Long to reign over us: A Case Study of Machine Translation and a New Monarch

Rebecca Knowles and **Samuel Larkin**
National Research Council Canada
{Rebecca.Knowles, Samuel.Larkin}@nrc-cnrc.gc.ca

Abstract

Novel terminology and changes in terminology are often a challenge for machine translation systems. The passing of Queen Elizabeth II and the accession of King Charles III provide a striking example of translation shift in the real world, particularly in translation contexts that have ambiguity. Examining translation between French and English, we present a focused case-study of translations about King Charles III as produced both by publicly-available MT systems and by a neural machine translation system trained specifically on Canadian parliamentary text. We find that even in cases where human translators would have adequate context to disambiguate terms from the source language, machine translation systems do not always produce the expected output. Where we are able to analyze the training data, we note that this may represent artifacts in the data, raising important questions about machine translation updates in light of real world events.

1 Introduction

With the passing of Queen Elizabeth II on September 8, 2022, King Charles III became the first King of Canada in over 70 years. Given official bilingualism (English and French) in Canada, this raised a natural question of how machine translation (MT) systems – particularly those trained on data collected from Canadian government sources, which forms a disproportionately large amount of publicly available data for this language pair (Bowker and Blain, 2022) – might perform on terminology related to the new sovereign. We hypothesized that systems trained on relatively recent parliamentary text might produce errors due to both linguistic features of French and English as well as the paucity of references to kings in the training data. We expand on this, showing that not only is this the case for MT systems trained solely on Canadian parliamentary data; these errors also appear (albeit less frequently) in the output of large publicly available MT systems. In this work we will distinguish between *errors*, where context (and world knowledge of the two sovereigns in question) would be sufficient for a human translator to translate correctly, and other potential artifacts of the data where there is insufficient context at the sentence level to translate unambiguously.

This work can be viewed as a narrowly-focused miniature challenge set (Isabelle et al., 2017), aiming to examine a specific intersection of MT challenges through a recent known example: world knowledge (or lack thereof) and changes in the state of the world, dataset imbalances, a subset of the different ways in which grammatical gender and the pronouns and inflections used for the referent affect translation for this language pair, and asymmetries in translation ambiguity.[1] By keeping this tight focus, we are able to point out some areas in which MT is not yet "solved," even for this highly-resourced language pair. On the other hand, this tight focus on both the language pair and the specific case of text about these two monarchs limits the scope of what this work addresses; we provide a brief discussion of more general related work in the following section and additional notes in the Limitations section.

2 Related Work

How to incorporate (new or updated) terminology into MT has long been an area of interest, from compound noun splitting and subword models (Koehn and Knight, 2003; Sennrich et al., 2016) to rapidly incorporating terminology from external sources like lexicons or dictionaries (Arthur et al., 2016; Kothur et al., 2018). Recently, there has been a focus on handling novel terminology resulting from the COVID-19 pandemic, including a shared task (Alam et al., 2021), the release of

[1]We release the annotated data as supplementary material.

Findings of the Association for Computational Linguistics: ACL 2023, pages 6589–6598

targeted datasets (Anastasopoulos et al., 2020), and evaluations of MT performance on related terminology (Bowker and Blain, 2022).

There has been work on bias, imbalance, and gender-inclusivity in coreference resolution (Rudinger et al., 2018; Zhao et al., 2018; Cao and Daumé III, 2020), on linguistic gender in MT (Vanmassenhove et al., 2018), on incorporating coreference into MT to improve pronoun and gender inflection translation (Miculicich Werlen and Popescu-Belis, 2017; Saunders et al., 2020), and on benchmarks for and analysis of gender in MT (Currey et al., 2022; Savoldi et al., 2021).

