university of groningen

faculty of science and engineering

Incorporating Automatically Generated Genre Labels into Neural Machine Translation Systems

Mălina Chichirău

university of groningen / faculty of science and engineering

University of Groningen

Credit Assignment for Incorporating Automatically Generated Genre Labels into Neural Machine Translation Systems

Master's Thesis

To fulfill the requirements for the degree of
Master of Science in Artificial Intelligence
at University of Groningen under the supervision of

Internal Supervisor: Prof. Dr. Stephen Jones (Bernoulli Institute, University of Groningen)
External Supervisors: Dr. Antonio Toral (CLCG, University of Groningen) and
Dr. Rik van Noord (CLCG, University of Groningen)

Mălina Chichirău (s3412768)

March 21, 2024

Contents

Acknowledgments

I would like to thank my supervisors for their continuous support and patience during the course of my research project. I am grateful to Taja Kuzman for sharing with me her work and expertise on genre classification. Additionally, I thank the Center for Information Technology of the University of Groningen for their support and for providing access to the Hábrók high-performance computing cluster. Finally, I want to thank my friends and family, especially my greatest supporter - my father - for inspiring me to study Artificial Intelligence and always encouraging me to pursue my academic goals.

Abstract

State-of-the-art neural machine translation (NMT) systems are often highly specialized for a certain type of text, referred to as a domain. However, the definition of a domain is still ambiguous in literature, with many studies focusing more on the provenance of the texts used for training rather than on their properties, under the assumption that texts from a single source have similar characteristics. Nevertheless, reliable information about the provenance of texts, especially in the case of web-crawled corpora, is not always available.

This study explores whether domains can be described based on text genres, defined by non-topical properties such as function, style, or register that can be automatically inferred from texts. We experiment with training genre-specific NMT systems for translating from English to Icelandic, Croatian, and Turkish. When tested on a holdout dataset, the genre-specific systems tend to outperform general NMT systems and NMT systems specialized in other genres, on their target genre. However, the results are not replicable on external datasets. Furthermore, we use special tokens that indicate the genres in the training data to train general genre-aware NMT systems. But, we find no significant difference compared to the equivalent genre-agnostic systems. Therefore, we conclude that genres are not sufficiently informative to define reliable translation domains that can be utilized across different corpora.

1 Introduction

Deep learning techniques enabled the development of powerful Neural Machine Translation (NMT) systems. However, these systems require large amounts of parallel training data to attain competitive performance and are often highly specialized for a certain **domain**. For instance, although several systems submitted to the Conference on Machine Translation (WMT) achieve human parity on the news translation task they were trained for, they still struggle to accurately translate biomedical texts (Saunders, 2021).

However, the definition of a **translation domain** is still ambiguous in the literature, with many studies focusing more on the provenance of the texts used for training rather than on their properties (i.e. genre, topic, style etc.). Even the test sets used for the WMT shared tasks, which are considered to be high-quality annotated data, are selected based on their provenance (from online newspapers)[1] and labeled as news despite being a combination of actual news reports and interviews or editorial pieces, and varying greatly in style and topics.

Therefore, **domains** are largely defined under the assumption that texts that originate from a single source have homogenous properties. Consequently, NMT systems trained on data from a particular source improve their knowledge of that domain, which leads to higher-quality translations. However, this assumption does not hold in the case of texts that are automatically collected from the Internet. The original source of web-crawled data is often uncertain, as texts can be easily fully or partially copied or translated. Furthermore, texts published on the Internet are not as strictly curated or moderated as they would be in printed press. Thus, the properties and quality of texts can vary greatly even if they are retrieved from the same website. Nevertheless, web-crawled data is relatively cheap and fast to collect, and it is particularly useful for building NMT systems for under-resourced languages (Kuzman & Ljubešić, 2023).

This thesis proposes to redefine **translation domains** based on text properties that can be automatically identified from the data itself. Therefore, we bypass the uncertainty of data provenance and focus instead on the form, function, and purpose of the texts, which we refer to as **genre** (Kuzman & Ljubešić, 2023). Genre information was shown to be useful in several natural language processing tasks such as part-of-speech tagging (Giesbrecht & Evert, 2009), summarization (Stewart & Callan, 2009), and zero-shot dependency parsing (Müller-Eberstein, van der Goot, & Plank, 2021). However, only a limited number of studies incorporated genres into machine translation research and they employed older, statistical machine translation systems (van der Wees, Bisazza, Weerkamp, & Monz, 2015; van der Wees, Bisazza, & Monz, 2018).

Therefore, we use state-of-the-art transformer-based NMT systems (Vaswani et al., 2017) to study the effectiveness of defining **translation domains** according to genre labels that we generate automatically from web-crawled data. First, to test whether genres are consistent and homogenous across different data sources, we are experimenting with training **genre-specific** systems on the web-crawled data and testing them on texts labeled with their target genres, from a holdout test set, but also from external datasets. Secondly, we investigate whether general NMT systems - trained on a variety of genres - benefit from incorporating genre information into their input. Consequently, we compare the quality of the translations produced by **genre-aware** and **genre-agnostic** systems. We test our

[1] According to the description of the data sets retrieved from: `https://www.statmt.org/wmt21/translation-task.html`.

approach to training **genre-aware** models in different scenarios: we train NMT models from scratch exclusively on the web-crawled data, we use the web-crawled data to further train an NMT baseline model (fine-tune), and we aggregate the web-crawled data into documents to further train an NMT baseline model.

1.1 Research Questions

This thesis focuses on two main research questions, with several secondary questions. Both questions inquire about the use of **automatically generated genre labels to define translation domains** in neural machine translation. The first question addresses training **genre-specific systems**, while the second one focuses on **comparing genre-aware and genre-agnostic systems** in different scenarios.

RQ 1. Can automatically generated genre labels be used to fine-tune **genre-specific** neural machine translation systems that outperform general systems on their target genre?

- RQ 1.1. Can these **genre-specific** systems perform well on external data sets classified with the same genre?
- RQ 1.2. Can the results be replicated across different languages?

RQ 2. Can automatically generated genre labels be used to train **genre-aware** neural machine translation systems that outperform equivalent **genre-agnostic** systems?

- RQ 2.1. Can we use the genre labels to train genre-aware systems **from scratch** that would outperform equivalent (genre-agnostic) systems?
- RQ 2.2. Can we use the genre labels to **fine-tune** genre-aware systems from pretrained (genre-agnostic) systems?
- RQ 2.3. Can we use the genre labels to **fine-tune** on *document-level* genre-aware systems from pretrained (genre-agnostic) systems?
- RQ 2.4. Can the results be replicated across different languages and across external data sets?

1.2 Thesis Outline

Relevant literature is discussed in Section 2, which offers a brief introduction to the machine translation field (Section 2.1), followed by a more in-depth analysis of the differences between genre and domain (Section 2.2), an explanation of domain adaptation techniques (Section 2.3) and an overview of studies that included genres in machine translation research (Section 2.4).

The Methods Section 3 covers data preprocessing (Section 3.1), genre classification (Section 3.2), the NMT models we use (Section 3.3), an overview of the experiments we conduct (Section 3.4) and an explanation of the evaluation metrics (Section 3.5).

The Results Section 4 follows the structure of the Experiments Section 3.4, first presenting the findings of the **genre-specific** models experiments (Section 4.1), followed by the results of the comparison between **genre-aware** and **genre-agnostic** models in different scenarios: trained from scratch (Section 4.2.1), fine-tuned on MaCoCu data (Section 4.2.2), and fine-tuned on document-level MaCoCu

data (Section 4.2.4).

Finally, the Discussion Section 5 analyses the findings of the **genre-specific** experiments (Section 5.1), followed by an analysis of the comparison between **genre-aware** and **genre-agnostic** models (Section 5.2), a discussion of the limitations of our research (Section 5.3) and some ideas for future research (Section 5.4).

2 Background Literature

This section provides an overview of the research works relevant to this paper. We briefly explain the development of machine translation, noting important and recent advancements in Section 2.1. Then, we investigate further the distinction between genres and domains in Section 2.2. Section 2.3 discusses domain-adaptation techniques for NMT systems. Lastly, Section 2.4 gives an overview of the few studies that incorporated genres into machine translation systems.

2.1 Machine Translation

Machine translation (MT) refers to the endeavor of using machines to translate written text or speech from one natural language into another. From a methodological standpoint, there are two main approaches to MT: rule-based methods and corpus-based methods (Wang, Wu, He, Huang, & Church, 2022). The rule-based methods require bilingual dictionaries and manually written rules (that dictate grammar, word order, etc.), which and are difficult or impossible to re-use for other language pairs. Corpus-based methods require larger amounts of data, in the form of bilingual corpora, and more computational power. These methods include example-based machine translation (EBMT), statistical machine translation (SMT), and neural machine translation (NMT).

EBMT works through analogies, by retrieving sentences that are similar to the source sentence (that has to be translated), and using the bilingual corpus to translate them (Nagao, 1984). Consequently, the quality of the EBMT heavily relies on finding similar sentences to the source, which is not always possible, since not all linguistic phenomena can be present in a corpus (Wang et al., 2022). SMT models were proposed by Brown et al. (1990), they learn the probability distributions of words or word combinations from the bilingual corpora. Therefore, they calculate the most likely translation, given a sentence. Although more versatile than previous methods, SMTs still struggle with translating polysemous words, different word order, and grammar between language pairs, statistical anomalies, out-of-vocabulary words, etc, and often employ additional models and heuristics to overcome their shortcomings (Wang et al., 2022).

NMT systems represent the latest developments in the field of machine translation. They use a single large neural network that directly translates a sentence from the source language into the target language (Sutskever, Vinyals, & Le, 2014). Bahdanau, Cho, and Bengio (2014) incorporated attention mechanisms into a recurrent neural network (RNN) architecture. A second breakthrough was represented by the Transformer architecture introduced by Vaswani et al. (2017), which relies solely on attention mechanisms to process, eliminating the need for recurrent connections in processing relations between words. Modern NMT systems map the source sentence into a vectorial representation that is used, along with attention mechanisms, to generate a translation (Wang et al., 2022). Therefore, both the vectorial representation and the translation knowledge (grammar, word order, relations between words, etc.) are learned from the training corpora. Consequently, NMT systems still rely on the quality of the training data and how similar it is to the testing data.

NMT systems learn a vectorial representation (referred to as “embedding“) for each *token* in their vocabulary. *Tokens* are often words, but they can also correspond to sub-words such as frequent character sequences and morphemes (e.g. un-, -atic, -al, -ly) that bound to words to change their meaning or functions. Additional special *tokens*, which are not inferred from the corpus, but imposed by the

programmer, are used to mark unknown words that cannot be derived from the existing *tokens*, the start and end of a sentence, or convey information about the domain of the source or target sentence (Stergiadis, Kumar, Kovalev, & Levin, 2021; Tars & Fishel, 2018). In this study, we will be using special tokens to signal the genre of the source sentences, in an attempt to "teach" the NMT systems that different genres require different linguistic features. The technical aspects of the NMT systems that are employed will be discussed in Section 3.3.

2.2 Genre and Domain

As previously discussed, research into enabling more versatile NMT systems has mainly focused on domain adaptation. Koehn and Knowles consider that a domain is *"a corpus from a specific source, and [it] may differ from other domains in topic, genre, style, level of formality, etc."* (2017, p.28).

However, Saunders (2021) argues that the *provenance* (i.e. source or origin) of a text should not be treated as the only domain marker, and should be instead considered alongside other text properties such as topic and genre. Firstly, the *provenance* of unseen test data might often be unknown, therefore selecting training data from an "appropriate" source to train a domain-specific model would be challenging. Secondly, the topic and genre are more reliable as domain markers since they can be identified by directly analyzing the texts, whereas *provenance* constitutes metadata that might be incomplete or inaccurate and impossible to recover. Finally, using *provenance* as the only domain marker would imply that language domains are discrete and exclusionary since a document can only be assigned a single source. However, texts can serve multiple purposes at once (e.g. both to promote a product and to inform readers how to use it), and multiple topics can be addressed even in the span of a single sentence. Therefore, textual properties such as topics and genres are more informative and descriptive as domain markers than the *provenance* of the text.

The topic of a text is regarded as the general subject addressed (van der Wees et al., 2015), and it can be determined on different levels from broad such as "sports" to narrow such as "tennis" or even the "2022 Wimbledon Championships". On the other hand, the genre is considered to be "complementary to the topic, covering the non-topical text properties function, style, and text type." (van der Wees et al., 2015, p. 561). Therefore, the genre plays a role in how the information is presented to a reader: subjectively or objectively, whether a text is persuasive or informative, whether it is entertaining or monotone, etc. Similarly to topics, genres can be broad such as "non-fiction" or narrower such as "editorial" or "news reports". Moreover, genres consistently differ in their use of language: type-token ratios, verb tenses, types of frequently used pronouns and modal verbs, the number of Wh-questions and Wh-clauses, etc. (Sharoff, 2020). Some of these genre-specific language features are consistent across languages, while others are language-specific (Sharoff, 2020).

Furthermore, the characteristics of genres change over time (Mehler, Sharoff, & Santini, 2010), even in the case of genres which often subject to editorial scrutiny. For instance, over the last two centuries, the rate of nouns increased while the rate of verbs decreased in academic writing (Biber & Gray, 2016). These changes are more obvious in the case of documents retrieved from the Internet, as conventional genres had to adapt to this new medium and new genres emerged. Due to the lack of uniform content moderation, there are also more variations within a single genre class, there is more overlap between genres, more hybrid texts (texts of different genres embedded into a single web page), and some texts might be too short to contain genre-specific language features (Kuzman &

Ljubešić, 2023).

This study will focus on adapting machine translation systems for different genres. Genre information was shown to be a useful addition in several natural language processing tasks (Giesbrecht & Evert, 2009; Stewart & Callan, 2009; Müller-Eberstein et al., 2021), however, only a limited number of machine translation studies used it (van der Wees et al., 2015, 2018). Therefore, we aim to look further into defining **domains** according to genes, rather than data **provenance**. We believe that the textual properties that genres encompass should be more informative and reliable when determining **translation domains**.

However, we will also be aware of the *provenance* of data when dividing it into training, development, and testing sets. Therefore, while the distribution of genres will be similar between the data sets, the *provenance* of the data (i.e. internet domain it originates from) will be different, such that we will test whether the genres alone are informative enough to train systems that generalize well on data from different sources.

2.3 Domain-Adaptation for NMT

Often there is insufficient domain-specific data available for training specialized NMT systems from scratch. However, as pointed out by Saunders (2021) domains are not discrete, they often overlap. Therefore, training on more data, which covers a wider range of domains can be beneficial to NMT systems. Such multi-domain models have been found to outperform domain-specific ones trained only on subsets of the available data (Britz, Le, & Pryzant, 2017).

However, domain-agnostic NMT models still struggle to identify by themselves the relevant linguistic features of the domains in the training data such that they can produce translations that fit a target domain (Saunders, 2021). For instance, Hovy, Bianchi, and Fornaciari (2020) found that commercial multi-domain NMT systems such as DeepL, Google Translate, and Bing, fail to preserve the stylistic features of the texts they translate, leading to translations that are more likely to be attributed to older males, than the original sentences. Furthermore, Emelin, Titov, and Sennrich (2019) found that multi-domain NMT systems fail to properly disambiguate words based on a deeper understanding of the domain of the target sentence, and instead rely heavily on inappropriate lexical correlations. Therefore, domain-adaptation techniques seek to train domain-specific NMT systems, starting from more general and versatile multi-domain systems.

Fine-tuning is a computationally efficient approach to domain adaptation that nevertheless leads to considerable improvements (Luong & Manning, 2015). It involves further training a multi-domain NMT system on a smaller dataset, belonging to the target domain. This is especially useful when there is little data available for the target domain. Vu and Moschitti (2021) trained a domain classifier to identify from a larger corpus the sentences that matched their target domain and used them to build a dataset to fine-tune their domain-specific models. In our case, the genre classifier will act in a similar manner, ensuring that our genre-specific models will be fine-tuned on relevant data, belonging to a given genre class. By using a single genre classifier on the entire dataset, we ensure consistency between the genres of the training and testing data.

Various studies used transformer-based NMT architectures (Vaswani et al., 2017), and included domain labels either as a single inline *token* or as an embedded feature that was combined with each *to-*

ken in the input (Kobus, Crego, & Senellart, 2017; Tars & Fishel, 2018), noting improved performance over domain-agnostic models. Pham, Crego, and Yvon (2021) compared several domain adaptation techniques, including Kobus et al.'s, on 7 domains which originated from different corpora.[2] While their results support the fact that the multi-domain models outperformed domain-agnostic ones on certain domains, they also found that the specialized models for each domain performed significantly better than multi-domain models. They also noted that the inline **tokens** approach achieved slightly better results than the embedded feature approach. Therefore, we will be following the methodology of Kobus et al. (2017) and Tars and Fishel (2018), and train multi-genre NMT models that incorporate genre-labels as inline *tokens*.

