translation
- Simple personalized SMT models using standard domain adaptation techniques offer a good approach for preserving gender traits in automatic translation

Future work

- State-of-the-art NMT models for personalization in translation
- Additional domains, datasets and language-pairs
- Additional authorial traits, e.g., age

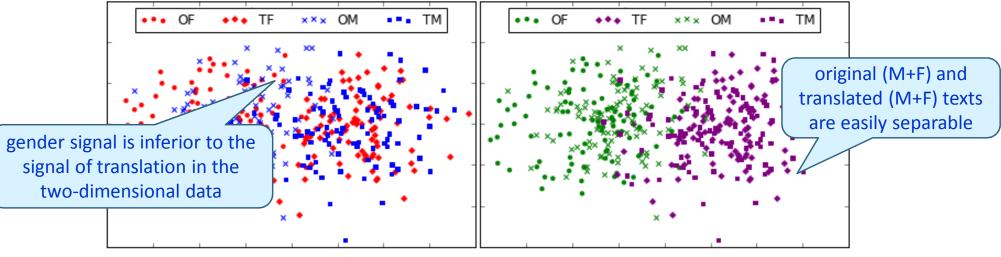
Backup

Preserving Gender Traits - Evaluation

Translations and original texts constitute distinct language variants

- Distinguishable by text classification techniques

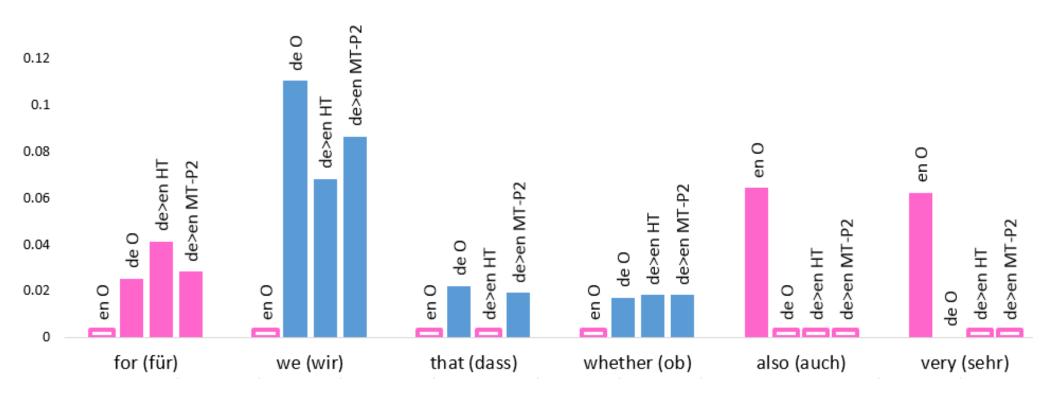
• We found that the signal of translation overshadows that of gender



Multivariate data color-separated by two dimensions (using function words as features) male vs female manually-translated vs original

- We therefore evaluate the signal of gender by classification of M vs F texts separately in original, human- and machine-translated texts
 - A gender classifier trained on originals fails to predict gender in translations

Analysis (en-de)



23

Capturing the personalization effect

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 O de-en HT	
	verpflichtend eingeführt wird, und zwar für Rinder und für Rindfleisch. Accordingly, I consider it essential that both the identification of cattle and the labelling of

The French "vraiment" in male utterance is translated as "really" by the gender-agnostic (and human) models, and as "exactly" by the personalized version; in German example, a female utterance is translated as English female marker "think", compared to the more neutral "believe" and "consider"