There has also been analysis of and attempts to mitigate language pair asymmetries in linguistically conveyed information, such as by incorporating factors (Koehn, 2005; Avramidis and Koehn, 2008; Mager et al., 2018). Here, while some of our examples might benefit from such approaches, many would require additional context beyond the sentence. The topic of additional context in MT and its evaluation remains an open area (Tiedemann and Scherrer, 2017; Junczys-Dowmunt, 2019; Castilho et al., 2020), and within this realm there has been work specifically done on anaphora resolution (Voita et al., 2018).

3 Linguistic and Grammatical Notes

In French, nouns are grammatically classed as masculine or feminine, and adjectives, articles, and determiners take inflected forms that agree with the nouns in terms of number and grammatical gender. The noun *Majesté* (majesty) is feminine (f). The form of address *Sa Majesté* on its own is ambiguous to translate into English, as the feminine form of the third person singular possessive determiner *Sa* agrees with the feminine noun *Majesté*, without regard to the specific referent. Depending on the referent's pronouns, *Sa* could be correctly translated as various singular third person pronouns such as *Her*, *His*, or *Their* (singular; for plural Their, the French source would be *Leurs Majestés*). Without additional context, like the sovereign's title and name, we expect current MT systems to almost always produce *Her Majesty* as a translation, due to the preponderance of that translation in the data. The question arises: will MT systems use information about words like King/*Roi* or the frequency with which the name Charles is associated with masculine pronouns to produce translations like *His Majesty King Charles III*? We anticipate more translation errors in the French–English translation direction, but examine both translation directions.

Table 1 illustrates five cases into which the examples in our data fall. In case A, a pair of words is unambiguously translated in either translation direction within this domain, such as *Reine* and Queen. Sometimes French has two forms of a noun like *souverain* (m)/*souveraine* (f) but English only has one unmarked form, sovereign, making the translation unambiguous in the French to English direction only (case B). In case C, the translation from English is unambiguous both because *Sa* is used for either He or She in our data and because its translation is governed by the grammatical gender of the noun *Majesté*, and does not depend on the referent. As described earlier, the reverse (case D) requires additional context when translating from French into English (due to the agreement between the possessive determiner and the grammatical gender of the noun in French, and the selection of the English pronoun based on the referent). The reverse direction of case B is case E, where additional context is required to translate the English word sovereign into French.[2]

4 MT Systems

4.1 Online Systems

We used MT output from two publicly available translation tools, Google Translate (`https://translate.google.com/`)[3] and Bing Translator (`https://www.bing.com/translator`). For the latter, we specify "French (Canada)". We do not know if they have been updated since September 8, 2022. All translations were re-run on January 13, 2023, to use recent versions.

4.2 Internal

We also use French-English (FR-EN) and English-French (EN-FR) MT systems trained on data from the Canadian Hansard (House of Commons),

[2]This highlights a subtle distinction between the last four cases. Those using the example of sovereign have a noun whose linguistic gender marking is selected based on the referent, whereas in the case of B and E, the English pronoun is selected based on the referent but the French determiner is selected based on the grammatical gender of the noun; e.g., if you wanted to describe "her path", the choice to use the translation *voie* (f) or *chemin* (m) would determine whether to translate *her* as *sa* or *son*, respectively.

[3]Google Translate offers (binary) gender-specific translations in some language pairs for some sentences (Johnson, 2020); while we did not test this for all sentences in our set, most did not appear to offer these options, even when it would be appropriate to do so (likely due to length/complexity).

A. *Reine*/Queen	bidirectionally unambiguous translation (EN↔FR)
B. *souverain(e)*	unambiguously translated as sovereign (FR→EN)
C. His Majesty	unambiguously translates as *Sa Majesté* (EN→FR)
D. *Sa Majesté*	requires context for *Sa*, e.g., *Sa Majesté la Reine* (Her Majesty the Queen)
E. sovereign	requires context to translate as *souverain* (m) or *souveraine* (f)

Table 1: Examples of unambiguous translations and translations that require context for disambiguation.

which we refer to as Internal. We trained Transformer models (Vaswani et al., 2017) using Sockeye (Hieber et al., 2018) version 2.3.14 on over 5.6 million lines of text drawn from sessions 39-1 (2006) to 43-2 (2021),[4] all predating the accession of the new sovereign. These systems were built for other projects, and were only used for decoding (no additional training was performed). See Appendix A for more details.

