On the Integration of (Extra-) Linguistic Information in Neural Machine Translation: 
A Case Study of Gender

Eva Vanmassenhove

This work has been supported by the Dublin City University Faculty of Engineering & Computing under the Daniel O’Hare Research Scholarship scheme and by the ADAPT Centre for Digital Content Technology, funded under the SFI Research Centres Programme (Grant 13/RC/2106) and co-funded under the European Regional Development Fund, by the European Commission as part of the FALCON project (contract number 610879).
Introduction
Introduction

What is NOW for NLP/NMT?

- ACL 2019, Florence
  - 660 papers accepted
  - THEN: embed, encode, attend and predict (2015-2018)
  - NOW: off-the-shelf pre-training (huge datasets), fine-tuning on an in-domain dataset
  - BERT(ology):
What is NEXT for NLP/NMT?

Has the new paradigm (pre-trained embeddings, fine-tune) trivialized previous modeling innovations? (SMT -> bi-LSTM -> Transformer…)

- **NEXT?**
  - Infusing more knowledge? Knowledge graphs? Linguistic analysis?
  - Better/more challenging test sets?
  - Tackling the hard(er) problems?
    - Biases in the datasets, loss of “linguistic” richness
    => require extra knowledge + solution for specific scenarios
Introduction

Human Parity/Superhuman(?!?) Performance...

Meanwhile while translating with Google Translate [16/08/2019]

<table>
<thead>
<tr>
<th>English</th>
<th>French</th>
</tr>
</thead>
<tbody>
<tr>
<td>I ran.</td>
<td>Iran.</td>
</tr>
<tr>
<td>Iran.</td>
<td>Iran.</td>
</tr>
<tr>
<td>I trained.</td>
<td>Il pleuvait.</td>
</tr>
<tr>
<td>It rained.</td>
<td>Il pleuvait.</td>
</tr>
<tr>
<td>It ripped.</td>
<td>J'ai trébuché.</td>
</tr>
<tr>
<td>I tripped.</td>
<td>J'ai trébuché.</td>
</tr>
</tbody>
</table>

Issues related specifically to our systems (NMT)
Human Parity/Superhuman(?!?) Performance...

Meanwhile while translating with Google Translate [16/08/2019]

Linguistic issues
Introduction

Human Parity/Superhuman Performance?
Meanwhile while translating with Google Translate [16/08/2019]

Anna fehlt ihrem Kater.
Anna fehlt ihrem Katze.

Anna is missing her hangover.
Anna is missing her cat.

Linguistic issues
Presentation Outline

- Linguistics and MT
- Gender and MT
- Related Work
- Compilation of Datasets
- Experimental Setup and Results
- Conclusions and Future Work
Linguistics and MT
Linguistics and MT

Hype vs Reality:
Where do linguistics come into play?

Bi-LSTM, Transformer
- **Subject-Verb Number Agreement in PB-SMT**
  
  *Solved for NMT?*
  
  We and you *sing*.
  
  *Nous chantons et nous* [GNMT]
  
  I always thought *you* were *nice people*.
  
  *J'ai toujours pensé que tu étais des gens sympas.* [GNMT]

- **Tense/Aspect in PB-SMT/NMT**
  
  *Solved for NMT?*
  
  I *liked* school. ⇔ At that moment, I *liked* school.
  
  *J'ai aimé l'école.* ⇔ *A ce moment-là, j'aimais l'école.* [GNMT]
- Integrating syntactic/semantic features into NMT
  Factored NMT integrating specific CCG-tags + more general supersenses
  => faster convergence, marginal BLEU score improvement at convergence, difficult to evaluate.

- Gender in NMT
  This man is a nurse. ⇔ Cet homme est une infirmière.
  She is our best surgeon. ⇔ Elle est notre meilleur chirurgien.

- Loss of Linguistic richness
  Loss of more specialized words, near-synonyms...
Gender in MT
Gender in MT

Gender Terminology

- **Natural Gender:**
  - Masculine, Feminine, Neuter

“Gender based on the **sex** or, for neuter, the lack of sex of the referent of a noun, as English girl (**feminine**) is referred to by the feminine pronoun she, boy (**masculine**) by the masculine pronoun he, and table (**neuter**) by the neuter pronoun it.”