2.4 Using Genres in Machine Translation

To our knowledge, genre information was not used for training NMT systems, which are generally adapted to new domains, defined based on the *provenance* of data sets (Pham et al., 2021; Tars & Fishel, 2018; Mino et al., 2020; Chu, Dabre, & Kurohashi, 2017), implicitly assuming that genres are uniform across texts originating from a given source. While this assumption holds reasonably well when it comes to data from curated corpora, it becomes questionable in the case of web-crawled corpora, where a more fine-grained definition of domains might be needed due to the lack of moderation and high variation in data quality. However, several studies did investigate the effects of genres on statistical machine translation, the previous mainstream paradigm in machine translation, and their findings will be discussed in this section.

van der Wees et al. (2015) trained a multi-domain Arabic-English statistical machine translation (SMT) system on a balanced data set, by controlling for two genres (user-generated and news) and five topics (culture, economy, health, politics, and security). When evaluating their model across the same genres and topics, they found both topic-related errors and genre-related errors, confirming that both the genres and topics influence the quality of machine translations.

Subsequently, van der Wees et al. (2018) experimented with genre-specific SMT systems for editorial, colloquial, news, and speech data. Their data was crawled from the internet, and they relied on keywords indicated by web pages to classify the data into genres. When cross-evaluating the genre-specialized systems, as expected, they performed best on the test sets of the genre they were trained on and outperformed the genre-agnostic baseline system. Therefore, it became apparent that choosing an appropriate genre-specific SMT system is important for the quality of translations.

Consequently, van der Wees et al. (2018) experimented with using a genre classifier on the test data, to simulate selecting a genre-specific system to translate unlabelled data. They trained an SVM genre classifier on a subset of their labeled dataset. They generally found the classifiers to be accurate, which led to similar translation quality as in the case of the labeled test data. For certain language pairs, the test sets included genres that were not present in the training data, which meant that the genre classifiers matched them with the most similar genre-specialized translation system. In such cases, the translation quality was slightly lower than that of the translations made by genre-agnostic systems but higher than that of the translations made by the other genre-specialized systems. There-

[2]UFAL Medical corpus for medical texts, the European Central Bank corpus for financial texts, The JRC-Acquis Communautaire corpus for legal texts, documentations for KDE, Ubuntu, GNOME, and PHP from Opus collection for IT texts, TED Talks for spoken texts, and the Koran for religious texts.

fore, they concluded that automatic genre classification is especially advantageous when translating data from domains/genres that do not perfectly match the training data.

Similar to van der Wees et al.'s (2018) approach, we will be using automatically-generated genre labels. However, we will be constructing the training, development, and testing sets according to the genre labels indicated by a pre-trained genre classifier, based on state-of-the-art language models. Therefore, the classifier is expected to be more accurate and unbiased by our training data. Furthermore, we will be using NMT systems rather than SMT and will be experimenting with genre-agnostic systems, multi-genre systems, and genre-specific systems.

3 Methods

3.1 Data

Our main dataset is the MaCoCu parallel corpora (Bañón et al., 2022). We chose as target languages for our experiments languages that are included in MaCoCu and belong to different language families. Therefore, we run experiments for English-Croatian, English-Turkish, and English-Icelandic.

Additionally, we use external test sets from Flores200 (Costa-Jussà et al., 2022) and WMT News Tasks from 2022 (Kocmi et al., 2022) for Croatian, from 2021 (Akhbardeh et al., 2021) for Icelandic, and from 2018 (Bojar et al., 2018) for Turkish. The Flores test sets contain Wikimedia articles, translated from English into other languages by professional translators. Therefore, the datasets are the same between languages. We also chose the most recent WMT test sets for each language included in our experiments. The test sets for Croatian and Icelandic are comprised of strictly news fragments, while the more recent Croatian test set is a general translation task. For the data set sizes and a breakdown of the genre within each test set, according to the X-GENRE classifier (Kuzman, 2022), see Table 4.

MaCoCu is a collection of parallel and monolingual corpora for under-resourced European languages. The data was gathered by automatically crawling top-level internet domains and was curated by filtering out sentences in non-target languages, boilerplates, duplicates, and low-quality texts. There are two releases of the MaCoCu corpora, with the second being smaller but aiming to be of higher quality. Therefore, when available, we use the second release of the parallel corpora.

The MaCoCu corpus contains the URLs of both the source and target sentences, the dates when the original documents were retrieved, and information about the position of the sentences within the documents. Since the sentence pairs were aligned automatically, using Bitextor (van der Linde, 2023), some pairs were retrieved from different websites and were aligned erroneously. Therefore, we use regular expressions to determine the internet domain from the URLs and we check that the domains coincide in each sentence pair. If this is not the case, we remove the pair from the data set. As a result, we filter leftover boilerplate texts, very short sentences that coincide between websites, but sometimes also correct matches between sentences retrieved from a current and an archived version of a website. We discarded 3.1% of the Croatian corpus, 8.1% of the Turkish corpus, and 10.7% of the Icelandic sentence pairs. Table 1 shows examples from the Icelandic MaCoCu parallel corpus of sentence pairs with mismatching domains.
Furthermore, a quality score is provided in the MaCoCu corpus, computed using the tool BicleanerAI (Zaragoza-Bernabeu, Ramírez-Sánchez, Bañón, & Ortiz Rojas, 2022), which indicates the likelihood of the sentence pairs being mutual translations. When dividing the corpus into training, development, and testing sets, we check that the distribution of the scores is similar between sets.

Additionally, the MaCoCU corpora contain an indication of the most probable translation direction between the sentence pairs, and whether the translation was made by a human or a translation system (van Noord, 2023). Previous research shows that translated text exhibits unique features and patterns that are not common in original texts and are referred to as *translationese* (Bizzoni et al., 2020). Furthermore, the presence of *translationese* in test data has been linked to an overestimation of the performance of NMT systems since *translationese* texts tend to be simplified and easier to automatically translate (Zhang & Toral, 2019; Bizzoni et al., 2020). Since the translator type and the

	Sentence	URL	Domain
English	The content of the amendment bill can be viewed here.	https://www.arsskyrsla.hugverk.is/articles-en/administrative-revocation-at-a-crossroads-a-look-back	arsskyrsla.hugverk.is
Icelandic	Unnt er að kynna sér nánar efni breytingarlaganna hér.	https://www.hugverk.is/um-okkur/frettasafn/stjornsysluleg-nidurfelling-senn-timamotum-litid-yfir-farinn-veg	hugverk.is
English	This website uses cookies to improve your experience.	http://1001arabian.net/media/magazines/bahrain_news.htm	1001arabian.net
Icelandic	Þessi vefsíða notar vefkökur (cookies) til að bæta upplifun þína og greina umferð um vefinn.	https://artasan.is/product/flourish-intimate-wash/	artasan.is

Table 1: Examples of mismatching domains, as inferred from URLs, for the English-Icelandic sentence pairs.

translation direction labels were generated automatically, and, therefore, do not represent the ground truth, no data is excluded based on these criteria. However, we check that the data splits are relatively balanced and there is no over-representation of likely translationese sentences in the testing and development datasets.

	Train	Dev	Test	Total
Sentence-level				
Croatian	1,250,976	14,046	19,011	1,284,033
Turkish	1,098,842	14,943	17,684	1,131,469
Icelandic	143,098	11,558	16,083	170,739
Document-level				
Croatian	130,750	1,286	1,587	133,623
Turkish	220,511	2,508	2,645	225,664
Icelandic	14,306	1,268	1,339	16,913

Table 2: Number of instances (sentence pairs or document pairs) from the MaCoCu corpora per data split.

The data split of the MaCoCu corpus is shown in Table 2. The data was split after it was labeled by the genre classifier - which is explained in Section 3.2. For the test sets to be meaningful, we tried to include at least 1000 sentences for each genre. In practice, this means that some genres tend to be over-represented in the testing and development data compared to the training data. This is often the case for *Prose/Lyrical* and *Forum* (Figure 1). Furthermore, the sources (web domains) of the sentences are different between splits, to simulate testing on sentences from different domains/corpora.

Some web domains contain several documents, labeled with different genres. Consequently, we included as many sentences from under-represented genres as possible, while avoiding discarding the sentences labeled with the over-represented genres from the same web domain. However, we use a threshold of 3000 sentences per genre for the development and test sets, to avoid the unnecessary use of computing resources. Therefore, we discarded 7.6% of the Croatian data, 1.4% of the Turkish corpus, and 3.1% of the Icelandic data. [3]

The document-level data is derived from the sentence-level data by aggregating the sentences in each data split according to their source (URL). However, due to how the data was preprocessed, there are likely gaps within the reconstructed documents, and the order of the sentences might not be correct. However, the order and the number of sentences are consistent between the source and the target data in our experiments. Furthermore, it is important to note that genres differ in the average document length. Particularly, *Prose/Lyrical* documents tend to be much longer than the average document, meaning they are under-represented in the dev and test sets now. Figure 10 from Appendix A shows the distribution of genres in the MaCoCu data set, aggregated on document level.

3.2 Genre Classification

The genre labels are produced by the X-GENRE classifier (Kuzman, 2022). The classifier was built by fine-tuning an XLM-RoBERTa language model (Conneau et al., 2019) on three annotated data sets: English CORE (Egbert, Biber, & Davies, 2015), English FTD (Sharoff, 2018), Slovene GINCO (Kuzman, Rupnik, & Ljubešić, 2022). Since each of these datasets uses different labeling conventions, the labels were aggregated into the following categories: *Information/Explanation, Instruction, Legal, News, Opinion/Argumentation, Promotion, Forum, Prose/Lyrical,* and *Other*. For a detailed correspondence between the X-GENRE labels and the English CORE, English FTD and Slovene GINCO labels, see Appendix B. Table 3 provides a short description of each genre, and a set of common features, adapted from the annotation guidelines of the GINCO corpus (Kuzman, Brglez, Rupnik, & Ljubešić, 2021).

In order to determine the genres in the MaCoCu corpus, the sentence pairs are aggregated into documents, and only the English documents are classified by the X-GENRE system. We first remove duplicated sentences by comparing the text and source listed for each sentence. We therefore try to preserve the integrity of documents, even though some sentences are duplicated between documents. Next, we aggregate sentences into documents according to their sources (ULRs) only if the English and the target language documents have the same number of sentences. We therefore avoid reconstructing documents from misaligned sources (e.g. a single URL is listed for the English sentences but there are two URLs listed for the corresponding Icelandic sentences).

The resulting documents are further filtered by imposing a threshold of at least 25 words per document. The documentation of the X-GENRE systems recommends that the classifier should be used on documents of at least 75 words. However, since documents in the MaCoCu corpus tend to be shorter, we experimented with other document sizes, as the alternative would have been discarding nearly half of the data. Therefore, we classify documents of at least 25 words. This means that we

[3]These values are computed after discarding sentence pairs with mismatching domains (explained above) and after discarding documents shorter than 25 words for genre classification (see Section 3.2).

Genre	Definition	Common Features
Forum	A text in which people discuss a certain topic in the form of comments.	subjective, 1st person informal language
Information/ Explanation	An objective text that describes or presents an event, a person, a thing, a concept etc. Its main purpose is to inform the reader about something	objective/factual, explains or defines a concept
Instruction	An objective text which instructs the readers on how to do something.	multiple steps/actions modality (must, need to)
Legal	An objective formal text that contains legal terms and is clearly structured. The name of the text type is often included in the headline (contract, rules, amendment, general terms and conditions, etc.).	objective/factual, 3rd person, specific terminology
News	An objective text which reports on an event recent at the time of writing or coming in the near future.	adverbs/adverbial clauses many proper nouns direct or reported speech
Opinion/ Argumentation	A subjective text in which the authors convey their opinion or their experience. It includes the promotion of an ideology and other non-commercial causes, but the main purpose of the text is not promotion.	exclamation marks subjective, 1st person
Other	A text that has no clear purpose or tangible features based on which it could be categorised.	quiz, survey, list, table of contents, worksheet
Promotion	A subjective text intended to sell or promote an event, product, or service. It addresses the readers, often trying to convince them to participate in something or to buy something.	usage of 2nd person comparative and superlative adjectives and adverbs
Prose/Lyrical	A text that consists of verses or a literary running text that consists of paragraphs. Has no other practical purpose than to give pleasure to the reader, it can be considered art.	Lyrics/poems/prayers figures of speech, adejectives

Table 3: Definition of genres and common features.

discarded 2.6% of the Croatian corpus, 4.5% of the Turkish corpus, and 2.6% of the Icelandic one.[4]. The resulting genre distribution (at the sentence level) is illustrated in Figure 1. Since the label *Other* does not refer to a cohesive genre class, and it is only intended to represent texts with no clear genre

[4]These percentages are computed after discarding sentence pairs with mismatching domains during pre-processing - see Section 3.1

markers, we excluded the data labeled with this label.

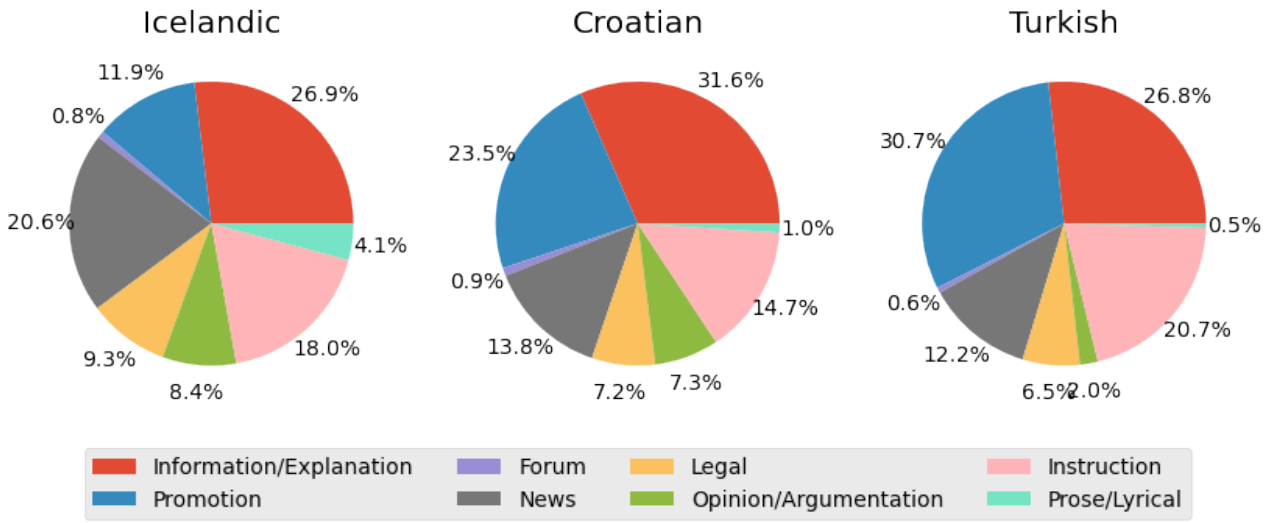

Figure 1: Genre distribution in the MaCoCu dataset.

In the case of the **external data sets**, we follow a similar procedure of first aggregating the sentences into documents using the URLs or sources provided by the data sets. However, we do not discard documents shorter than 25 words, as they belong to benchmark datasets. Similarly, documents labeled with *Other* are not discarded. A single document in the Flores dev set does not meet the length requirement.

The exact genre distribution of the external test sets is illustrated in Table 4. No documents were labeled as either *Legal* or *Prose/Lyrical*, which is not surprising since these datasets are mostly comprised of web articles (Flores) and news articles (WMT). Consequently, most of the instances from the Icelandic WMT 2021 News (92%) and Turkish WMT 2018 News (84%) test sets were appropriately labeled as *News*, with the second most popular genre being *Opinion/Argumentation* (7% and 11%, respectively). In the case of the Croatian test set from the WMT 2022 General MT shared task, the most popular genre is still *News* (38%), followed by *Opinion/Argumentation* (32%), *Forum* (19%), and *Promotion* (6%). The main genres in the Flores test sets are *Information/Explanation* (44% dev and 45% devtest), *News* (31% dev and 29% devtest), *Instruction* (16.5% dev and 13% devtest) and *Opinion/Argumentation* (6% dev and 10% devtest). Furthermore, most instances labeled as *News* were retrieved from the *Wikinews* web domain and most *Information/Explanation* were retrieved from *Wikibooks* and *Wikivoyage*. Overall, the genre labels are consistent with the sources listed by the test sets.

Furthermore, for some genres, there are very few examples (< 50) in the external data (highlighted in red in Table 4). Therefore, we are not evaluating our genre-specific models on these genres, as the results would not be reliable on such a small sample. However, they are still present in the test sets used for the multi-genre models, as we are evaluating the overall translation quality, not on particular genres.