5 Experiments

We collect a small amount of existing parallel text from several sources: the text of the Prime Minister's statement regarding King Charles III's accession to the throne, text from the Canadian Hansard (proceedings of the House of Commons), and the Royal Anthem (*God Save the Queen/King*).[5]

From these, we manually extract terms that vary in at least one language based on whether they would refer to Queen Elizabeth II or King Charles III. This includes pronouns/determiners, adjectives and nouns that are grammatically marked for gender, and their names and titles. After translation, an author of this paper annotated each term in context to mark if it had been translated as expected. This was done via first automatically checking for string matches, followed by a manual check of all examples and notes on the cases where the expected translation was not found. Table 2 shows a summary of the Hansard and Prime Minister's Announcement settings in which at least one system produced a translation error.

5.1 Prime Minister's Announcement

The text of the prime minister's announcement on the accession to the throne is 7 lines long and contains 24 terms that we examine. Of these, 10 are bidirectionally unambiguous (e.g., "Queen Elizabeth II"). In the English to French direction another 11 are unambiguous, while the other 3 have enough context that a human translator could translate them unambiguously. In the French to English direction, another 3 are unambiguous, and the remaining 11 have sufficient context for a human translator.

	Bing	Google	Internal
EN→FR PM Ann.	24/24	23/24	24/24
FR→EN Hansard-King	3/3	3/3	1/3
FR→EN PM Ann.	22*/24	23/24	17/24

Table 2: Fraction of accurate term translations. Anthem and sets where all systems performed perfectly omitted. *In the case of FR→EN PM Announcement, Bing produces one translation that is rephrased such that a pronoun is not needed; we count this as correct.

In the English to French direction, across all the systems and terms, there is only one case where the correct translation is not produced: an instance of Google producing *souverain* where it ought to produce *souveraine* in a sentence that references both monarchs (see Table 3).[6]

As expected, it is in the French to English direction that we see the most errors. All systems perform accurately on the 13 unambiguous translations. On the 11 remaining terms that have adequate context for translation, the Bing system correctly translates 8 (also producing two instances of "Her Majesty" rather than "His," and one valid translation that is rephrased such that a pronoun is not needed), the Google system accurately translates 10 (with the same Her/His Majesty substitution), and the Internal system only accurately translates 4 (with 6 Her/His Majesty substitutions and 1 substitution of them for him).

5.2 Hansard

We selected sentences from the Hansard, all of which referenced the Queen. There were 9 from the training data and 2 from held out data. Across these sentences, there are a total of 13 terms that we examine. Two of the terms are bidirectionally unambiguous to translate. In the English to French

[4]Hansard text is available at https://www.ourcommons.ca/documentviewer/en/house/latest/hansard.

[5]The title quote comes from the Royal Anthem, which can be found online at https://www.canada.ca/en/canadian-heritage/services/royal-symbols-titles/royal-anthem.html along with its French translation.

[6]The Internal system produces *souveraine* twice in a row in the same sentence, but a full discussion of all types of translation errors is beyond the scope of this short paper.