Gender Terminology

- **Grammatical Gender** (~ Noun class):
  - Masculine - Feminine - Neutral
  - Animate - Inanimate

“It is a bleak Day. Hear the Rain, how he pours, and the Hail, how he rattles; and see the Snow, how he drifts along, and oh the Mud, how deep he is! Ah the poor Fishwife, it is stuck fast in the Mire; it has dropped its Basket of Fishes; and its Hands have been cut by the Scales as it seized some of the falling Creatures; and one Scale has even got into its Eye. And it cannot get her out. “

Mark Twain, “A Tramp Abroad”,
“The Awful German Language”
A Quick Problem Sketch

■ A simple example:

I am happy!
A Quick Problem Sketch

- A simple example:

I am happy!

Je suis heureux! 50%

Je suis heureuse! 50%

[Natural Gender]
A Quick Problem Sketch

- A simple example:

  I am happy!

  Je suis heureux!

  Je suis heureuse!

- “Let the data decide”?

  50% 50%

  [Natural Gender]
A Quick Problem Sketch

- A simple example:

```
I am happy!
```

```
Je suis heureux!
```

```
Je suis heureuse!
```

50% 50%

- “Let the data decide”? 

Europarl EN-FR

```
33% 67%
```

[Natural Gender]
“Let the data speak”

I am beautiful.
I am a surgeon.
I am a beautiful surgeon.
I am good-looking.
I am a teacher.
I am a good-looking teacher.

Open in Google Translate
“Let the data speak”

I am beautiful.
I am a surgeon.
I am a beautiful surgeon.
I am good-looking.
I am a teacher.
I am a good-looking teacher.

Soy hermoso.
Soy cirujano.
Soy una hermosa cirujana.
Soy guapo.
Yo soy un profesor.
Soy una profesora guapa.
"Let the data speak"

I am beautiful.
I am a surgeon.
I am a beautiful surgeon.
I am good-looking.
I am a teacher.
I am a good-looking teacher.

Soy hermoso.
Soy cirujano
Soy una hermosa cirujana.

I am smart.
I am beautiful.
I am beautiful but not smart.

Je suis intelligent.
Je suis beau.
Je suis belle mais pas intelligente.
“Let the data speak”

I am beautiful.
I am a surgeon.
I am a beautiful surgeon.
I am good-looking.
I am a teacher.
I am a good-looking teacher.

Soy hermoso. (M)
Soy cirujano (M)
Soy una hermosa cirujana. (F)

I am smart.
I am beautiful.
I am beautiful but not smart.

Je suis intelligent. (M)
Je suis beau. (M)
Je suis belle mais pas intelligente. (F)
Introduction

“Let the data speak”

I am beautiful.
I am a surgeon.
I am a beautiful surgeon.
I am good-looking.
I am a teacher.
I am a good-looking teacher.

Open in Google Translate

I am smart.
I am beautiful.
I am beautiful but not smart.

I’m happy with my pink toy.
I’m happy with my blue toy.
Introduction

“Let the data speak”

I am beautiful.
I am a surgeon.
I am a beautiful surgeon.
I am good-looking.
I am a teacher.
I am a good-looking teacher.

Soy hermoso.
Soy cirujano.
Soy una hermosa cirujana.

I am smart.
I am beautiful.
I am beautiful but not smart.

Je suis intelligent.
Je suis beau.
Je suis belle mais pas intelligente.

I’m happy with my pink toy.
I’m happy with my blue toy.
Other (related) Issues:

щастлива съм. Edit
shtastliva süm.

Je suis heureux
Other (related) Issues:

щастлива съм. Edit

shtastliva süm.

Je suis heureux (M)
Introduction

Other (related) Issues:

<table>
<thead>
<tr>
<th>Bulgarian</th>
<th>French</th>
</tr>
</thead>
<tbody>
<tr>
<td>Щастлива съм.</td>
<td>Je suis heureux</td>
</tr>
</tbody>
</table>
### Other (related) Issues:

- The speaker is a woman.
- The speaker is my wife.
- The speaker is my sister.
- The speaker is Maja.

- The politician is a woman.
- The nurse is a man.