	Flores dev	Flores devtest	Croatian WMT General MT 2022	Icelandic WMT News 2021	Turkish WMT News 2018
Forum	4	2	316	0	0
Information/Explanation	447	459	10	4	32
Instruction	165	133	58	0	33
News	309	294	635	919	2525
Opinion/Argumentation	56	104	530	77	348
Other	2	4	16	0	62
Promotion	13	15	106	0	0
Total	996	1011	1671	1000	3000

Table 4: Genre distribution in the external data sets, at sentence level. There are no sentences labeled as either *Legal* or *Prose/Lyrical*.

3.3 NMT Models

For our experiments, we use OPUS-MT models (Tiedemann & Thottingal, 2020), based on the Marian-NMT framework (Junczys-Dowmunt et al., 2018). However, we use the Huggingface PyTorch implementation, as we are also using Python code to run the experiments and data pre-processing. For the fine-tuning experiments, we use the OPUS-MT models pre-trained on the OPUS corpus, and for the models trained from scratch, we use the same architectures and re-initialize the weights.

The OPUS-MT models implement a transformer-based architecture (Vaswani et al., 2017) that uses both encoders and decoders. The models for Croatian and Icelandic experiments utilize encoders and decoders with 6 layers and 8 attention heads and a hidden size of 2048. The vocabulary size is 58647 tokens for English-Icelandic and 58879 tokens for English-Croatian, and both accept a maximum input length of 512. On the other hand, the model for English-Turkish is slightly larger, using 6 layers for the encoder and the decoder, but with 16 attention heads, and thus a hidden size of 4096. The vocabulary size is 57060, and although it can handle input sequences up to 1024 tokens, we do not change the default size which is set to only 512. All models use shared vocabularies between source and target languages.

There is no OPUS-MT model for Croatian alone, therefore, we use a multilingual model trained for several Slavic languages: Belarusian, Croatian, Macedonian, Czech, Russian, Polish, Bulgarian, Ukrainian, and Slovenian. Since this is a multilingual model, it requires language codes to tokenize the data correctly. The language code used for Croatian is $>>$ *hrv* $<<$, which is appended in front of the input English data. The language code is added in front of the genre tokens, in the case of genre-aware models. Furthermore, the language code is removed by the pre-trained tokenizer of the OPUS-MT models, as it is only used by the decoder when converting the output from ids to tokens and cleaning up the tokenization.

We perform hyperparameter tuning manually, experimenting with several values for learning rate, batch size, and gradient accumulation steps and testing on the dev split. The parameter values used

for training are shown in Table 5. Croatian models trained from scratch were trained for 15 epochs and Turkish models trained from scratch were trained for 8 epochs. All experiments involving fine-tuning were run for 5 epochs. We do not use early stopping, as we do not want the training time to be a confounding factor for our results. For reliability, we averaged results over 3 runs, thus each model was trained 3 times using seeds 1-3. Experiments were run on the high-performance computing cluster of the University of Groningen, using Nvidia A100 and Nvidia V100 GPUs.

	Huggingface model name	**From Scratch**	**Fine-tuning**
Croatian	opus-mt-en-sla	lr:1e-4 bsz:16 gac:2	lr:1e-5 bsz:16 gac:2
Turkish	opus-mt-tc-big-en-tr	lr:1e-5 bsz:16 gac:2	lr:1e-4 bsz:16 gac:2
Icelandic	opus-mt-en-is	–	lr:1e-5 bsz:16 gac:2

Table 5: Parameter values used during training from scratch and fine-tuning experiments. Due to limited data availability (see Table 2), Icelandic models cannot be trained from scratch. Abbreviations: lr = learning rate, bsz = batch size, gac = gradient accumulation steps.

3.4 Experiments

3.4.1 Genre-Specific Models

The first experiment relates to the first **research question RQ1**, and it involves fine-tuning pre-trained OPUS-MT models to become **genre-specific** models. In order to make a fair comparison between the performance of the **genre-specific** models for the different genres, we fine-tune the models on equal-sized train and dev sets for several genres.

Since there is very little data for *Forum* and *Prose/Lyrical* (see Figure 1), we do not train **genre-specific** models for these genres. However, we want to fine-tune models for *Legal*, as this genre is very different from the others. Therefore, we under-sampled the data from the other genres to match the amount of *Legal* data from the train and dev sets. Consequently, we train **genre-specific** models for *Information/Explanation*, *News*, *Promotion*, *Instruction* and *Legal* for all languages, and we additionally train a model for *Opinion/Argumentation* for Croatian, since the data is more abundant for this language.

To account for the fact that models often improve with additional training time, regardless of the nature of training data, we also fine-tune models on a randomized dataset, equal in size to the genre datasets. The data is sampled using the Pandas *sample* method, the exact distributions and data sizes are shown in Table 9, in Appendix D.

3.4.2 Genre-Aware vs. Genre-Agnostic Models

The second experiment corresponds to the **research question Q2** and aims to compare **genre-aware models** and **genre-agnostic models**, in several scenarios. This section first explains the differences between the types of genre-aware models and genre-agnostic models and then explains the scenarios

in which these models are compared.

Genre-aware models are trained using special tokens that indicate the genre of the training data. The tokens are added in front of the input English sentences and followed by a single whitespace character. Genre tokens are of the form $< promo >$, using a sequence of at most five letters between the "less than" and "greater than" signs, generally an abbreviation of the genre label (see Appendix C for a complete list of the tokens used). There are two types of genre-aware models implemented across all experiments, which will be referred to as **genre aware models** and **genre aware + tokens models**. In the case of the **genre-aware + tokens** models, the special genre tokens are manually added to their vocabulary to prevent the pre-trained tokenizers from splitting them or treating them as unknown tokens. In turn, the embeddings are also resized to match the new vocabulary size. All other aspects of the training process remain unchanged between the **genre aware** and **genre aware + tokens models**.

Genre-agnostic models are trained on the same data as the **genre-aware models**, but without the special genre tokens. Therefore, they are unaware of the genre of each sentence and have to infer the differences and similarities between genres from the data.

Genre-Aware vs Genre-Agnostic Models Trained on MaCoCu
This experiment compares models trained **from scratch** on the MaCoCu dataset, and it corresponds to **RQ 2.1**. As previously mentioned, these models use the architectures of the pre-trained OPUS-MT models, but the weights are re-initialized and trained from scratch on the MaCoCu data. Models are trained for 15 epochs, using the parameters shown in Table 5. We also experiment with training tokenizers on the MaCoCu data for Croatian. [5] We use a byte-pair encoding algorithm (Sennrich, Haddow, & Birch, 2016), implemented by the Python module of the SentencePiece framework (Kudo & Richardson, 2018), and the same vocabulary size as the pre-trained tokenizers. This experiment cannot be conducted for Icelandic, since there is too little data available to train an NMT model from scratch.

Genre-Aware vs Genre-Agnostic Models Fine-Tuned on MaCoCu
This experiment corresponds to **RQ 2.2**, and it compares OPUS-MT models fine-tuned on the MaCoCu data, either as **genre-agnostic**, **genre aware** or **genre aware + tokens** models. These models are trained on the entire MaCoCu datasets, but instead of being trained from scratch (**Experiment 2.1**), they use the OPUS-MT models as a baseline.

Genre-Aware vs Genre-Agnostic Models Fine-Tuned on Document-Level
Similarly to the previous experiment, we fine-tune OPUS-MT models, but the datasets are aggregated into documents. Therefore, we aim to see whether the genre labels are more informative for longer sequences, which are also more likely to contain genre markers. This experiment corresponds to **RQ 2.3**.

3.5 Evaluation Metrics

We are evaluating our models using standard metrics for translation, namely, COMET (Rei, Stewart, Farinha, & Lavie, 2020) and BLEU (Papineni, Roukos, Ward, & Zhu, 2002) scores. BLEU scores

[5] Since we performed the Croatian experiments first and found no benefits for training tokenizers, we did not experiment with training tokenizers for Turkish anymore to avoid wasting computing resources.

measure the similarity between machine-translated texts and a reference text using n-gram precision. We use the sacrebleu (Post, 2018) implementation, version 2.3.1.[6] COMET scores are generated by comparing candidate translations with a reference translation, while also taking the source text into account. They are produced by deep learning models, trained to predict machine translation quality. We use the (currently) default COMET model *Unbabel/wmt-22-comet-da*, which was trained on the direct assessments of the submissions to the WMT Machine Translation Conference from 2017 to 2020.

[6]The version signature is the following: nrefs:1|case:mixed|eff:no|tok:13a|smooth:exp|version:2.3.1.

4 Results

4.1 Genre-Specific Models

This section describes the findings of the first experiment, which involves fine-tuning OPUS-MT models to become **genre-specific** NMT systems. As previously mentioned, we fine-tune models for *Information/Explanation*, *Intruction*, *Legal*, *News*, *Promotion* for Croatian, Icelandic, and Turkish, and we fine-tune an additional model on *Opinion/Argumentation* for Croatian. Besides the genre-specific models, we fine-tune two general NMT models on a data set with randomly selected instances, equal in size to the genre-specific datasets. The distribution of genres in these randomized datasets can be seen in Table 9, in Appendix D. The genre-aware model we use is a **genre-aware + tokens** model, as discussed in Section 3.4.

Table 6 shows the difference in COMET scores between the baseline OPUS-MT models for Croatian, Turkish, and Icelandic and our fine-tuned models when tested on the MaCoCu test set and on the Flores devtest set. Furthermore, Appendix E includes plots of all models' performance across the MaCoCu and the external test sets, aggregated by genres. We organise our findings for the first experiment in the following sections: Section 4.1.1 focuses on the comparison between the **genre-specific** and general NMT models and Section 4.1.2 discusses the results of our models on the external test sets. The comparison between the **genre-aware** and **genre-agnostic** models used as control conditions is discussed in Section 4.2.3, as it is more relevant to the second experiment and research question.

4.1.1 Genre-Specific vs General NMT Models

The **genre-specific** models were expected to outperform the general models. However, the differences between the models are very small, with the **genre-specific** models consistently producing slightly higher quality translations only for *Legal, News*, *News*, and *Promotion*, across all three languages. This suggests that the benefits of fine-tuning on a specific genre are minimal compared to fine-tuning on a dataset that includes a variety of genres.

Generally, we noticed that the **genre-specific** models tend to achieve higher scores than the baseline models across all genres, not only for their target genre, indicating that there is still useful information even in sub-optimal datasets. The only exception (across all languages) seems to be the models fine-tuned for *Legal* texts, which perform worse than the baseline models when tested on other genres (see complete results tables in Appendix E). This is not surprising, since *Legal* texts tend to use very specific terminology and be more formal in style compared to the other genres. The largest difference in style and terminology might be between the *Legal* texts and the *Prose/Lyrical* (see Appendix B). Consequently, the models fine-tuned on *Legal* texts consistently achieve significantly lower scores when tested on *Prose/Lyrical* texts than other genre-specific models (Figure 12, Figure 16 and Figure 20, in Appendix E). We notice larger improvements over the OPUS-MT baseline in the case of Icelandic than in the case of Croatian and Turkish, where differences were smaller than 1 COMET point. This is to be expected, as there is less data available for Icelandic, not only in the MaCoCu dataset but also in the OPUS Corpus, which was used for pre-training. [7]

[7] There are about 33,000,000 sentence pairs for English-Icelandic, but 170,000,000 for English-Turkish and 130,000,000 for English-Croatian (source: https://opus.nlpl.eu/).

	Croatian		Turkish		Icelandic	
	MaCoCu	**Flores devtest**	**MaCoCu**	**Flores devtest**	**MaCoCu**	**Flores devtest**
Information/Explanation						
Genre-Specific	0.62	0.78	0.21	-0.13	2.8	0.11
Genre-Agnostic	**0.65**	0.05	**0.29**	**0.03**	2.6	0.18
Genre-Aware	0.54	**1**	-1.1	-2.3	**3**	**2.9**
Instruction						
Genre-Specific	0.17	0.44	**0.27**	0.63	**6.8**	**3.2**
Genre-Agnostic	**0.34**	**1.2**	0.21	**0.72**	6.5	2.3
Genre-Aware	0.26	1	-1.5	-1.4	5.7	2.7
News						
Genre-Specific	**0.87**	**1.2**	**0.87**	-0.42	**5.3**	1.6
Genre-Agnostic	0.71	-0.82	0.43	**-0.07**	4.4	1.8
Genre-Aware	0.48	1.1	-0.73	-2.6	3.5	**5.5**
Opinion/Argumentation						
Genre-Specific	**0.74**	0.49	–	–	–	–
Genre-Agnostic	0.44	0.33	0.17	**0.22**	4.1	2.2
Genre-Aware	0.72	**0.74**	-1.8	-1.4	**5.1**	**3.4**
Legal						
Genre-Specific	**1.9**	–	**0.92**	–	**4.8**	–
Genre-Agnostic	1.5	–	0.39	–	4.4	–
Genre-Aware	0.72	–	-0.19	–	3.5	–
Promotion						
Genre-Specific	**1.9**	–	**0.45**	–	**7.2**	–
Genre-Agnostic	1.5	–	**0.45**	–	6.4	–
Genre-Aware	1.2	–	-0.39	–	6.1	–

Table 6: Difference in COMET score between the pre-trained OPUS-MT models and the models fine-tuned on subsets of genre-specific MaCoCu data or on an equal-sized randomized dataset, as genre-aware or genre-agnostic models. The largest improvements over the baseline models are written in bold font. Due to low data availability, we cannot reliably evaluate our models on *Legal* and *Promotion* texts from external data sets.

4.1.2 Genre-Specific Models Tested on External Datasets

Furthermore, when testing on external datasets, the **genre-specific** models are sometimes outperformed on their target genre by other **genre-specific** models. For instance, the Icelandic model fine-tuned on *Instruction* texts scores higher than the other genre-specific models across all genres in Flores dev, Flores devtest, and WMT21 News (see Appendix E.2). Moreover, in the case of Turkish, our genre-specific models tend to produce lower quality translations than the OPUS-MT baseline on the WMT18 News test set (Figure 23), on the Flores dev set (Figure 21), and on the *News* and *Information/Explanation* examples from Flores devtest (Figure 18).

When testing the Croatian **genre-specific** models on the *News* examples from external datasets, only the **genre-specific** model for *News* outperforms the OPUS-MT baseline (Figure 13, Figure 14 and Figure 15, from Appendix E.2). Furthermore, when testing on the *News* examples from the MaCoCu test set, the **genre-specific** *News* model still scores highest (+0.87 COMET). In contrast, the textbfgenre-specific *Instruction* and *Promotion* are still outperformed by the baseline, and the models fine-tuned for *Information/Explanation* and *Opinion/Argumentation* improve by less than 0.4 COMET points (Figure 12). Therefore, the **genre-specific** *News* models for Croatian tend to consistently outperform the other genre-specific models by a considerable margin, on every test set. Consequently, this genre seems to be defined more precisely in Croatian than in the other languages, as we do not see this trend in the case of Icelandic or Turkish.

4.2 Genre-Aware vs Genre-Agnostic Models

This section presents the findings of our second experiment, which compares **genre-aware** and **genre-agnostic** NMT models in different scenarios: trained from scratch on the MaCoCu dataset (Section 4.2.1), fine-tuned on the MaCoCu dataset 4.2.2), and trained on document-level on the MaCoCu dataset 4.2.4). Additionally, Section 4.2.3 briefly addresses the comparison of the **genre-aware** and **genre-agnostic** models fine-tuned on a subset of the MaCoCu dataset, as part of the control condition in the first experiment.

4.2.1 Genre-Aware vs Genre-Agnostic Models Trained on MaCoCu

We compare **genre-aware** and **genre-agnostic** models trained from scratch on the MaCoCu datasets for Croatian and Turkish. Figure 2 shows a comparison of the COMET scores achieved by our Croatian models on several test sets. We find that all models achieve higher scores on the MaCoCu test set than on the external test sets. However, the differences between models are rather small (< 0.5 COMET points) on a given dataset, with the exception of the **genre-aware + token** model that uses a custom-trained tokenizer. This model scores 2-3 points lower than the other models, despite our hypothesis that it would outperform them.

Since we did not find any benefits in training new tokenizers, we only used the pre-trained OPUS-MT tokenizers alongside the Turkish models we trained from scratch (Figure 3). We again find very small differences between models (< 0.6 COMET points), indicating that the **genre-agnostic** models are on par with the *genre-aware* ones. Furthermore, our models seem to produce better translations for the external datasets, especially for the Flores sets, than for the MaCoCu test set, despite being trained exclusively on the MaCoCu corpus.