English	*French*	*MT*
While we continue to mourn the loss of Canada's longest-reigning **sovereign**, **Her Majesty Queen Elizabeth II**, we also look to the future with the proclamation of the accession of **His Majesty King Charles III** as **Sovereign** of Canada.	Alors que nous continuons de pleurer la perte de la **souveraine** qui a régné le plus longtemps sur le Canada, **Sa Majesté la reine Elizabeth II**, nous nous tournons vers l'avenir au moment de la proclamation de l'accession au trône de **Sa Majesté le roi Charles III**, **souverain** du Canada.	Alors que nous continuons à pleurer la perte du plus ancien ***souverain*** du Canada, **Sa Majesté la reine Elizabeth II** [...] (Google) [...] the proclamation of ***Her Majesty*** **King Charles III**, the **sovereign** of Canada. (Internal)

Table 3: Examples of translation errors. Terms in bold, errors in red and italics.

direction, the remaining 11 are all also unambiguous to translate. In the French to English direction, 10 would require additional context to guarantee translation accuracy, while 1 has sufficient context for a human translator to translate it accurately. For the two bidirectionally unambiguous translations and for the one contextually informed translation in the French to English direction, we also produce alternative versions of the same segments modified to reference King Charles III.

In translating English to French, all terms are translated correctly for both monarchs by all MT systems. In translating French to English, all translations of text about Queen Elizabeth II are correct (modulo capitalization or apostrophe differences) for all systems. All 10 of the sentences that would require additional context to guarantee translation accuracy were examples with *Sa Majesté*, and all were translated as "Her Majesty" by all three MT systems. Note that we would especially expect this to be true of the training data for the Internal MT system, since this training data had already been observed and possibly memorized by the system, but it is also the case for the one sentence with this phrase from the held out data. The one sentence where the context would have been sufficient for a human translator included the phrase *Sa Majesté le roi Charles III*; both publicly available systems handled this correctly, while the Internal system translated it as "Her Majesty King Charles III." The internal system also once left *Roi* untranslated.

Nevertheless, these results are somewhat weakened by the fact that much of the data is from the training data for the Internal system, and may also be incorporated in the public MT systems; possibly implicating memorization.

5.3 Anthem

The Royal Anthem has a number of references to the Queen or King (depending on the version) as well as pronouns and (in the case of French) inflected adjectives. As song lyrics, the MT output is often adequate (the Internal system struggles the most) but not poetic. We present only the following high-level comments: when translated line by line, all systems default to masculine inflections of the adjectives, but when lines are merged to provide additional coreferent context, the adjectives are inflected to match the referent.

6 Discussion and Conclusions

Perhaps unlike the introduction of COVID-19 terminology (where an entire new topic or domain is rapidly introduced to the translation landscape), the accession of a new monarch may cause a shift in terminology in an existing domain, in this case one with 70 years of history.[7] As we expected, ambiguous terms tend to be translated in a way that likely corresponds to the imbalance in the training data (i.e., in the feminine, as referencing Queen Elizabeth II); this also highlights the need for context (whether document-level or external) that is often required for accurate translation when there is an asymmetry in what information is (un)marked across a language pair. Though they likely contain many Canadian translations (see Bowker and Blain (2022)), we cannot examine the public system training data, only the Internal system data. While there are thousands of mentions of the Queen in the Hansard training data, there are only hundreds of references to kings, and only 36 instances of the term "His Majesty" as compared to 882 instances of "Her Majesty". In our Internal system, an additional consequence of this is subword segmentation of words like *roi*: the word was fully segmented into its three characters, rather than appearing as a single token in the vocabulary, likely contributing to observed errors. We also found that even in sentences that would have adequate context for a human translator (with knowledge of the forms

[7] The recent terminology shift in English from Turkey to Türkiye may provide another example for study; as of May 2, 2023, Google and Bing exhibited different results when translating the country's name from French into English.

of address for the two monarchs), the MT systems sometimes made errors. Without examining the inner workings of the systems, the fact that this occurred primarily in sentences with references to both monarchs leaves open the question of whether this is a problem of erroneous implicit coreference resolution, imbalance in the training data around these particular terms, or a combination of the two. Nevertheless, while accuracy in term translation is high overall, these striking errors where context ought to be sufficient serve as a warning that even in high-resource language pairs, history and data maintain a strong influence.