<table>
<thead>
<tr>
<th>English</th>
<th>French</th>
</tr>
</thead>
<tbody>
<tr>
<td>The speaker is a woman.</td>
<td>L'orateur est une femme.</td>
</tr>
<tr>
<td>The speaker is my wife.</td>
<td>L'orateur est ma femme.</td>
</tr>
<tr>
<td>The speaker is my sister.</td>
<td>L'orateur est ma sœur.</td>
</tr>
<tr>
<td>The speaker is Maja.</td>
<td>L'orateur est Maja.</td>
</tr>
<tr>
<td>The politician is a woman.</td>
<td>Le politicien est une femme.</td>
</tr>
<tr>
<td>The nurse is a man.</td>
<td>L'infirmière est un homme.</td>
</tr>
</tbody>
</table>

- स्वागतिक स्वभाव.
- I am happy.

- Je suis heureux. (M)
- shtastliva sūm.
Introduction

Other (related) Issues:

- The speaker is a woman.
- The speaker is my wife.
- The speaker is my sister.
- The speaker is Maja.

- The politician is a woman.
- The nurse is a man.

- We are beautiful.
- We are very beautiful.

- Le politique est une femme.
- L'infirmière est un homme.
- Nous sommes beaux.
- Nous sommes très belle.
Other (related) issues:

- The speaker is a woman.
- The speaker is my wife.
- The speaker is my sister.
- The speaker is Maja.
- The politician is a woman.
- The nurse is a man.

- Nous sommes beaux. (M)
- Nous sommes très belle (F sg)
Related Work
Related Work

Gender in Linguistics
- Differences male/female language

Gender in Computational Linguistics
- Bias (Gender, Racial…) in NLP:
  - ‘Debiasing’ techniques for word embeddings
  - Counterfactual Data Augmentation (CDA)
- Personalization
  - Domain adaptation
Related Work

Male and Female Language

■ “Language and Woman’s place” (Lakoff, 1973)
  ○ Male and Female discourse
    - Female discourse: more warm, compassionate, polite… (Park et al., 2016)
    - Different preferences syntactic structures (Mondorf, 2002; Newman et al., 2008; Coates, 2015)

=> Contradictory evidence (Price and Graves, 1980), quantitative experiments remain controversial (Hellinger and Motschenbacher, 2015)

! Cross-genre gender prediction
Bias in Natural Language Processing

- **Problem:**
  - Bias in Word Embeddings
  - Bias in Algorithms (?)
    Prates et al. (2017), Zhao et al. (2017), Lu et al. (2018),

- **Proposed Solutions**
  - Debiasing Word Embeddings
    Bolukbasi et al. (2016)
  - (Counterfactual) Data Augmentation
    Zhao et al. (2019)

- **Problems Proposed Solutions**
  - Only superficial removal
    Gonen and Goldberg (2019)
Related Work

Personalization in Machine Translation

- **Statistical Machine Translation**
  - Need for more personalized SMT (Mirkin et al., 2014)
  - Domain Adaptation (Rabinovich et al., 2017)
    - Gender as a domain
    - No improvements over baseline
    - Analysis of gender traits in HT and MT

- **Neural Machine Translation**
  - Female/Male tags (Vanmassenhove, 2018)
  - Extreme Personalization (Michel and Neubig, 2018)
    Adapting the bias of the output softmax to each particular user of the MT system
  - Google Translate (2018)

**More recently [ACL 2019, Gender bias in NLP]**
Moryossef et al. (2019), Stanovsky et al (2019), Habash et al. (2019)
Related Work

- **Google Translate**
  - December 2018
  - Masculine/Feminine translations for:
    - (Most) single words:
      - English $\Rightarrow$ French, Spanish, Italian and Portuguese
    - Phrases and sentences:
      - Turkish $\Rightarrow$ English
  - Google AI Blog:
    - Turkish $\Rightarrow$ English:

  ![Diagram](image)

  1. Detect gender-neutral queries
  2. Generate gender-specific translations
  3. Check for accuracy

  Gender-neutral or not? $\langle 2\text{MALE} \rangle$
  $\langle 2\text{FEMALE} \rangle$
  Identical apart from gender-related changes?
Compilation of Datasets
Compilation of Datasets

- Large dataset with speaker information
  - Europarl source files:
    - Speaker tags
    - Date of session
    - MEPs meta-information available (Rabinovich et al. 2017)
  => Retrieve gender, age …