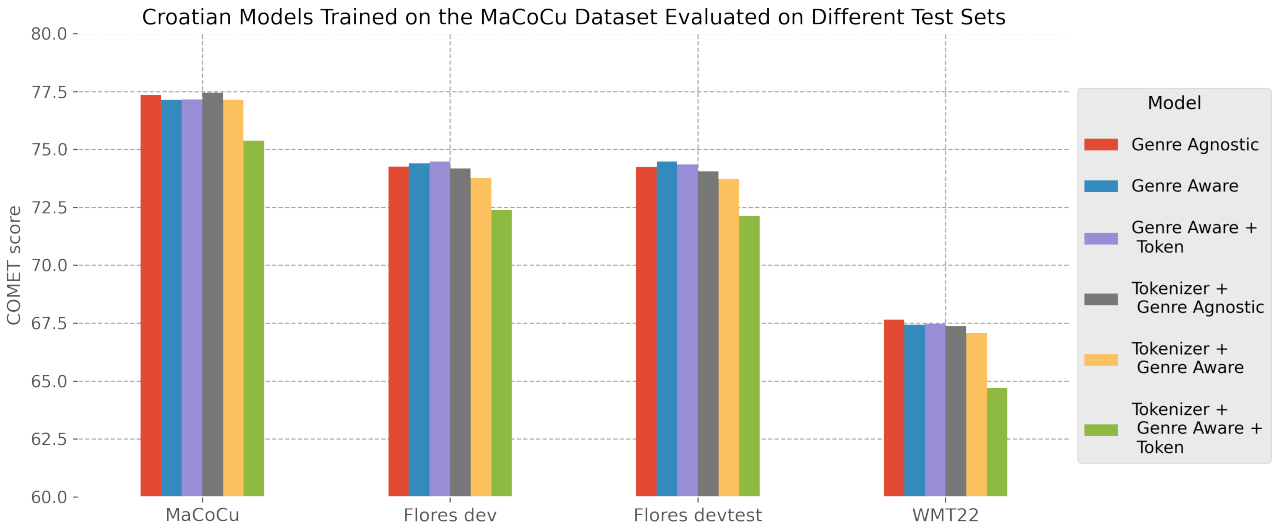

Figure 2: COMET scores of the Croatian models trained from scratch on the MaCoCu dataset. The results are averaged over 3 runs.

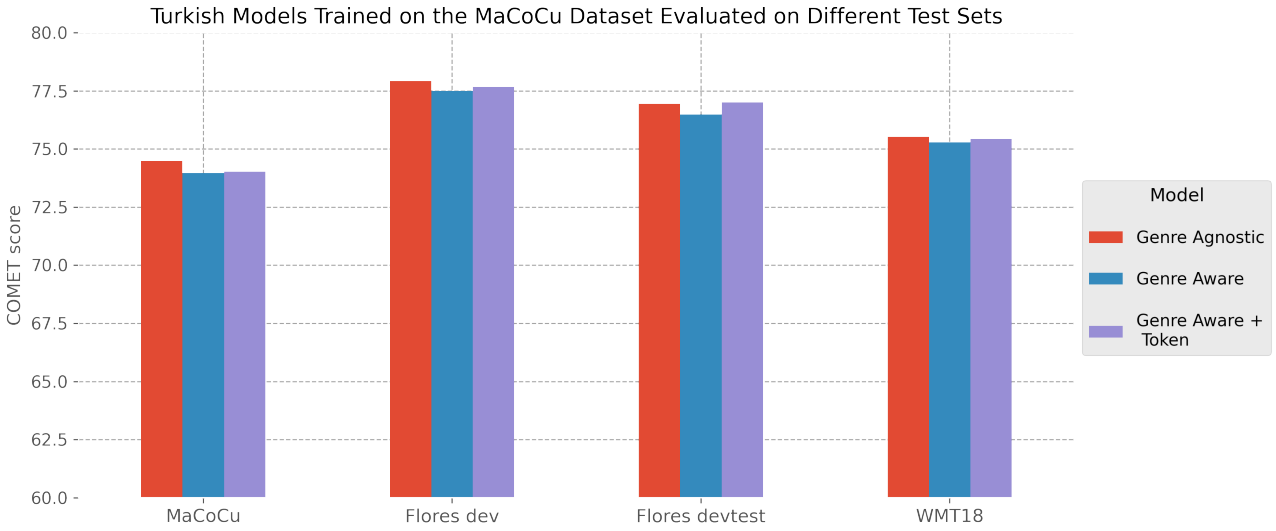

Figure 3: COMET scores of the Turkish models trained from scratch on the MaCoCu dataset. The results are averaged over 3 runs.

4.2.2 Genre-Aware vs Genre-Agnostic Models Fine-Tuned on MaCoCu

Next, we compare models that are fine-tuned as either **genre-aware** or **genre-agnostic**. We used OPUS-MT models as the baseline models, and therefore we plot the results as a difference between the COMET scores of the baseline models and the fine-tuned models, as in the first experiment. Note that the dataset available for Croatian was roughly ten times the size of that available for Icelandic and Turkish (Table 2).

Figure 4 shows the models fine-tuned for Icelandic, tested on both the MaCoCu test set and the external test sets. The differences between models are very small on the external datasets (< 0.2), and on the MaCoCu test set, the **genre-agnostic** model seems to outperform both **genre-aware** models by around 0.4 COMET points.

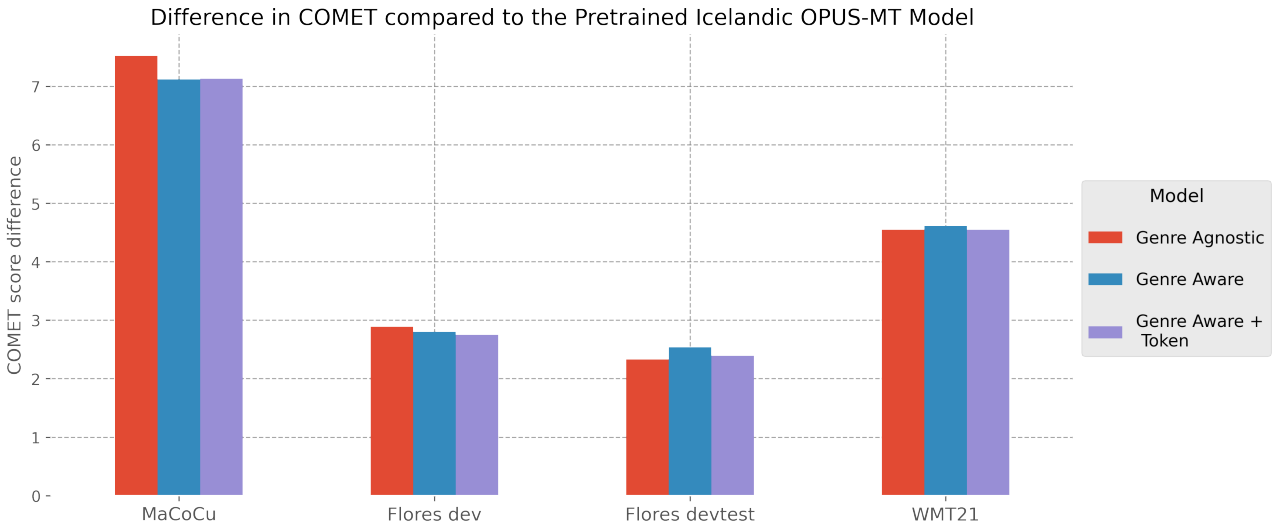

Figure 4: The difference in COMET scores between the baseline OPUS-MT model for Icelandic and the models fine-tuned on the entire MaCoCu corpus. Models are evaluated on several datasets, the results are averaged over 3 runs.

In the case of the Croatian fine-tuned models (Figure 5), the **genre-agnostic** models consistently produced higher quality translations than the **genre-aware** models, by a margin of around 0.5. Therefore, we find no evidence that fine-tuning **genre-aware** models should be preferred over fine-tuning **genre-agnostic** models, regardless of the size of the dataset used for fine-tuning.

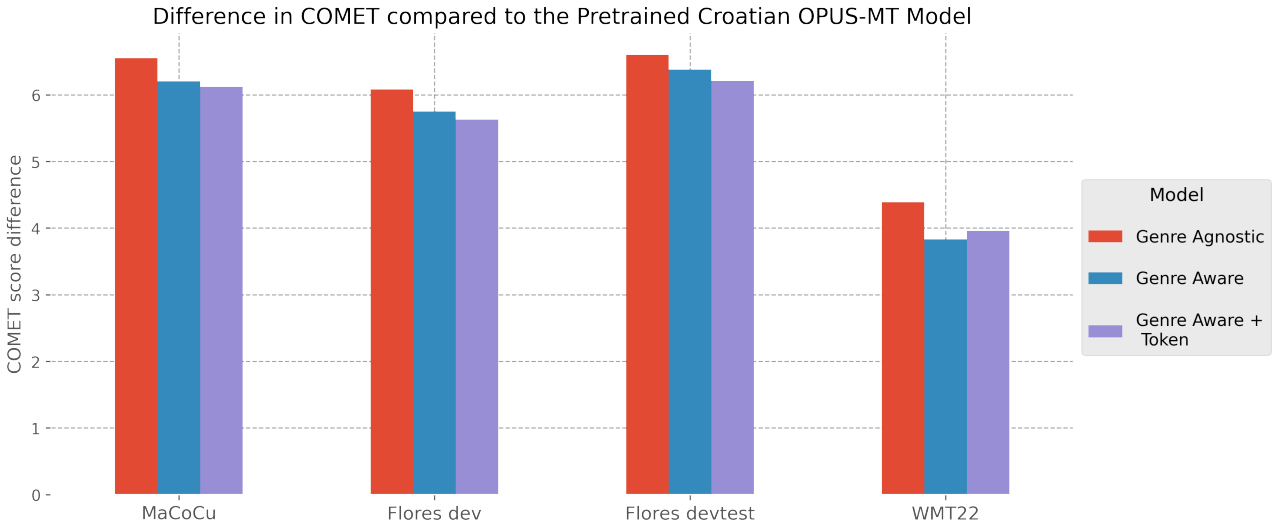

Figure 5: The difference in COMET scores between the baseline OPUS-MT model for Croatian and the models fine-tuned on the entire MaCoCu corpus. Models are evaluated on several datasets, the results are averaged over 3 runs.

Surprisingly, when fine-tuning the OPUS-MT model for Turkish on the MaCoCu data, we find that they perform worse than the baseline model on both Flores sets and on the WMT18 News test set

(Figure 6). Only on the MaCoCu test set, fine-tuning leads to a small improvement of 0.4 COMET points, for the **genre-aware** model. In fact, fine-tuning seems to be more detrimental to the **genre-aware** models, as they do not improve significantly for MaCoCu data (about 0.05), and they perform worse than the **genre-agnostic** ones on the external data. However, overall the impact of fine-tuning is more modest for Turkish than for Croatian and Icelandic, which improved by 2 to 7 COMET points, while Turkish OPUS-MT baselines scores were altered by less than 0.8 COMET in either direction.

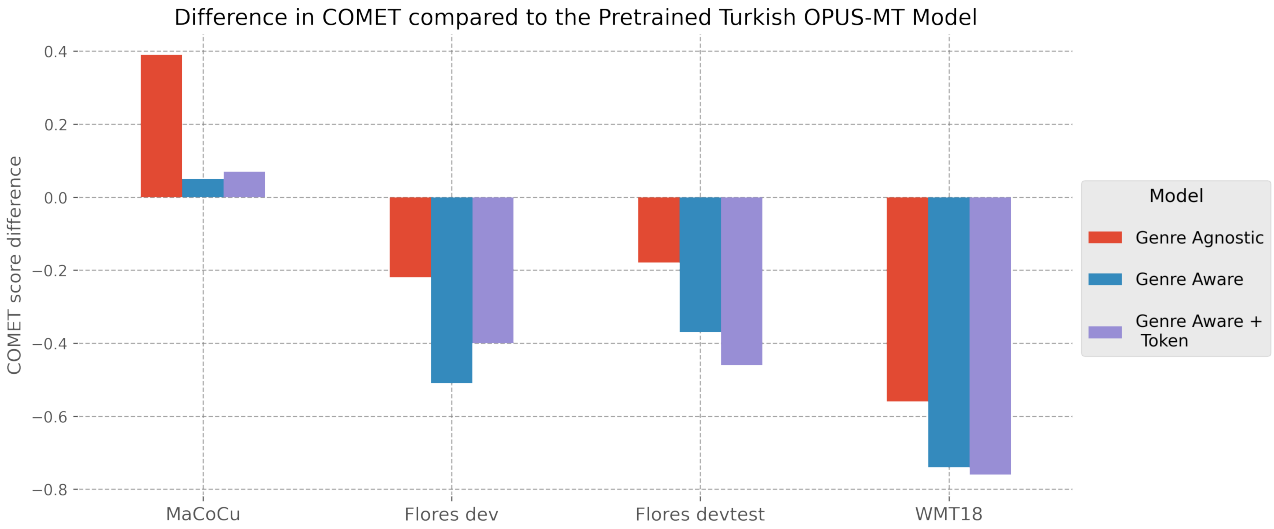

Figure 6: The difference in COMET scores between the baseline OPUS-MT model for Turkish and the models fine-tuned on the entire MaCoCu corpus. Models are evaluated on several datasets, the results are averaged over 3 runs.

4.2.3 Genre-Aware vs Genre-Agnostic Models Fine-Tuned on a Subset of MaCoCu

This experiment compares the two control conditions used in the first experiment - a general **genre-aware** model and a general **genre-agnostic** model, both trained on a subset of MaCoCu data. The examples in the training data set are chosen randomly, the genre distribution is illustrated in Appendix D. The **genre-aware** model is trained by manually adding the special tokens to the vocabulary (see 3.4), thus it is a **genre-aware + tokens** model.

When comparing the general models used in the first experiment (Table 6), it was expected that the **genre-aware** models would outperform the **genre-agnostic** ones. However, when testing only on MaCoCu data, **genre-agnostic** models seem to score higher than the **genre-aware** models. In fact, **genre-aware** models outperform **genre-agnostic ones** only in the case of *Opinion/Argumentation* examples, for Croatian and Icelandic, and this trend occurs when testing on Flores devtest as well. Moreover, in the case of Turkish, not only do the **genre-aware** models consistently perform worse than the **genre-agnostic** ones, but they also score lower than the OPUS-MT baseline, on all test sets (see Appendix E.3).

4.2.4 Genre-Aware vs Genre-Agnostic Models on Document-Level

Finally, we fine-tuned OPUS-MT models as either **genre-aware** or **genre-agnostic**, using the MaCoCu training data set aggregated into documents according to their source URL. We tested these models on the MaCoCu test set and on the Flores dev and devtest sets, which we also aggregated into documents according to their sources. For this experiment, we present BLEU scores, instead of COMET scores, as COMET scores are computed on individual sentences.

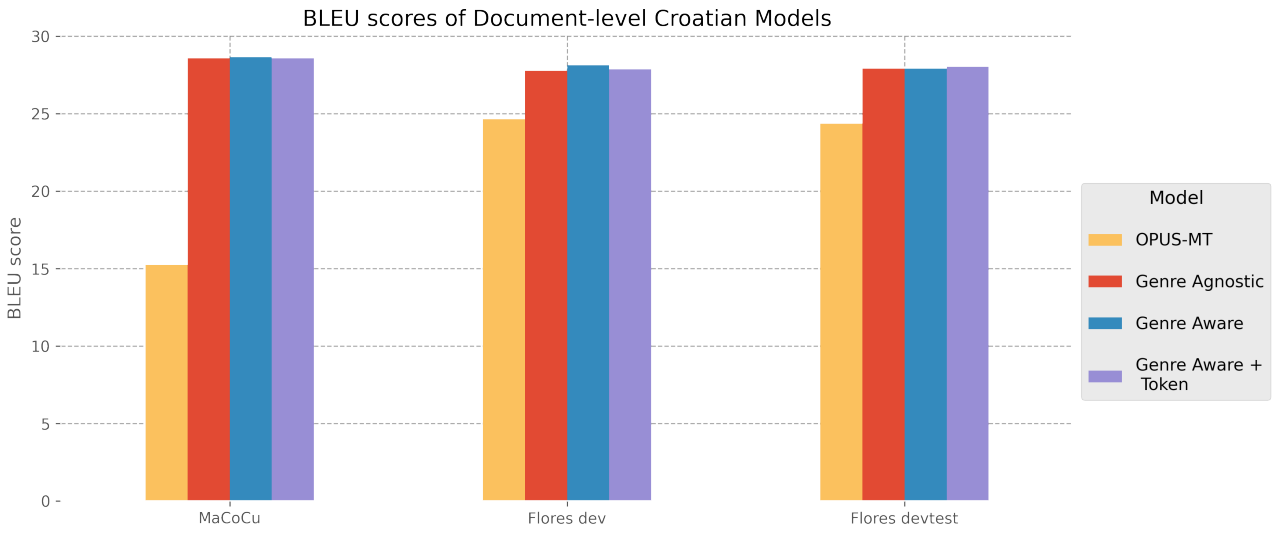

Figure 7: BLEU scores of the baseline OPUS-MT system for Croatian, and the models fine-tuned on document-level MaCoCu data, as either genre-aware and genre-agnostic.