Limitations

This work has a narrow focus: small-scale analysis, translation between one language pair (French and English), examining terminology around two real-world public figures (whose forms of address are both highly prescribed and publicly documented),[8] in a specific newsworthy event (the accession to the throne of a new king after over 70 years of data and translation about a queen). First, the scale of the analysis is quite small, so it does not examine in detail questions of frequency of errors, distributions of errors, or statistical significance. While this work raises issues that may be relevant for consideration across other language pairs, the relevance of the specific linguistic conventions discussed here will vary across language pairs, and certainly do not cover the full range of asymmetries in linguistically encoded information (see, e.g., Mager et al. (2018)). Due to the prescribed forms of address of the two monarchs in question, this work only examined translations related to a small subset of terms (e.g., "His"/"Her", *Reine*/*Roi*) and does not examine performance on terms used related to other individuals or to other third person singular pronouns or forms of address that could be used by a monarch. The specific circumstances (a 70 year reign of a sovereign of a country with an official bilingualism policy and this particular set of linguistic features) means that we may not expect these results to generalize to other potentially comparable scenarios. Lastly, we cannot examine the training data used for the public models, so we can only draw conclusions related to training data about the internal system.

[8] See, e.g., https://www.canada.ca/en/canadian-heritage/services/protocol-guidelines-special-event/styles-address.html

Ethics Statement

This work included data collection, specifically the selection of test sentences from public-facing Canadian government websites as well as the annotation of machine translation errors. This was performed by one of the authors, who reads both languages and received confirmation on French-related questions from fluent colleagues.

While this work does focus on two identifiable individuals, these two individuals are public figures and the data sources that we select are official sources of public information about them (in fact, produced by governments of which they were/are the Heads of State). There is discussion in the NLP and MT literature of the harms of misgendering and of treating gender as a binary or immutable feature (Cao and Daumé III, 2020; Saunders et al., 2020). In this work, we focus on some aspects of grammatical gender that can be unrelated to an individual referent (e.g., Sa Majesté), as well as some aspects of linguistic gender that do have a tie to the referent (e.g., pronouns, inflection of adjectives). By choosing this particular case study of the accession of King Charles III after the passing of Queen Elizabeth II, this paper does focus on only two linguistic genders in French and English, because the current and past official formal forms of address of these two particular individuals are well-documented in this language pair by sources from their governments. We use the most recent available information for this, as linked in the footnote in the previous section. For a broader discussion of gender-inclusive language related to translation and this particular language pair, there are various sources on the topic,[9] and some of these conventions are changing.

From a computational cost perspective, this paper reused existing neural MT systems (publicly available systems and internal systems) rather thank training systems from scratch, and translated a very small amount of text.

Acknowledgements

We thank our colleagues and the anonymous reviewers for their feedback on this paper.

[9] E.g., https://www.noslangues-ourlanguages.gc.ca/en/writing-tips-plus/gender-inclusive-writing-correspondence

References

Md Mahfuz Ibn Alam, Ivana Kvapilíková, Antonios Anastasopoulos, Laurent Besacier, Georgiana Dinu, Marcello Federico, Matthias Gallé, Kweonwoo Jung, Philipp Koehn, and Vassilina Nikoulina. 2021. Findings of the WMT shared task on machine translation using terminologies. In *Proceedings of the Sixth Conference on Machine Translation*, pages 652–663, Online. Association for Computational Linguistics.

Antonios Anastasopoulos, Alessandro Cattelan, Zi-Yi Dou, Marcello Federico, Christian Federmann, Dmitriy Genzel, Francisco Guzmán, Junjie Hu, Macduff Hughes, Philipp Koehn, Rosie Lazar, Will Lewis, Graham Neubig, Mengmeng Niu, Alp Öktem, Eric Paquin, Grace Tang, and Sylwia Tur. 2020. TICO-19: the Translation initiative for COvid-19. arXiv:2007.01788.

