Screening Gender Transfer in Neural Machine Translation
Guillaume Wisniewski, Lichao Zhu, Nicolas Ballier, François Yvon

To cite this version:
Guillaume Wisniewski, Lichao Zhu, Nicolas Ballier, François Yvon. Screening Gender Transfer in Neural Machine Translation. BlackBoxNLP 2021, Nov 2021, Punta Cana, Dominican Republic. hal-03993451

HAL Id: hal-03993451
https://hal.science/hal-03993451
Submitted on 16 Feb 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution 4.0 International License
**Screening Gender Transfer in Neural Machine Translation**

Guillaume Wisniewski\(^1\), Lichao Zhu\(^1\), Nicolas Ballier\(^2\), François Yvon\(^3\)

\(^1\)LLF, Université de Paris, F-75013 Paris, France
\(^2\)CLILLAC-ARP, Université de Paris, F-75013 Paris, France
\(^3\)LISN, Université Paris-Saclay & CNRS, 91403 Orsay, France

---

**Information flow within an encoder/decoder architecture**

- **key steps in interpreting NMT systems**
  - which informations are captured by the decoder?
  - which informations are transferred from the source to the target?
- **how:** study the transfer of gender information from French to English
  - using probes to find where this information is represented;
  - using causal models to determine when this information is used.

**Gender Transfer between French and English**

- **challenge**
  - in French: gender = property of all nouns; agreement rules within noun phrase
  - in English: gender = only in rare constructs involving human agents and pronoun reference
- **focus on the following pattern:**
  - [DET] \[N\] a terminé son travail.
  - The [N] has finished [PRO] work.
  - [DET] = occupational noun either feminine or masculine
  - [PRO] = French possessor
  - Dataset of 3,394 parallel sentences following this pattern
- **Hypothetical paths for transferring gender information from French to English**
  - (a) direct influence = cross-lingual attention;
  - (b) indirect influence = monolingual encoding of gender in the representation of the English noun;
  - (c) indirect influence = cross-lingual attention to the French possessive adjective.

**In the source**

| layer | a | terminé | son | travail | eos |
|-------|---|---------|-----|---------|----|
| 1     | 80.4% | 75.1% | 80.6% | 76.4% | 59.5% | 73.3% |
| 2     | 85.8% | 80.8% | 81.6% | 78.3% | 87.6% | 88.3% |
| 3     | 89.5% | 88.2% | 89.2% | 82.0% | 86.5% | 87.6% |
| 4     | 90.8% | 89.3% | 90.6% | 85.9% | 87.5% | 85.6% |
| 5     | 90.4% | 89.3% | 90.4% | 85.5% | 86.4% | 85.2% |
| 6     | 91.0% | 89.3% | 90.0% | 86.0% | 86.4% | 85.1% |

**In the target**

- **target sentence not as ‘regular’ as source sentences**
  - (predicted by MT system)

  ⇒ accuracy of the probe computed over all tokens of the translation hypothesis

| layer | decoder | the | all | tokens |
|-------|---------|-----|-----|--------|
| 1     | 89.5%  | 71.6% |
| 2     | 92.0%  | 76.3% |
| 3     | 91.8%  | 78.1% |
| 4     | 90.9%  | 79.1% |
| 5     | 89.5%  | 82.4% |
| 6     | 87.7%  | 84.7% |

**Manipulating Representations**

- **linguistic probe:** predict the gender of the French occupational noun from a source/target word representation
  - simple binary classification problem
  - evaluation: accuracy

- **evaluation:** distribution of pronoun in translation hypothesis

| intervention | English pronoun % | sentences |
|--------------|-------------------|-----------|
| none         | her 13.4%         |           |
| his          | 57.1%             |           |
| other        | 29.5%             |           |
| feminine     | her 17.3%         |           |
| his          | 56.8%             |           |
| other        | 25.9%             |           |
| gender-neutral her 13.2%       | |           |
| other        | 29.4%             |           |
| his          | 57.4%             |           |
| masculine    | her 13.8%         |           |
| other        | 29.2%             |           |
| his          | 57.0%             |           |

- representations of son are not the only evidence used during the generation of the translation hypothesis
  - path (c) has only a limited influence

**Conclusions**

- **Contributions:** new dataset \(\oplus\) two techniques (probing & manipulating)
- **Conclusions:**
  - gender information in the representation of all tokens representations built by the encoder and the decoder
  - choice of English pronoun distributed
- **future work:**
  - generalization to other language & syntactic divergences
  - identify which information is used to choose the English pronoun

**Acknowledgments**

This work was partially funded by the NeuroViz project subsidized by the Île-de-France Region, and supported in part by the 2020 émergence research project SPEC-TRANS.

**Contact information**

guillaume.wisniewski@u-paris.fr

**Code & Corpus**

https://github.com/neuroviz/neuroviz/tree/main/blackbox2021