We find almost no difference (< 0.1) between the BLEU scores of the **genre-aware** and **genre-agnostic** models, in any of the language pairs English-Croatian (Figure 7), English-Icelandic (Figure 8) and English-Turkish (Figure 9). Consequently, we also do not find differences between the two types of **genre-aware** models we experimented with.

Furthermore, we noticed that the baseline OPUS-MT models achieve low scores on the MaCoCu test set – 15 for Croatian, and 5 for Icelandic and Turkish. Therefore, after fine-tuning we find the largest improvements when testing on the MaCoCu test set – more than 10 BLEU points – while for the Flores sets we observe improvements of around 2.5 BLEU for Croatian and 3 BLEU for Icelandic (but almost no improvements for Turkish). This could be explained by the fact that the MaCoCu documents are longer on average. For instance, the documents in the Croatian MaCoCu test set are on average 678 words long ($std = 231$) compared to 74.5 words ($std = 26.2$) for Croatian Flores dev set (see Table 7 from Appendix A. Although the median word number per document is more similar between the test sets (119 for MaCoCu and 72 for Flores), the scores we present are also computed as an average and therefore are affected more by the outliers.

As previously mentioned, the OPUS-MT models for Icelandic and Croatian improve on the Flores sets after fine-tuning. However, the benefits of fine-tuning are marginal (0.1 BLEU) for the Turkish OPUS-MT model (Figure 9). Despite this, the baseline Turkish model improves the most (13 BLEU points) on the MaCoCu test set after fine-tuning. This suggests that the OPUS corpus that was used to pre-train the model is more similar to the Flores data sets than to the MaCoCu corpus. This is consis-

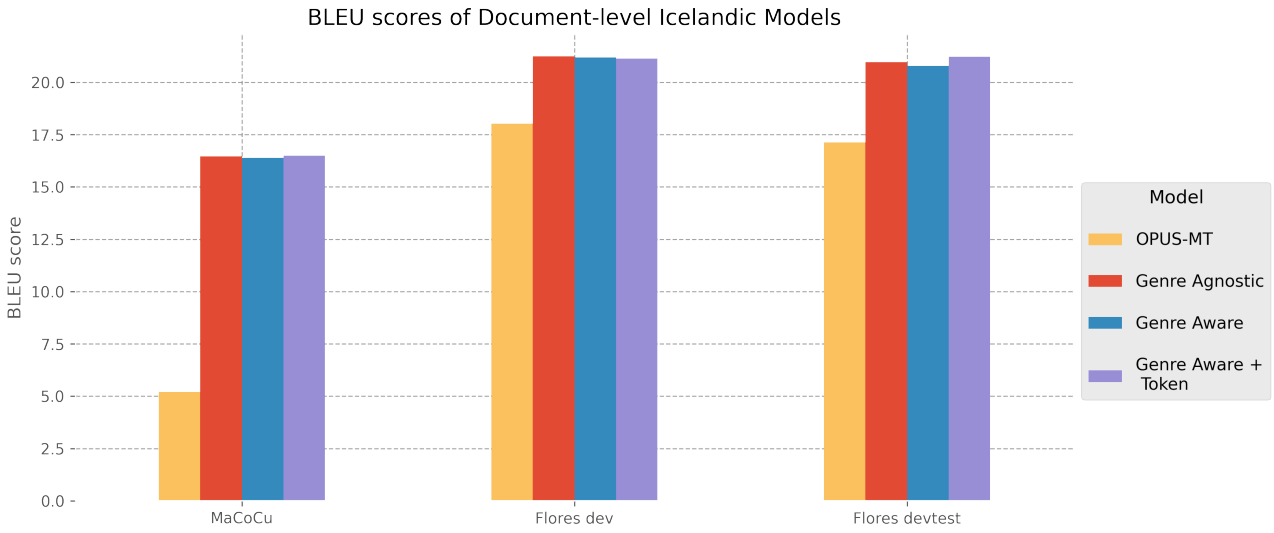

Figure 8: BLEU scores of the baseline OPUS-MT system for Icelandic, and the models fine-tuned on document-level MaCoCu data, as either genre-aware and genre-agnostic.

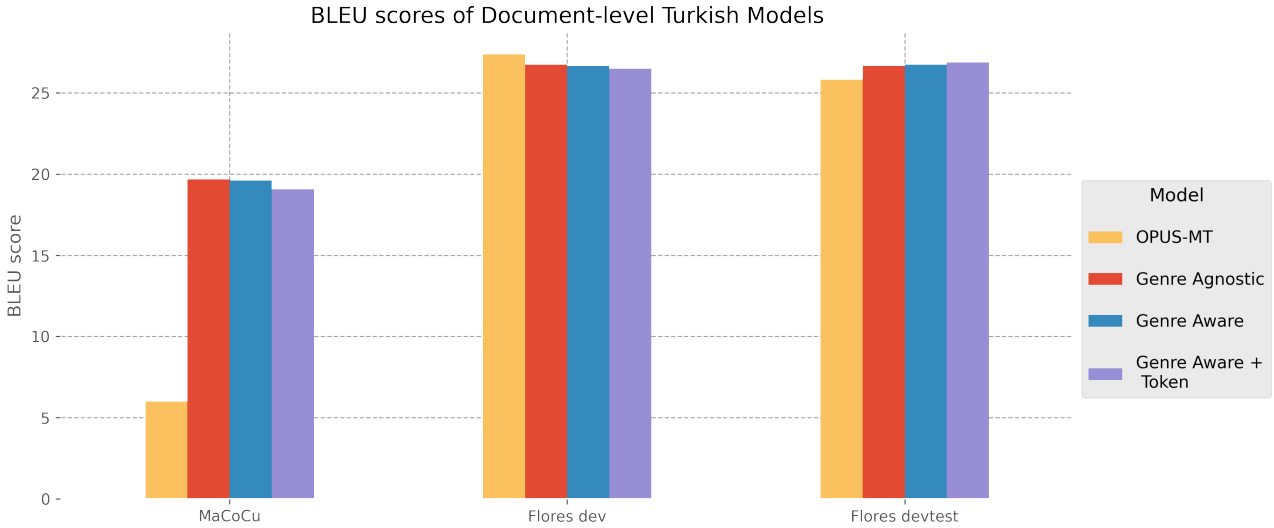

Figure 9: BLEU scores of the baseline OPUS-MT system for Turkish, and the models fine-tuned on document-level MaCoCu data, as either genre-aware and genre-agnostic.

tent with the results of the previous experiment, which found that fine-tuning the Turkish OPUS-MT model on sentence-level MaCoCu data is slightly detrimental to its performance on external datasets.

Lastly, the MaCoCu dataset for Croatian is roughly ten times larger than the one for Icelandic (see Table 2). However, we found that the improvements in BLEU scores after fine-tuning were similar in magnitude for all languages ($10-13$ BLEU on MaCoCu and 3 BLEU on Flores). Therefore, we find that a relatively smaller dataset is sufficient for fine-tuning document-level models, and there might be diminishing returns for training on larger datasets.

5 Discussion

5.1 Genre-Specific Models

The first research question **RQ1** inquired about fine-tuning **genre-specific NMT systems** that were expected to consistently produce better translations for their target genre, compared to other **genre-specific** models and the control condition - a model trained on a dataset with randomly selected instances. We indeed found this to generally be the case when testing our **genre-specific** models on the MaCoCu dataset - which was also used for fine-tuning them. Therefore, we found the genres within the MaCoCu datasets to be rather consistent, and the X-GENRE classifier, used to identify them, to be fairly reliable. Furthermore, we found benefits for fine-tuning on any genre (except for *Legal*). Consequently, models fine-tuned on a randomized dataset, which roughly followed the distribution of genres in the entire corpus, scored similarly to the **genre-specific** models. This suggests that relatively little **genre-specific** data might be sufficient for fine-tuning for a particular genre.

However, when testing on external datasets, the **genre-specific** models do not consistently outperform other **genre-specific** models on their target genre, and neither does the control condition consistently perform well across all genres. Therefore, the MaCoCu might not be consistent with the genres found in the Flores and WMT test sets, despite the classifier labeling them as such. Since all MaCoCu data is web-crawled, this might already be considered a broad, overarching genre. Furthermore, "online" genres tend to differ from their conventional counterparts in certain aspects (Kuzman & Ljubešić, 2023), as they were adapted to a new medium and audience. Therefore, although the classifier finds common aspects between the genres in the MaCoCu and in the external datasets, there might still be fundamental differences between them, which hinder the generalizability of the **genre-specific** NMT models trained on the MaCoCu data.

5.2 Genre-Aware vs. Genre-Agnostic Models

Furthermore, our second research question inquired about comparing **genre-aware** and **genre-agnostic** models in different scenarios: trained from scratch (**RQ2.1**), fine-tune (**RQ2.2**) or fine-tuned on document-level (**RQ2.3**). Across our experiments, we did not find significant differences between the performance of **genre-aware** and **genre-agnostic** models. One possible explanation would be that texts are labeled on document-level, but not all sentences within a document contain genre markers, especially the shorter ones (Kuzman & Ljubešić, 2023). Consequently, when training our models using sentence-level data, some genre tags might be noisy or unreliable, in practice.

However, we would then expect to see more evident differences between the **genre-aware** and **genre-agnostic** models we fine-tuned on document-level. But, we still find that **genre-aware** and **genre-agnostic** models achieve similar scores. This might point to the fact that the attention mechanisms employed by state-of-the-art transformer-based NMT models are sufficiently equipped to identify and conserve the particularities of each genre. Thus, incorporating genre labels in the input of such systems might be redundant, especially in the case of pre-trained models.

Pham et al. (2021) conducted a similar experiment, comparing domain-specific models with domain-aware and domain-agnostic models, their domains being defined according to data provenance. They found that domain-aware models significantly outperformed the domain-agnostic ones only for their

"religious" texts domain. We find a similar trend for our *Opinion/Argumentation* genre, but only for Croatian and Icelandic. Furthermore, they found that the specialized models for each domain performed significantly better than the general models. We find a similar trend only when testing on the MaCoCu test set. However, they do not test their models on external data, and they train their models from scratch, and further fine-tune them on the same data. Therefore, it is unknown whether their models are more robust than ours.

5.3 Limitations

It is possible that incorporating genre labels is redundant since the genres are not distinct enough. As we noted in the case of genre-specific models, in the first experiment, by fine-tuning on any genre - except for *Legal* - the translation quality across all genres tends to improve. Therefore, the genre markers of the majority of genres we included in our experiments might overlap too much or might not be salient enough to justify categorizing our data into genres.

Furthermore, a confounding factor in our experiments might be that genre markers differ between languages (Sharoff, 2020). The labels for all our datasets were generated using only the English side of the parallel corpora. Therefore, the genre markers may be more distinct in English, than in the target languages, and thus **genre-aware models** might be uninformative and redundant. For instance, *Instruction* seems to be a very general genre in Icelandic, such that fine-tuning models on this genre considerably improves their performance across all genres much more than fine-tuning on other genres. On the other hand, *News* seems to be very specific in Croatian, as fine-tuning on any other single genre is detrimental, leading to COMET scores lower than the OPUS-MT baseline, for the external datasets.

5.4 Future Work

A possible direction for future research would be studying the interference between genres as topics, following up on van der Wees et al. (2015), who controlled both the genres and the topics present in their test sets. As mentioned previously, we used data from different Internet domains in our data splits. Therefore, the topics found in the *News* examples from the training data might differ from those found in the *News* examples from the test data. Consequently, the models might fail to encounter relevant vocabulary during training. Of course, this is less likely to happen in the case of more informal genres, which do not require special terminology. but in the case of *Legal* or medical texts different sub-fields might employ distinct terminology. Therefore, it would be interesting to look further into identifying the possible genres for which topics might be more important for defining translation domains. However, since the interference between genres and topics is likely language-specific, and defining very narrow translation domains is tedious the possible benefits of this approach might not outweigh the costs.

5.5 Conclusion

This thesis attempted to answer two main research questions. First, *can automatically generated genre labels be used to fine-tune **genre-specific** neural machine translation systems that outperform general systems on their target genre?* We found that on a holdout dataset, the **genre-specific** models tend to perform marginally better than general systems. However, our findings indicate that the genre labels are not reliable enough for training genre-specific models that perform well on external data. Second,

*can automatically generated genre labels be used to train **genre-aware** neural machine translation systems that outperform equivalent **genre-agnostic** systems?* We did not find benefits to incorporating the genre information into the input of NMT systems, in any of the scenarios we experimented with. Therefore, we conclude that the textual properties that can be automatically derived from texts and categorized into genres are not sufficiently informative to define reliable translation domains that can be utilized across different corpora.

Bibliography

Akhbardeh, F., Arkhangorodsky, A., Biesialska, M., Bojar, O., Chatterjee, R., Chaudhary, V., ... Zampieri, M. (2021, November). Findings of the 2021 conference on machine translation (WMT21). In L. Barrault et al. (Eds.), *Proceedings of the sixth conference on machine translation* (pp. 1–88). Online: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2021.wmt-1.1

Bahdanau, D., Cho, K., & Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Bañón, M., Esplà-Gomis, M., Forcada, M. L., García-Romero, C., Kuzman, T., Ljubešić, N., ... Zaragoza, J. (2022, June). MaCoCu: Massive collection and curation of monolingual and bilingual data: focus on under-resourced languages. In *Proceedings of the 23rd annual conference of the european association for machine translation* (pp. 303–304). Ghent, Belgium: European Association for Machine Translation. Retrieved from https://aclanthology.org/2022.eamt-1.41

Biber, D., & Gray, B. (2016). *Grammatical complexity in academic english: Linguistic change in writing (studies in english language)*. Cambridge: Cambridge University Press.

Bizzoni, Y., Juzek, T. S., España-Bonet, C., Dutta Chowdhury, K., van Genabith, J., & Teich, E. (2020, July). How human is machine translationese? comparing human and machine translations of text and speech. In M. Federico et al. (Eds.), *Proceedings of the 17th international conference on spoken language translation* (pp. 280–290). Online: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2020.iwslt-1.34 doi: 10.18653/v1/2020.iwslt-1.34

Bojar, O., Federmann, C., Fishel, M., Graham, Y., Haddow, B., Huck, M., ... Monz, C. (2018, October). Findings of the 2018 conference on machine translation (WMT18). In O. Bojar et al. (Eds.), *Proceedings of the third conference on machine translation: Shared task papers* (pp. 272–303). Belgium, Brussels: Association for Computational Linguistics. Retrieved from https://aclanthology.org/W18-6401 doi: 10.18653/v1/W18-6401

Britz, D., Le, Q., & Pryzant, R. (2017, September). Effective domain mixing for neural machine translation. In *Proceedings of the second conference on machine translation* (pp. 118–126). Copenhagen, Denmark: Association for Computational Linguistics. Retrieved from https://aclanthology.org/W17-4712 doi: 10.18653/v1/W17-4712

Brown, P. F., Cocke, J., Della Pietra, S. A., Della Pietra, V. J., Jelinek, F., Lafferty, J. D., ... Roossin, P. S. (1990). A statistical approach to machine translation. *Computational Linguistics*, *16*(2), 79–85. Retrieved from https://aclanthology.org/J90-2002

Chu, C., Dabre, R., & Kurohashi, S. (2017, July). An empirical comparison of domain adaptation methods for neural machine translation. In *Proceedings of the 55th annual meeting of the association for computational linguistics (volume 2: Short papers)* (pp. 385–391). Vancouver, Canada: Association for Computational Linguistics. Retrieved from https://aclanthology.org/P17-2061 doi: 10.18653/v1/P17-2061

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., ... Stoyanov, V. (2019). Unsupervised cross-lingual representation learning at scale. *CoRR*, *abs/1911.02116*. Retrieved from http://arxiv.org/abs/1911.02116

Costa-Jussà, M. R., Cross, J., Çelebi, O., Elbayad, M., Heafield, K., Heffernan, K., ... Wang, J. (2022). *No language left behind: Scaling human-centered machine translation.*

Egbert, J., Biber, D., & Davies, M. (2015). Developing a bottom-up, user-based method of web register classification. *Journal of the Association for Information Science and Technology*, *66*(9), 1817-1831. Retrieved from https://asistdl.onlinelibrary.wiley.com/doi/abs/10.1002/asi.23308 doi: https://doi.org/10.1002/asi.23308

Emelin, D., Titov, I., & Sennrich, R. (2020). Detecting word sense disambiguation biases in machine translation for model-agnostic adversarial attacks. *CoRR*, *abs/2011.01846*. Retrieved from https://arxiv.org/abs/2011.01846

Giesbrecht, E., & Evert, S. (2009, 01). Is part-of-speech tagging a solved task? an evaluation of pos taggers for the german web as corpus. *Web as Corpus Workshop (WAC5)*.