Philip Arthur, Graham Neubig, and Satoshi Nakamura. 2016. Incorporating discrete translation lexicons into neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1557–1567, Austin, Texas. Association for Computational Linguistics.

Eleftherios Avramidis and Philipp Koehn. 2008. Enriching morphologically poor languages for statistical machine translation. In *Proceedings of ACL-08: HLT*, pages 763–770, Columbus, Ohio. Association for Computational Linguistics.

Lynne Bowker and Frédéric Blain. 2022. When French becomes Canadian French: The curious case of localizing covid-19 terms with Microsoft Translator. *The Journal of Internationalization and Localization*, 9(1):1–37.

Yang Trista Cao and Hal Daumé III. 2020. Toward gender-inclusive coreference resolution. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4568–4595, Online. Association for Computational Linguistics.

Sheila Castilho, Maja Popović, and Andy Way. 2020. On context span needed for machine translation evaluation. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3735–3742, Marseille, France. European Language Resources Association.

Anna Currey, Maria Nadejde, Raghavendra Reddy Pappagari, Mia Mayer, Stanislas Lauly, Xing Niu, Benjamin Hsu, and Georgiana Dinu. 2022. MT-GenEval: A counterfactual and contextual dataset for evaluating gender accuracy in machine translation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4287–4299, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Felix Hieber, Tobias Domhan, Michael Denkowski, David Vilar, Artem Sokolov, Ann Clifton, and Matt Post. 2018. The sockeye neural machine translation toolkit at AMTA 2018. In *Proceedings of the 13th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 200–207, Boston, MA. Association for Machine Translation in the Americas.

Pierre Isabelle, Colin Cherry, and George Foster. 2017. A challenge set approach to evaluating machine translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2486–2496, Copenhagen, Denmark. Association for Computational Linguistics.

Melvin Johnson. 2020. A scalable approach to reducing gender bias in google translate. https://ai.googleblog.com/2020/04/a-scalable-approach-to-reducing-gender.html.

Marcin Junczys-Dowmunt. 2019. Microsoft translator at WMT 2019: Towards large-scale document-level neural machine translation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 225–233, Florence, Italy. Association for Computational Linguistics.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of Machine Translation Summit X: Papers*, pages 79–86, Phuket, Thailand.

Philipp Koehn and Kevin Knight. 2003. Empirical methods for compound splitting. In *10th Conference of the European Chapter of the Association for Computational Linguistics*, Budapest, Hungary. Association for Computational Linguistics.

Sachith Sri Ram Kothur, Rebecca Knowles, and Philipp Koehn. 2018. Document-level adaptation for neural machine translation. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pages 64–73, Melbourne, Australia. Association for Computational Linguistics.

Manuel Mager, Elisabeth Mager, Alfonso Medina-Urrea, Ivan Vladimir Meza Ruiz, and Katharina Kann. 2018. Lost in translation: Analysis of information loss during machine translation between polysynthetic and fusional languages. In *Proceedings of the Workshop on Computational Modeling of Polysynthetic Languages*, pages 73–83, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Lesly Miculicich Werlen and Andrei Popescu-Belis. 2017. Using coreference links to improve Spanish-to-English machine translation. In *Proceedings of the 2nd Workshop on Coreference Resolution Beyond OntoNotes (CORBON 2017)*, pages 30–40, Valencia, Spain. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

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.

Danielle Saunders, Rosie Sallis, and Bill Byrne. 2020. Neural machine translation doesn't translate gender coreference right unless you make it. In *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*, pages 35–43, Barcelona, Spain (Online). Association for Computational Linguistics.

Beatrice Savoldi, Marco Gaido, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2021. Gender Bias in Machine Translation. *Transactions of the Association for Computational Linguistics*, 9:845–874.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Jörg Tiedemann and Yves Scherrer. 2017. Neural machine translation with extended context. In *Proceedings of the Third Workshop on Discourse in Machine Translation*, pages 82–92, Copenhagen, Denmark. 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, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, pages 5998–6008.