“Madam President, as a former health care professional…”
## Datasets

<table>
<thead>
<tr>
<th>Languages</th>
<th># sents</th>
<th>Languages</th>
<th># sents</th>
</tr>
</thead>
<tbody>
<tr>
<td>EN–BG</td>
<td>306,380</td>
<td>EN–IT</td>
<td>1,297,635</td>
</tr>
<tr>
<td>EN–CS</td>
<td>491,848</td>
<td>EN–LT</td>
<td>481,570</td>
</tr>
<tr>
<td>EN–DA</td>
<td>1,421,197</td>
<td>EN–LV</td>
<td>487,287</td>
</tr>
<tr>
<td>EN–DE</td>
<td>1,296,843</td>
<td>EN–NL</td>
<td>1,419,359</td>
</tr>
<tr>
<td>EN–EL</td>
<td>921,540</td>
<td>EN–PL</td>
<td>478,008</td>
</tr>
<tr>
<td>EN–ES</td>
<td>1,419,507</td>
<td>EN–PT</td>
<td>1,426,043</td>
</tr>
<tr>
<td>EN–ET</td>
<td>494,645</td>
<td>EN–RO</td>
<td>303,396</td>
</tr>
<tr>
<td>EN–FI</td>
<td>1,393,572</td>
<td>EN–SK</td>
<td>488,351</td>
</tr>
<tr>
<td>EN–FR</td>
<td>1,440,620</td>
<td>EN–SL</td>
<td>479,313</td>
</tr>
<tr>
<td>EN–HU</td>
<td>251,833</td>
<td>EN–SV</td>
<td>1,349,472</td>
</tr>
</tbody>
</table>

**Table 1**: Overview of annotated parallel sentences per language pair
Compilation of Datasets

- Analysis EN-FR Annotated Dataset

  Sentences per age group

![Bar chart showing the distribution of sentences per age group for males and females.]

- 67.39% (M) vs 32.61% (F)

  Age groups:

  - Male: 20-30 (60%), 30-40 (60%), 40-50 (60%), 50-60 (60%), 60-70 (60%), 70-80 (60%), 80-90 (60%)
  - Female: 20-30 (10%), 30-40 (10%), 40-50 (10%), 50-60 (10%), 60-70 (10%), 70-80 (10%), 80-90 (10%)

  Total:

  - Male: 0.71%
  - Female: 43.76%
Experimental Setup
Experimental Setup

- **Experimental Setup**
  - OpenNMT-py toolkit (Klein et al. 2017)
  - Sequence to sequence encoder-decoder with LSTM as the recurrent unit (Bahdanau et al. 2014; Cho et al., 2014; Sutskever et al., 2014)
  - BPE to reduce OOV
  - Trained for 13 epochs: Best system selected for evaluation

- **Datasets**
  - EN-FR  EN-DA
  - EN-IT  EN-DE
  - EN-PT  EN-NL
  - EN-ES  EN-FI
  - EN-EL  EN-SV

- **Example**
  - “FEMALE Madam President, as a…”
  - “MALE …”
Results
### General Testsets:

<table>
<thead>
<tr>
<th>Systems</th>
<th>EN</th>
<th>EN-TAG</th>
</tr>
</thead>
<tbody>
<tr>
<td>FR</td>
<td>37.82</td>
<td>39.26*</td>
</tr>
<tr>
<td>ES</td>
<td>42.47</td>
<td>42.28</td>
</tr>
<tr>
<td>EL</td>
<td>31.38</td>
<td>31.54</td>
</tr>
<tr>
<td>IT</td>
<td>31.46</td>
<td>31.75*</td>
</tr>
<tr>
<td>PT</td>
<td>36.11</td>
<td>36.33</td>
</tr>
<tr>
<td>DA</td>
<td>36.69</td>
<td>37.00*</td>
</tr>
<tr>
<td>DE</td>
<td>28.28</td>
<td>28.05</td>
</tr>
<tr>
<td>FI</td>
<td>21.82</td>
<td>21.35*</td>
</tr>
<tr>
<td>SV</td>
<td>35.42</td>
<td>35.19</td>
</tr>
<tr>
<td>NL</td>
<td>28.35</td>
<td>28.22</td>
</tr>
</tbody>
</table>

**Table 2:** BLEU scores for the 10 baseline (denoted with EN) and the 10 gender-enhanced NMT (denoted with EN-TAG) systems. Entries labeled with * present statistically significant differences (p < 0.05). Statistical significance was computed with the MultEval tool (Clark et al., 2011).
Male vs Female test sets