Hovy, D., Bianchi, F., & Fornaciari, T. (2020, July). "you sound just like your father" commercial machine translation systems include stylistic biases. In *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 1686–1690). Online: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2020.acl-main.154 doi: 10.18653/v1/2020.acl-main.154

Junczys-Dowmunt, M., Grundkiewicz, R., Dwojak, T., Hoang, H., Heafield, K., Neckermann, T., ... Birch, A. (2018, July). Marian: Fast neural machine translation in C++. In *Proceedings of acl 2018, system demonstrations* (pp. 116–121). Melbourne, Australia: Association for Computational Linguistics. Retrieved from http://www.aclweb.org/anthology/P18-4020

Kobus, C., Crego, J., & Senellart, J. (2017, September). Domain control for neural machine translation. In *Proceedings of the international conference recent advances in natural language processing, RANLP 2017* (pp. 372–378). Varna, Bulgaria: INCOMA Ltd. Retrieved from https://doi.org/10.26615/978-954-452-049-6$_0$49 doi: 10.26615/978-954-452-049-6$_0$49

Kocmi, T., Bawden, R., Bojar, O., Dvorkovich, A., Federmann, C., Fishel, M., ... Popović, M. (2022, December). Findings of the 2022 conference on machine translation (WMT22). In P. Koehn et al. (Eds.), *Proceedings of the seventh conference on machine translation (wmt)* (pp. 1–45). Abu Dhabi, United Arab Emirates (Hybrid): Association for Computational Linguistics. Retrieved from https://aclanthology.org/2022.wmt-1.1

Koehn, P., & Knowles, R. (2017, August). Six challenges for neural machine translation. In *Proceedings of the first workshop on neural machine translation* (pp. 28–39). Vancouver: Association for Computational Linguistics. Retrieved from https://aclanthology.org/W17-3204 doi: 10.18653/v1/W17-3204

Kudo, T., & Richardson, J. (2018). *Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing.*

Kuzman, T. (2022). *Comparison of genre datasets: CORE, GINCO and FTD.* https://github.com/TajaKuzman/Genre-Datasets-Comparison. GitHub.

Kuzman, T., Brglez, M., Rupnik, P., & Ljubešić, N. (2021). *Slovene web genre identification corpus GINCO 1.0.* Retrieved from http://hdl.handle.net/11356/1467 (Slovenian language resource repository CLARIN.SI)

Kuzman, T., & Ljubešić, N. (2023, Nov 16). Automatic genre identification: a survey. *Language Resources and Evaluation*. Retrieved from https://doi.org/10.1007/s10579-023-09695-8 doi: 10.1007/s10579-023-09695-8

Kuzman, T., Rupnik, P., & Ljubešić, N. (2022, June). The GINCO training dataset for web genre identification of documents out in the wild. In *Proceedings of the thirteenth language resources and evaluation conference* (pp. 1584–1594). Marseille, France: European Language Resources Association. Retrieved from https://aclanthology.org/2022.lrec-1.170

Luong, M.-T., & Manning, C. (2015, December 3-4). Stanford neural machine translation systems for spoken language domains. In M. Federico, S. Stüker, & J. Niehues (Eds.), *Proceedings of the 12th international workshop on spoken language translation: Evaluation campaign* (pp. 76–79). Da Nang, Vietnam. Retrieved from https://aclanthology.org/2015.iwslt-evaluation.11

Mehler, A., Sharoff, S., & Santini, M. (2010). Riding the rough waves of genre on the web. *Genres on the Web. Computational Models and Empirical Studies*, 3–30.

Mino, H., Tanaka, H., Ito, H., Goto, I., Yamada, I., & Tokunaga, T. (2020, May). Content-equivalent translated parallel news corpus and extension of domain adaptation for NMT. In *Proceedings of the twelfth language resources and evaluation conference* (pp. 3616–3622). Marseille, France: European Language Resources Association. Retrieved from https://aclanthology.org/2020.lrec-1.445

Müller-Eberstein, M., van der Goot, R., & Plank, B. (2021, November). Genre as weak supervision for cross-lingual dependency parsing. In M.-F. Moens, X. Huang, L. Specia, & S. W.-t. Yih (Eds.), *Proceedings of the 2021 conference on empirical methods in natural language processing* (pp. 4786–4802). Online and Punta Cana, Dominican Republic: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2021.emnlp-main.393 doi: 10.18653/v1/2021.emnlp-main.393

Nagao, M. (1984). A framework of a mechanical translation between japanese and english by analogy principle.. Retrieved from https://api.semanticscholar.org/CorpusID:18366233

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002, July). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the association for computational linguistics* (pp. 311–318). Philadelphia, Pennsylvania, USA: Association for Computational Linguistics. Retrieved from https://aclanthology.org/P02-1040 doi: 10.3115/1073083.1073135

Pham, M., Crego, J. M., & Yvon, F. (2021). Revisiting multi-domain machine translation. *Transactions of the Association for Computational Linguistics*, *9*, 17–35. Retrieved from https://aclanthology.org/2021.tacl-1.2 doi: 10.1162/tacl$_{a0}$0351

Post, M. (2018, October). A call for clarity in reporting BLEU scores. In *Proceedings of the third conference on machine translation: Research papers* (pp. 186–191). Belgium, Brussels: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/W18-6319

Rei, R., Stewart, C., Farinha, A. C., & Lavie, A. (2020, November). COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 conference on empirical methods in natural language processing (emnlp)* (pp. 2685–2702). Online: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2020.emnlp-main.213 doi: 10.18653/v1/2020.emnlp-main.213

Saunders, D. (2021). Domain adaptation and multi-domain adaptation for neural machine translation: A survey. *CoRR*, *abs/2104.06951*. Retrieved from https://arxiv.org/abs/2104.06951

Sennrich, R., Haddow, B., & Birch, A. (2016, August). Neural machine translation of rare words with subword units. In K. Erk & N. A. Smith (Eds.), *Proceedings of the 54th annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 1715–1725). Berlin, Germany: Association for Computational Linguistics. Retrieved from https://aclanthology.org/P16-1162 doi: 10.18653/v1/P16-1162

Sharoff, S. (2018). Functional text dimensions for the annotation of web corpora. *Corpora*, *13*(1), 65-95. Retrieved from https://doi.org/10.3366/cor.2018.0136 doi: 10.3366/cor.2018.0136

Sharoff, S. (2020, 12). Genre annotation for the web: text-external and text-internal perspectives. *Register Studies*, *3*. doi: 10.1075/rs.19015.sha

Stergiadis, E., Kumar, S., Kovalev, F., & Levin, P. (2021). Multi-domain adaptation in neural machine translation through multidimensional tagging. *CoRR*, *abs/2102.10160*. Retrieved from https://arxiv.org/abs/2102.10160

Stewart, J. G., & Callan, J. (2009). *Genre oriented summarization* (Unpublished doctoral dissertation). Carnegie Mellon University, Language Technologies Institute, School of ....

Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to sequence learning with neural networks. *Advances in neural information processing systems*, *27*.

Tars, S., & Fishel, M. (2018). Multi-domain neural machine translation. *CoRR*, *abs/1805.02282*. Retrieved from http://arxiv.org/abs/1805.02282

Tiedemann, J., & Thottingal, S. (2020). OPUS-MT — Building open translation services for the World. In *Proceedings of the 22nd annual conferenec of the european association for machine translation (eamt).* Lisbon, Portugal.

van der Linde, K. . S. L. P. . Z. J. . B. M., Jelmer Heafield. (2023). *Bitextor.* https://github.com/bitextor/bitextor. GitHub.

van der Wees, M., Bisazza, A., & Monz, C. (2018). Evaluation of machine translation performance across multiple genres and languages. In *International conference on language resources and evaluation.*

van der Wees, M., Bisazza, A., Weerkamp, W., & Monz, C. (2015, July). What's in a domain? analyzing genre and topic differences in statistical machine translation. In *Proceedings of the 53rd annual meeting of the association for computational linguistics and the 7th international joint conference on natural language processing (volume 2: Short papers)* (pp. 560–566). Beijing, China: Association for Computational Linguistics. Retrieved from https://aclanthology.org/P15-2092 doi: 10.3115/v1/P15-2092

van Noord, R. (2023). *Translation direction.* https://github.com/RikVN/TranslationDirection. GitHub.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... Polosukhin, I. (2017). Attention is all you need. *CoRR*, *abs/1706.03762*. Retrieved from http://arxiv.org/abs/1706.03762

Vu, T., & Moschitti, A. (2021). Machine translation customization via automatic training data selection from the web. In *European conference on information retrieval.* Retrieved from https://api.semanticscholar.org/CorpusID:231986294

Wang, H., Wu, H., He, Z., Huang, L., & Church, K. W. (2022). Progress in machine translation. *Engineering*, *18*, 143-153. Retrieved from https://www.sciencedirect.com/science/article/pii/S2095809921002745 doi: https://doi.org/10.1016/j.eng.2021.03.023

Zaragoza-Bernabeu, J., Ramírez-Sánchez, G., Bañón, M., & Ortiz Rojas, S. (2022, June). Bicleaner AI: Bicleaner goes neural. In *Proceedings of the thirteenth language resources and evaluation conference* (pp. 824–831). Marseille, France: European Language Resources Association. Retrieved from https://aclanthology.org/2022.lrec-1.87

Zhang, M., & Toral, A. (2019, August). The effect of translationese in machine translation test sets. In O. Bojar et al. (Eds.), *Proceedings of the fourth conference on machine translation (volume 1: Research papers)* (pp. 73–81). Florence, Italy: Association for Computational Linguistics. Retrieved from https://aclanthology.org/W19-5208 doi: 10.18653/v1/W19-5208

Appendices

A Document-Level Genre Distribution

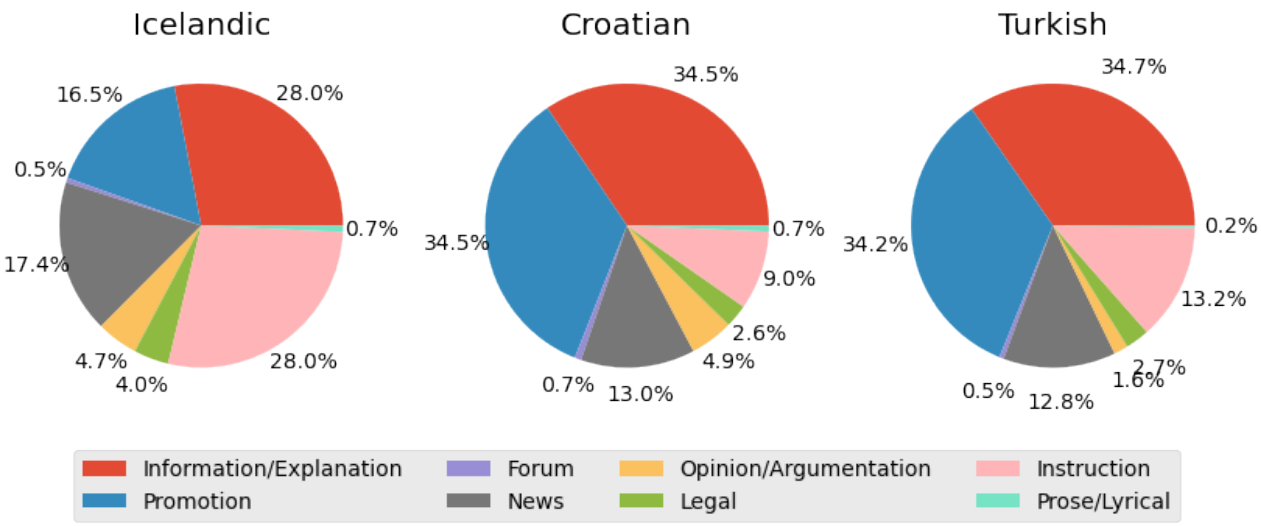

Figure 10: Genre distribution in the (entire) MaCoCu corpora. The sentence-level data in each data split is aggregated according to the source URLs, as explained in Section 3.1.

	Flores dev	**Flores devtest**	**Croatian MaCoCu**	**Icelandic MaCoCu**	**Turkish MaCoCu**
Forum	1	1	48	4	200
Information/Explanation	121	126	368	153	530
Instruction	44	36	209	76	346
Legal	0	0	59	15	69
News	96	87	257	106	699
Opinion/Argumentation	15	26	161	29	168
Other	1	1	0	0	0
Promotion	4	3	420	81	624
Prose/Lyrical	0	0	65	10	107
Total Documents	281	281	1,587	1,339	2,645
Average Number of Words	74.5	77.9	231	167	170
Standard Deviation	26.2	27	678.1	363.2	322.3
Median Number of Words	72	74	119	76	80

Table 7: Document-level genre distribution in our test sets. The average, the median, and the standard deviation of the number of words per document are computed on the Croatian, Icelandic, and Turkish sides for the MaCoCu test sets and on the source side (English) for the Flores test sets.

B Genre Labels Schema

Joint mapping category	Initial categories
Information/Explanation	A14 (academic) (FTD), A16 (information) (FTD), Encyclopedia Article (CORE), Research Article (CORE), Information/Explanation (GINCO), Research Article (GINCO)
Instruction	A7 (instruction) (FTD), How-to (CORE), Recipe (CORE), Technical Support (CORE), Instruction (GINCO), Recipe (GINCO)
Legal	A9 (legal) (FTD), Legal terms (CORE), Legal/Regulation (GINCO)
News	A8 (news) (FTD), News Report/Blog (CORE), Sports Report (CORE), News/Reporting (GINCO), Opinionated News (GINCO), Announcement (GINCO)
Opinion/Argumentation	A11 (personal) (FTD), Editorial (CORE), Formal Speech (CORE), Letter to Editor (CORE), Opinion Blog (CORE), Personal Blog (CORE), Persuasive Article or Essay (CORE), Reviews (CORE), Opinion/Argumentation (GINCO), Review (GINCO)
Promotion	A12 (promotion) (FTD), Advertisement (CORE), Invitation (GINCO), Promotion (GINCO), Promotion of a Product (GINCO), Promotion of Services (GINCO)
Forum	Discussion Forum (CORE), Forum (GINCO)
Prose/Lyrical	A4 (fiction) (FTD), Prayer (CORE), Short Story (CORE), Song Lyrics (CORE), Prose (GINCO), Lyrical (GINCO)
Other	Interview (CORE), TV/Movie Script (CORE), Call (GINCO), Correspondence (GINCO), Interview (GINCO), Other (GINCO), Script/Drama (GINCO)
discarded	A1 (argumentative) (FTD), A17 (review) (FTD), Advice (CORE), Course Materials (CORE), Description of a Person (CORE), Description of a Thing (CORE), Description with Intent to Sell (CORE), FAQ about How-to (CORE), FAQ about Information (CORE), Historical Article (CORE), Information Blog (CORE), Magazine Article (CORE), Other Forum (CORE), Other Information (CORE), Other Informational Persuasion (CORE), Other Opinion (CORE), Other Spoken (CORE), Poem (CORE), Question/Answer Forum (CORE), Reader/Viewer Responses (CORE), Religious Blogs/Sermons (CORE), Technical Report (CORE), Transcript of Video/Audio (CORE), Travel Blog (CORE), Other Narrative (CORE), Other Lyrical (CORE), Other How-to (CORE), FAQ (GINCO), List of Summaries/Excerpts (GINCO)

Figure 11: Mapping between the genre labels used by X-GENRE classifier and English Core, English FTD and Slovene GINCO datasets. Image source: `https://github.com/TajaKuzman/Genre-Datasets-Comparison/blob/main/Creation-of-classifiers-and-cross-prediction/figures/GINCORE-schema-plus-FTD.png`

C Special Genre Tokens

Forum	Information/ Explanation	Instruction	Legal	News	Opinion/ Argumentation	Promotion	Prose/ Lyrical
$< forum >$	$< info >$	$< instr >$	$< law >$	$< news >$	$< arg >$	$< promo >$	$< lit >$

Table 8: Special tokens used for training the **genre-aware** systems.

D Genre Distribution in Randomized Train Sets

	Croatian	Turkish	Icelandic
Forum	803	395	50
Information/Explanation	28182	19139	3577
Instruction	13398	14917	2291
Legal	6496	4685	1133
News	12448	8531	2732
Opinion/Argumentation	6649	1408	996
Promotion	20713	22015	1480
Prose/Lyrical	595	132	526
Total	89284	71222	12785

Table 9: Genre distribution in the randomized datasets used for fine-tuning the control condition of the **genre-specific models** experiment.