Elena Voita, Pavel Serdyukov, Rico Sennrich, and Ivan Titov. 2018. Context-aware neural machine translation learns anaphora resolution. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1264–1274, Melbourne, Australia. 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.

A Internal System Details

We trained Transformer models (Vaswani et al., 2017) using Sockeye (Hieber et al., 2018) version 2.3.14 and cuda-10.1. We used Sockeye's default value of 6 encoder/ 6 decoder layers, 8 attention heads, a model size of 512 units with a FFN size of 2048, the Adam (Kingma and Ba, 2015) optimizer, label smoothing of 0.1 and a cross-entropy-without-softmax-output loss. The whole validation set (2000 sentences) is used during validation. We optimized for BLEU (Papineni et al., 2002) using Sockeye's default of `sacreBLEU-1.4.14` (Post, 2018). Every 1000 updates, we evaluate BLEU on the validation and perform early stopping if there is no improvement after 32 checkpoints. Only sentence pairs with at most 200 tokens on both source and target side are used during training. Gradient clipping was set to absolute, the initial learning rate set to 0.0002, batch size set to 8192 tokens and we used weight tying and vocabulary sharing. Training was performed on 4 Tesla V100s, while inference used 1. During inference, the beam size is set to 5. The training data consisted of over 5.6 million lines of text drawn from sessions 39-1 (2006) to 43-2 (2021), with validation and additional held out data drawn exclusively from 43-2. Hansard text is publicly available at `https://www.ourcommons.ca/documentviewer/en/house/latest/hansard`. These systems were built for other projects, and were simply used to decode the selected texts (no additional training was performed for this paper).

B Test Data Sets

The Prime Minister's statement (with link to the French version) is found at:

https://pm.gc.ca/en/news/statements/2022/09/10/statement-prime-minister-proclamation-accession-his-majesty-king-charles

The segments from House of Commons were subselected from sentences available at https://www.ourcommons.ca/documentviewer/en/house/latest/hansard

The Royal Anthem data is collected from https://www.canada.ca/en/canadian-heritage/services/royal-symbols-titles/royal-anthem.html

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
1 (Introduction), 6 (Discussions and Conclusion), Limitations (unnumbered section)

☑ A2. Did you discuss any potential risks of your work?
Ethics section

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Abstract and 1 (Introduction)

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

Annotated data released.

☑ B1. Did you cite the creators of artifacts you used?
4 & 5, Appendix B

☑ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Annotated data release provides link to terms of use regarding unofficial, non-commercial reproduction of House of Commons text.

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
We do discuss in the limitations section what conclusions should not be drawn from our annotated data.

☑ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
The data is public data (press release, Parliamentary text, anthem) that uniquely identifies two individual public figures. We do not anonymize it because the study focuses on the translation of those figures' titles and coreferents. We do not use any data that provides private personal information.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
1,3,5

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
4,5

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

C ☑ Did you run computational experiments?

We performed machine translation.

☒ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
We present information about the parameters of the models used (Appendix A) but do not include full details of computational budget, as these MT systems were trained for prior unrelated work and only used here to decode an extremely small set of test sentences.

☐ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Not applicable. Not applicable; using existing MT systems and analyzing output.

☐ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Not applicable. This is an extremely small-scale study. We report counts of errors.

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Appendix A

D ☑ Did you use human annotators (e.g., crowdworkers) or research with human participants?

5. One of the authors annotated the data.

☒ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
One of the authors performed the data annotation, and did not provide self with a written set of instructions for annotation.

☑ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
One of the authors performed the data annotation. We did not provide information about the author's salary or demographic information.

☐ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
Not applicable. An author performed the annotation in awareness of the use of the data.

☒ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Author annotated MT errors.

☒ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
There was only one annotator (one of the authors).