<table>
<thead>
<tr>
<th>Test Sets</th>
<th>EN</th>
<th>EN-TAG</th>
</tr>
</thead>
<tbody>
<tr>
<td>FR (M)</td>
<td>37.58</td>
<td>38.71*</td>
</tr>
<tr>
<td>FR (F)</td>
<td>37.75</td>
<td>38.97*</td>
</tr>
<tr>
<td>FR (M1)</td>
<td>39.00</td>
<td>39.66*</td>
</tr>
<tr>
<td>FR (F1)</td>
<td>37.32</td>
<td>38.57*</td>
</tr>
</tbody>
</table>

Table 3: BLEU-scores on EN–FR comparing the baseline (EN) and the tagged systems (EN–TAG) on 4 different test sets: a test set containing only male data (M), only female data (F), 1st person male data (M1) and first person female data (F1). All the improvements of the EN-TAG system are statistically significant (p < 0.5), as indicated by *.
## Successes and failures

<table>
<thead>
<tr>
<th>(Ref)</th>
<th>En tant que <em>vice-président</em>...</th>
<th>(M)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(BASE)</td>
<td>En tant que <em>vice-présidente</em>...</td>
<td>(F)</td>
</tr>
<tr>
<td>(TAG)</td>
<td>En tant que <em>vice-président</em>...</td>
<td>(M)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>(Ref)</th>
<th>... je suis <em>heureuse que</em>...</th>
<th>(F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(BASE)</td>
<td>... je suis <em>heureux que</em>...</td>
<td>(M)</td>
</tr>
<tr>
<td>(TAG)</td>
<td>... je suis <em>heureuse que</em>...</td>
<td>(F)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>(Ref)</th>
<th>je suis <em>gênée que</em>...</th>
<th>(M)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(BASE)</td>
<td>je suis <em>embarassé que</em>...</td>
<td>(F)</td>
</tr>
<tr>
<td>(TAG)</td>
<td>je suis <em>embarassé que</em>...</td>
<td>(F)</td>
</tr>
</tbody>
</table>
Results

■ Side-effects

(Ref) Je *pense* que … (Ref) J’ ai plusieurs *remarques*…
(BASE) Je *crois* que… (BASE) J’ ai un nombre de *commentaires*…
(TAG) Je *pense* que… (TAG) J’ ai plusieurs *remarques*..

Remarks
- Both correct translations
- Enriched system picks ‘preferred’ variant
  ~ Different preferences: constructions, word choices etc.
  ~ Frequency list:
    => “penser” (more neutral) vs “croire” (~male)

- Gender of the translator?
Side-effects

(Ref) Je *pense* que … (Ref) J’ ai plusieurs *remarques*…

(BASE) Je *crois* que… (BASE) J’ ai un nombre de *commentaires*…

(TAG) Je *pense* que… (TAG) J’ ai plusieurs *remarques*..

Remarks
- Both correct translations
- Enriched system picks ‘preferred’ variant
  ~ Different preferences: constructions, word choices etc.
  ~ Frequency list:
    => “penser” (more neutral) vs “croire” (~male)

**IS THIS SOMETHING WE WANT?**
- Gender of the translator?
Results

- Problem with approach
  - Controllability:
    Syntactic agreement with natural gender of the speaker
    => Inconsistencies in the output
  - Side-effects:
    Word choices, (syntactic constructions?)
    => Difficult to systematically analyse differences between
    male/female speech, different topics covered, ‘bag-of-words’
    analysis...

- Why?
  - Syntax harder to learn than semantics?
  - Fundamental issue of distributional semantics?

- What now?
  - Hybrid approaches (knowledge!)
Conclusions and Future Work
Is gender a symptom of another underlying issue?

Overgenerating of male/female nouns (referring to professions) even when taking into account the already existing bias in the data (here demographic data)? $\sim \rightarrow$ algorithmic bias?

$\Rightarrow$ overgeneralization of observed ‘patterns’
$\Rightarrow$ loss of lexical/linguistic richness?, lack of diversity in our translations?
Conclusions

- **BLEU:**
  Significant improvements over sota baseline systems:
  - For some language pairs

- **Manual Analysis:**
  - Not always consistent
  - Different word choices?

Future Work

- **Compile different dataset:**
  - e.g. OpenSubtitles, TED Talks…

- **Alternative approaches:**
  - Post-processing? Incorporating syntactic knowledge? Hybrid systems?

- What about ‘you’, ‘we’, ‘they’ and the other related issues?
Thank you for your attention!