E Additional Results of the Genre-Specific Models

E.1 Croatian

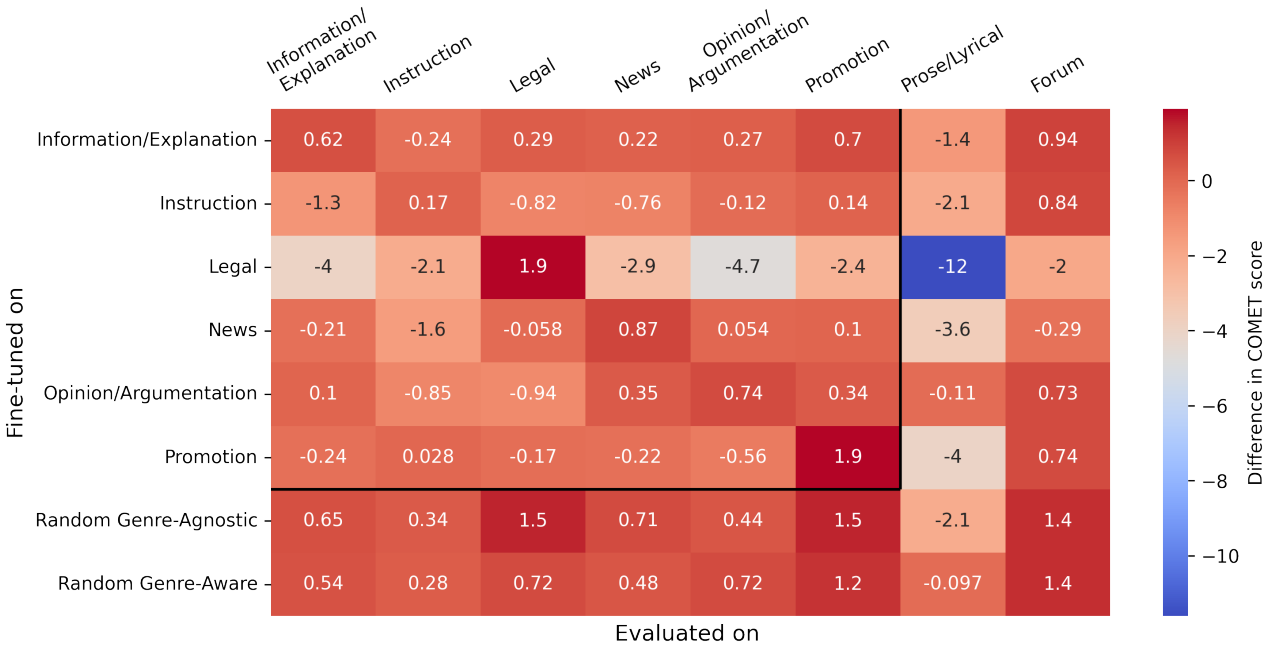

Figure 12: Difference in COMET score between Genre-specific models and the OPUS-MT baseline on the MaCoCu test set.

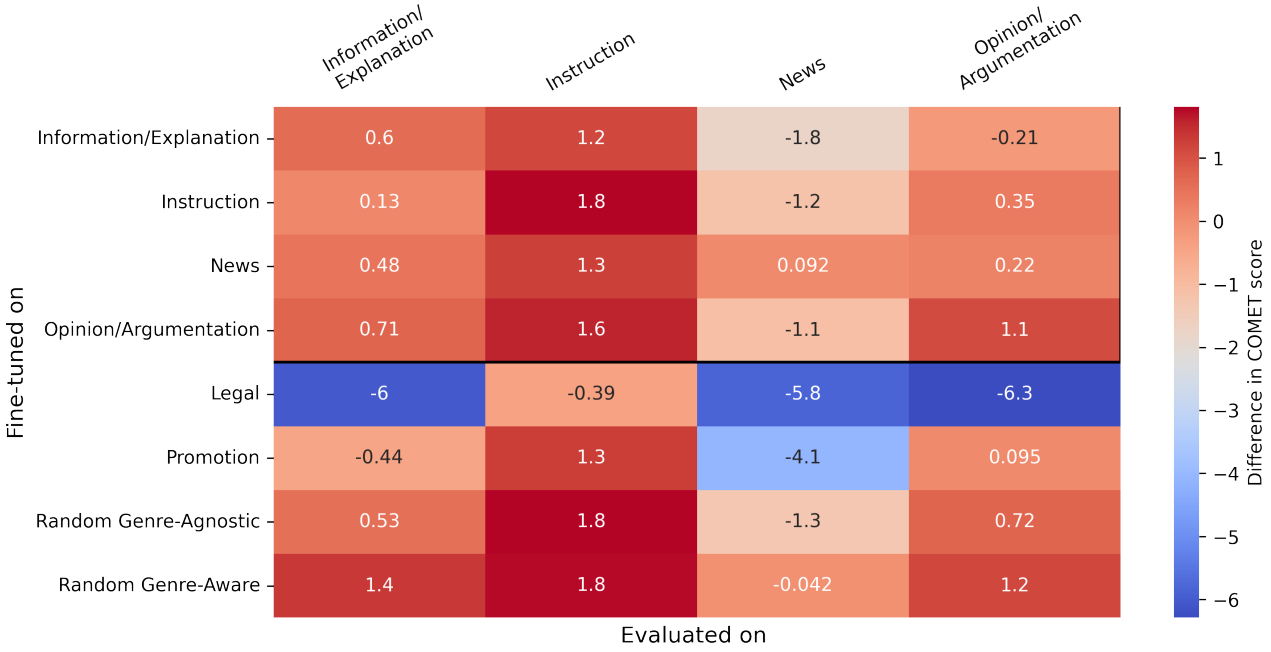

Figure 13: Difference in COMET score between Genre-specific models and the OPUS-MT baseline on Flores dev.

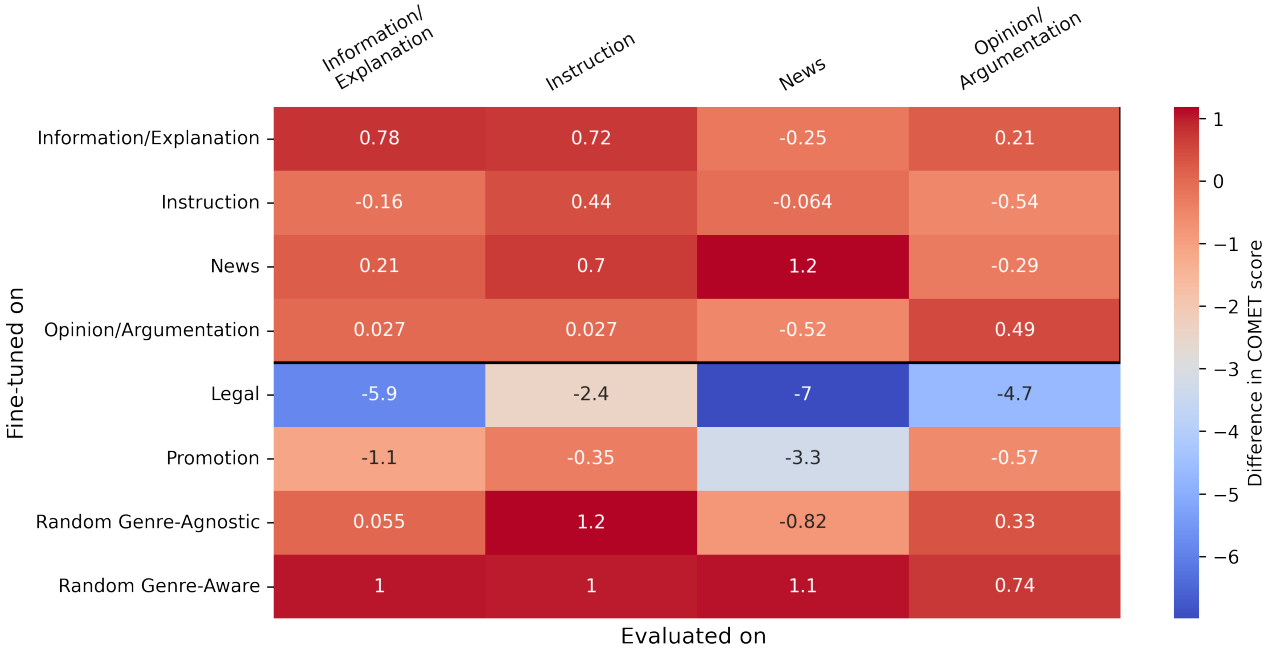

Figure 14: Difference in COMET score between Genre-specific models and the OPUS-MT baseline on Flores devtest.

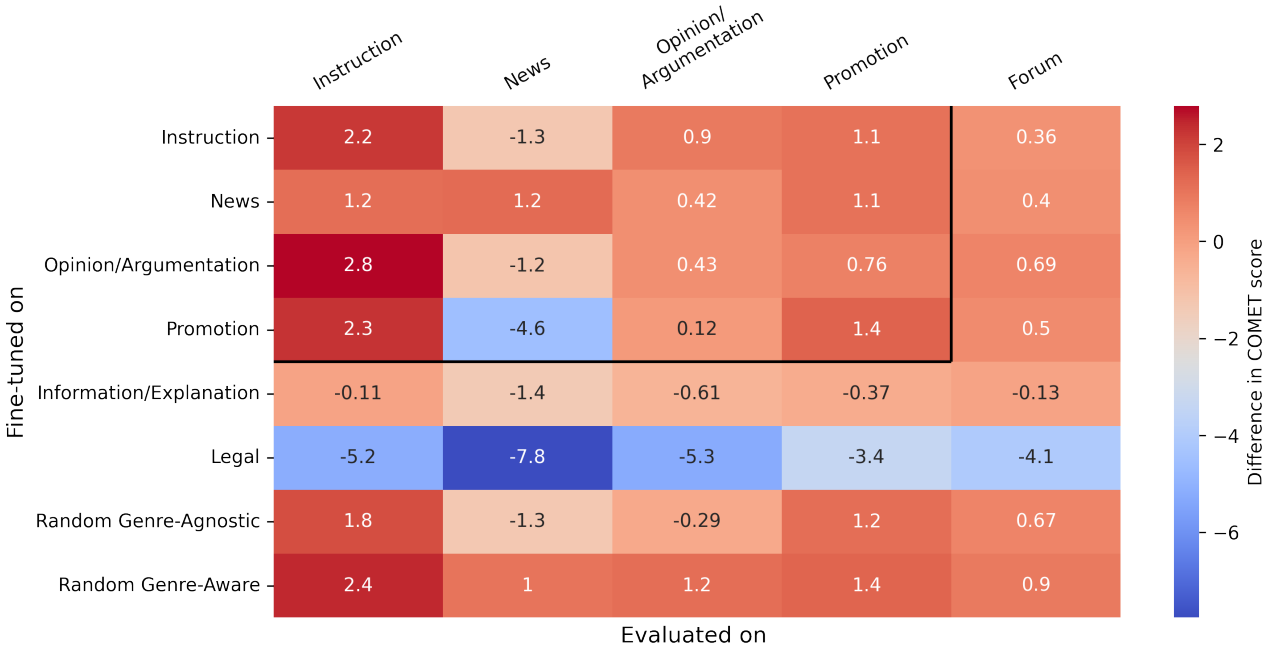

Figure 15: Difference in COMET score between Genre-specific models and the OPUS-MT baseline on the WMT22 test set.

E.2 Icelandic

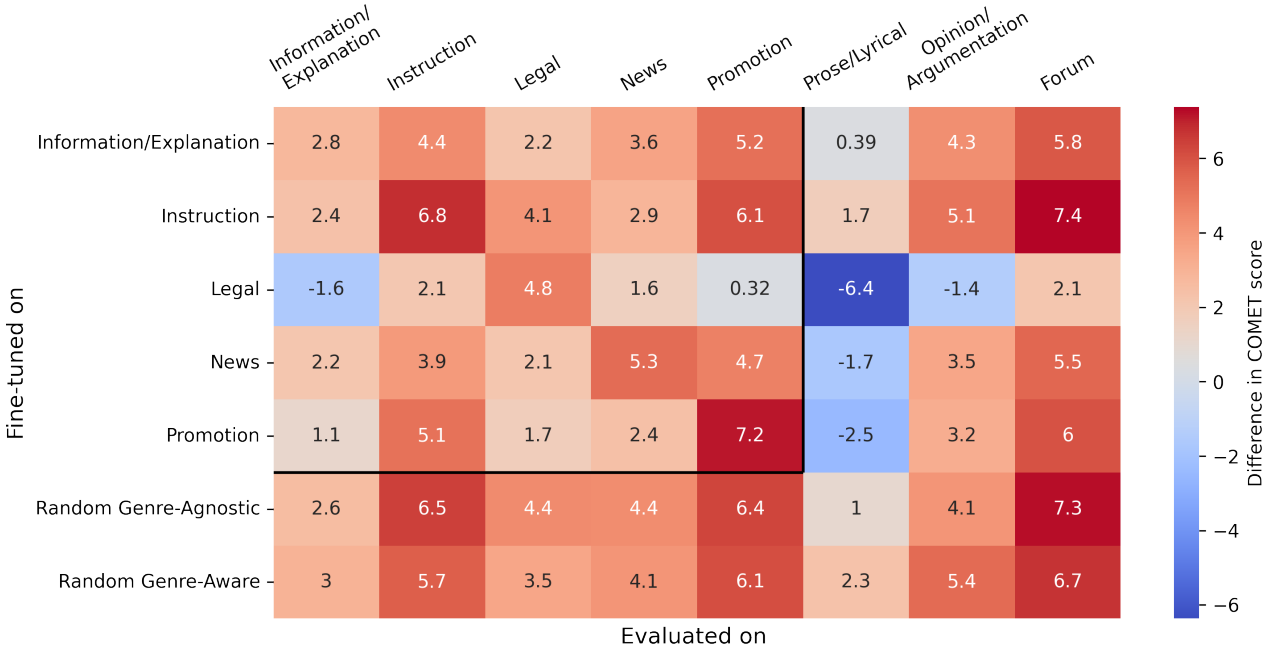

Figure 16: Difference in COMET score between Genre-specific models and the OPUS-MT baseline on the MaCoCu test set.

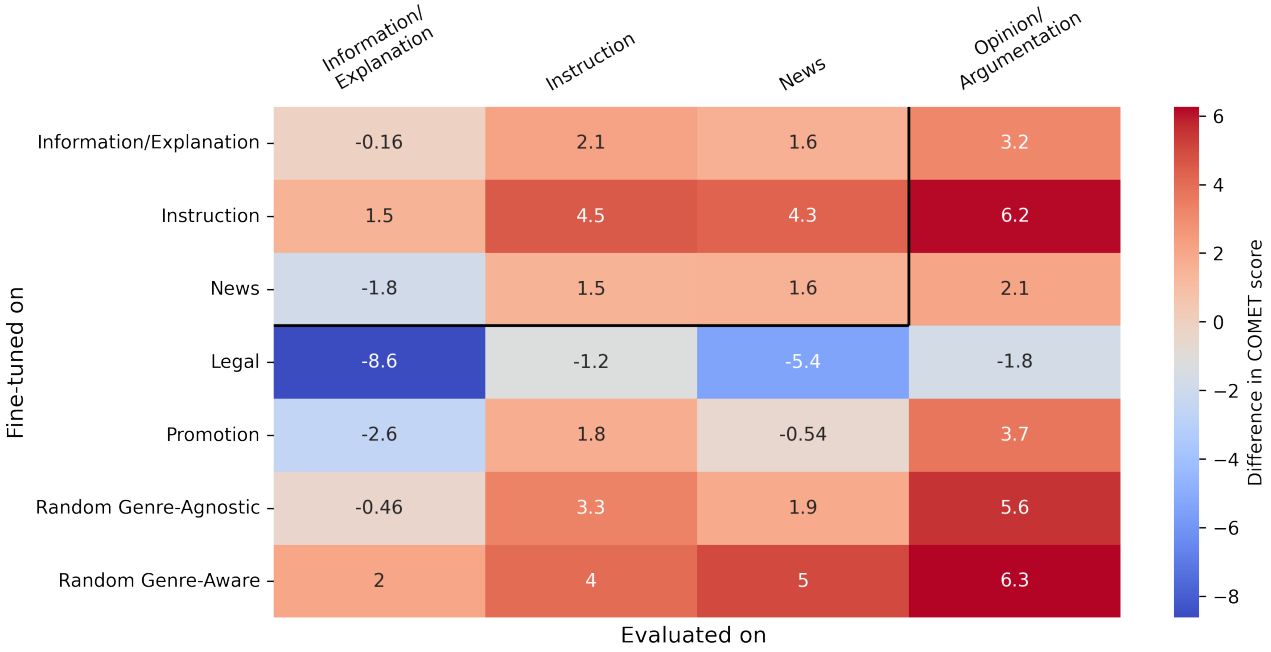

Figure 17: Difference in COMET score between Genre-specific models and the OPUS-MT baseline on the Flores dev set.

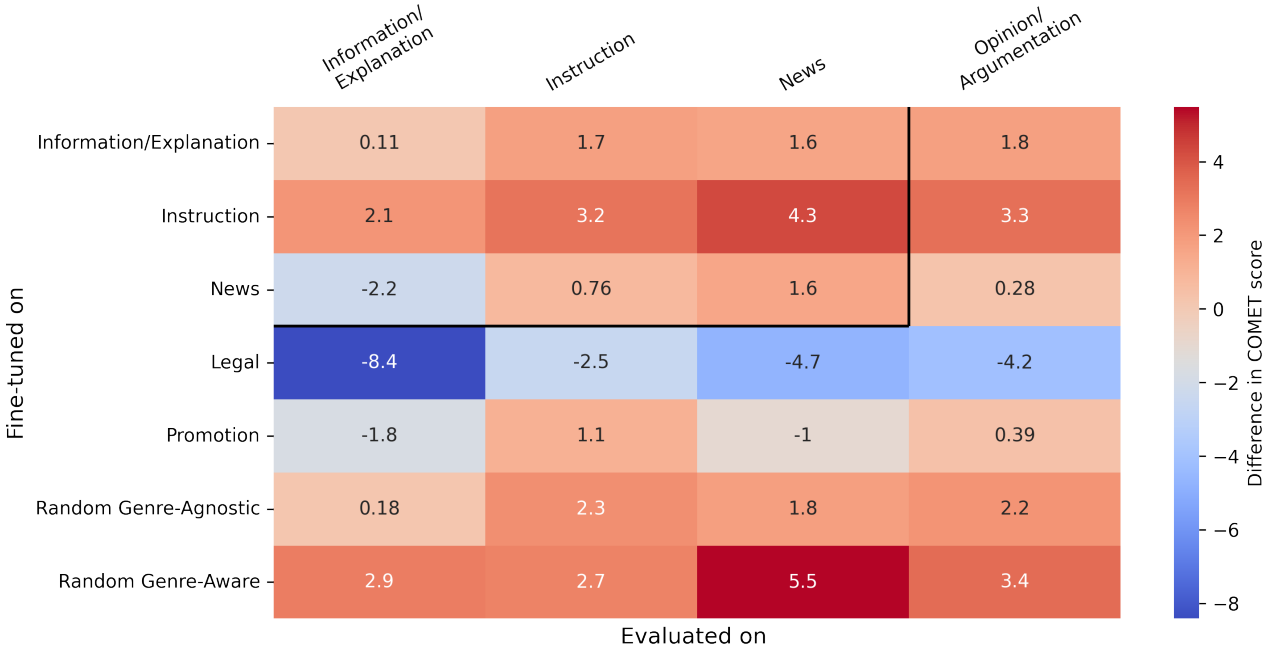

Figure 18: Difference in COMET score between Genre-specific models and the OPUS-MT baseline on the Flores devtest set.

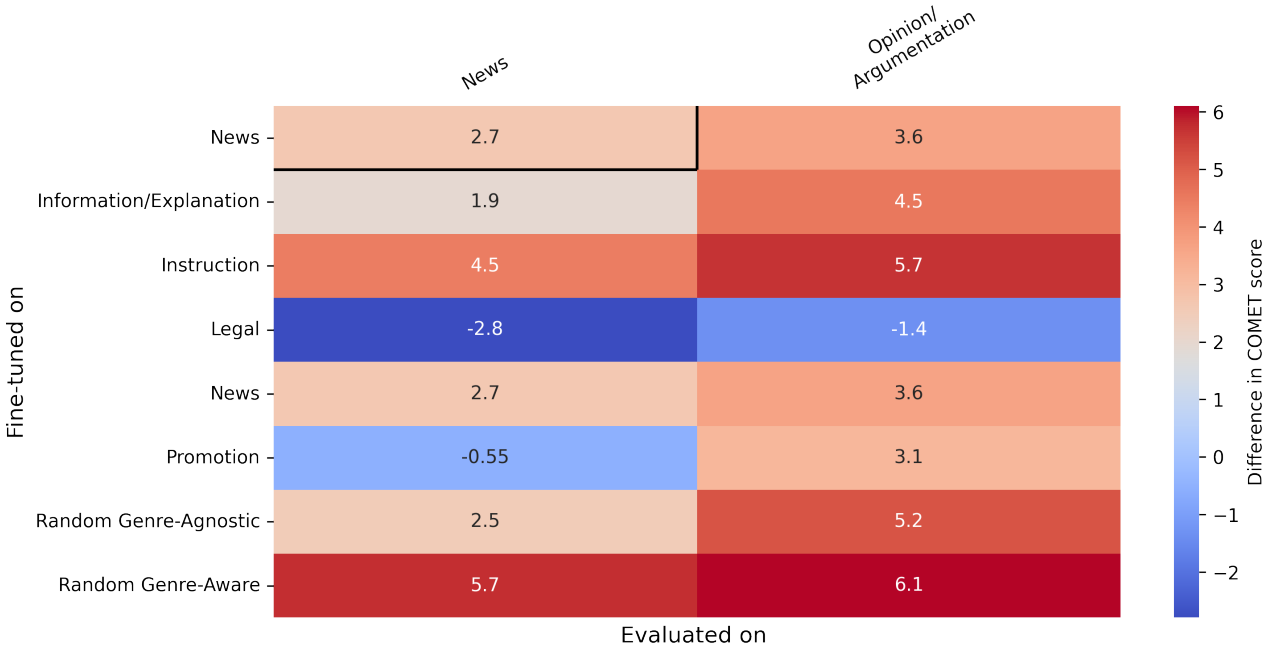

Figure 19: Difference in COMET score between Genre-specific models and the OPUS-MT baseline on the WMT21 test set.

E.3 Turkish

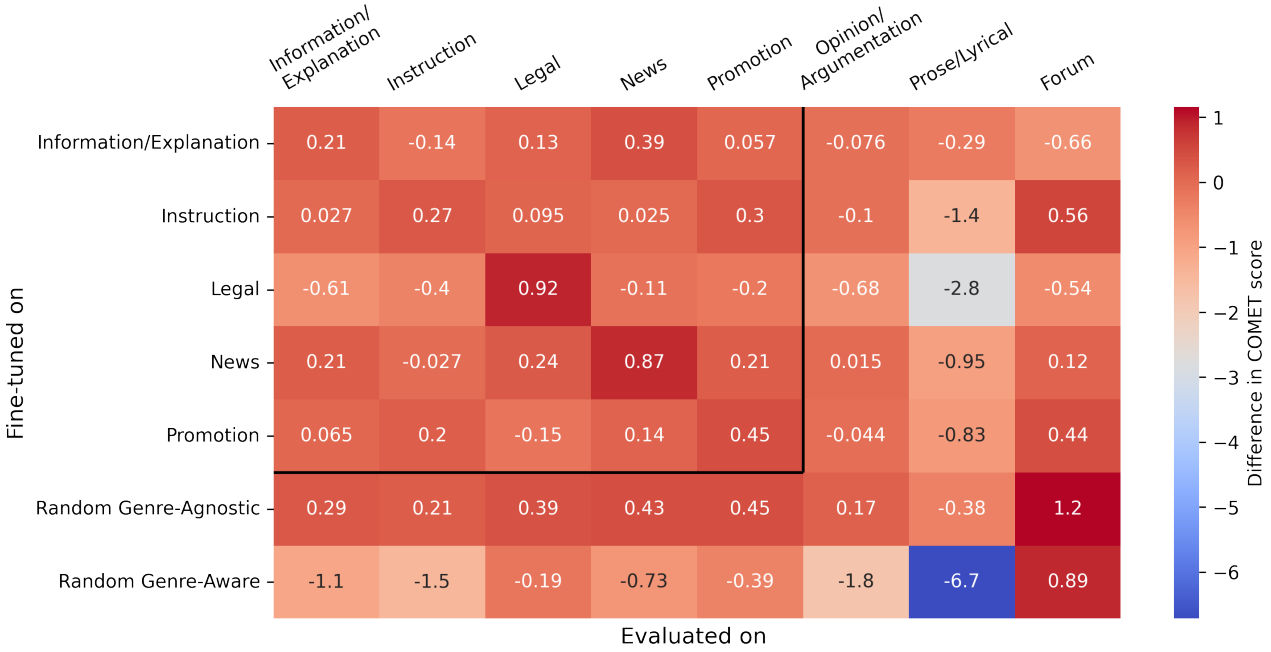

Figure 20: Difference in COMET score between Genre-specific models and the OPUS-MT baseline on the MaCoCu test set.

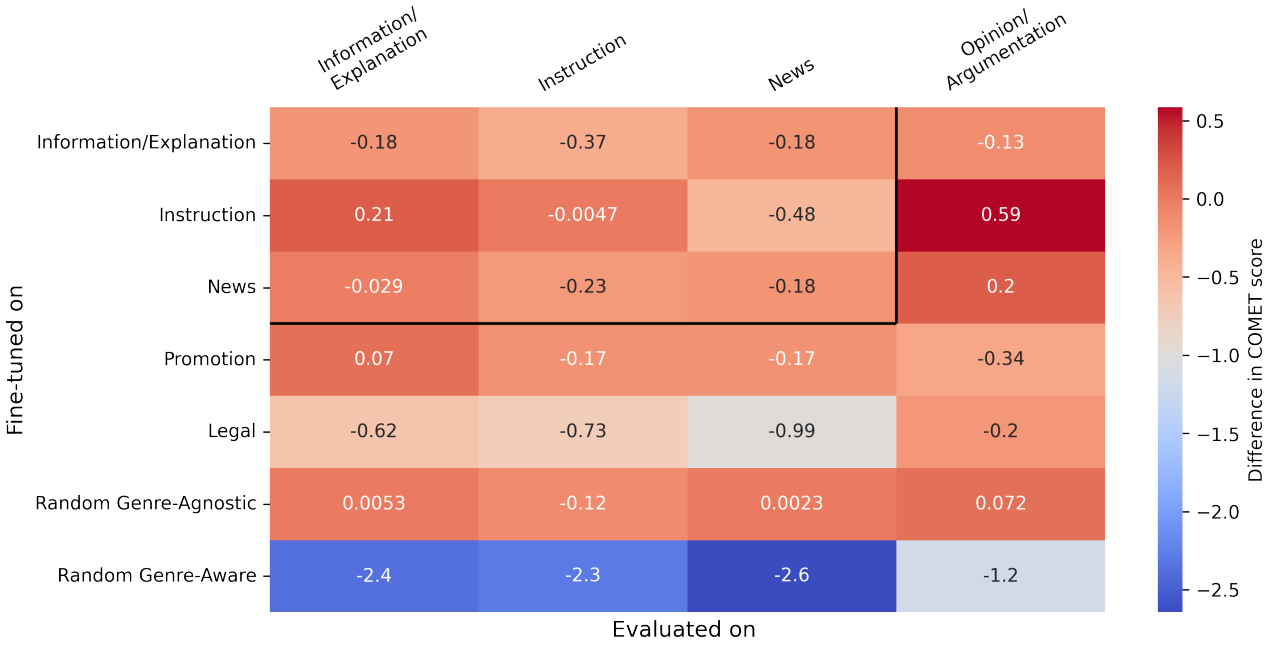

Figure 21: Difference in COMET score between Genre-specific models and the OPUS-MT baseline on the Flores dev set.

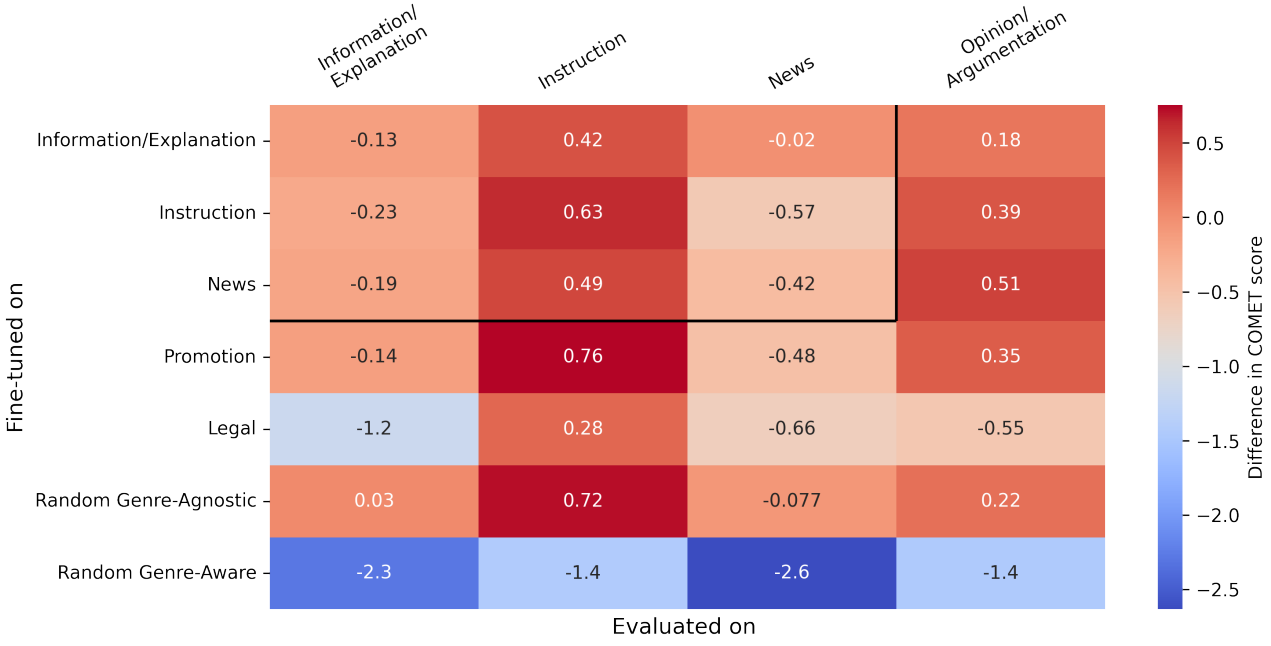

Figure 22: Difference in COMET score between Genre-specific models and the OPUS-MT baseline on the Flores devtest set.

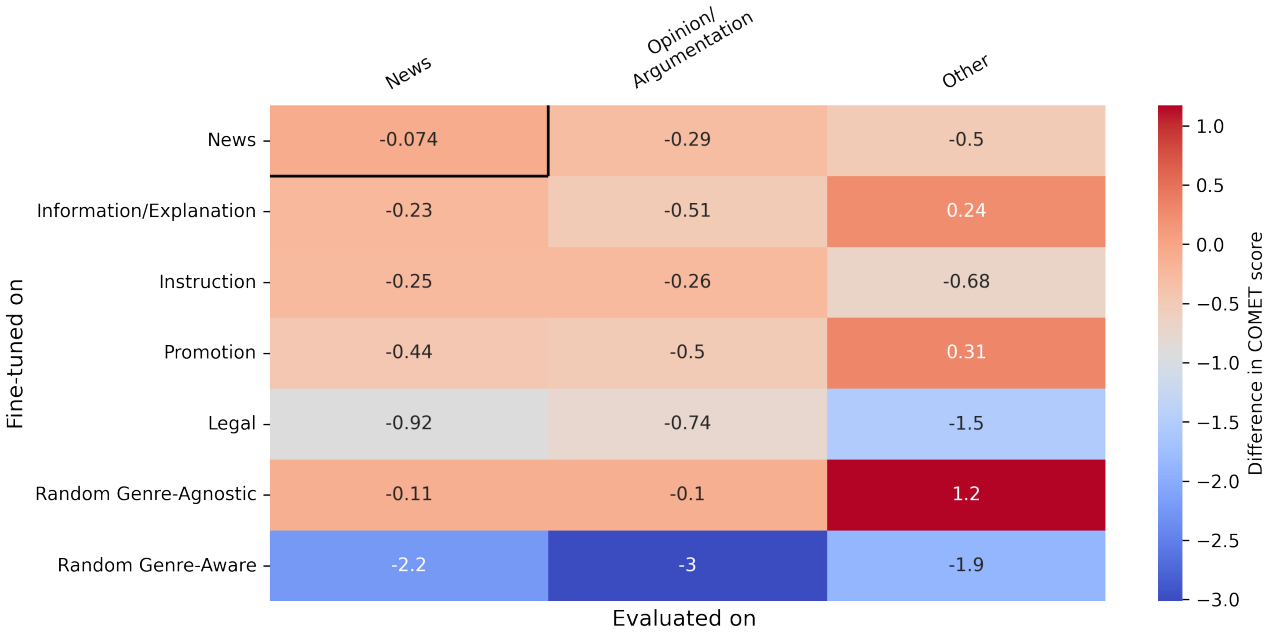

Figure 23: Difference in COMET score between Genre-specific models and the OPUS-MT baseline on the WMT18 test set.