Measuring Publication Bias in Foreign Language Editions of Russian State-Owned Media Company RT (www.rt.com)

Benjamin Figueroa
Claremont Mckenna College
bfigueroa20@cmc.edu

Mike Izbicki
Claremont Mckenna College
mizbicki@cmc.edu

Abstract

State-owned newspapers publish articles that support their sponsoring state's foreign policy. Political scientists therefore use the contents of these newspapers to better understand a state's policy objectives. Often, these newspapers publish multiple editions. These editions are written in different languages and target different regions of the world. By analyzing which articles get published in which language, we can get a more fine-grained understanding of the sponsoring state's policy towards each region.

We introduce the *Missing Content* task for performing this analysis, and present a fully automated solution based on the *Multilingual Universal Sentence Encoder* (Yang et al., 2019). As a motivating example, we analyze publication bias in Russian state-owned media company RT. We are able to quantitatively show that the English and Spanish editions of RT contain similar content, but that the English and Arabic editions contain different content, which could indicate that Russia has similar foreign policy objectives in the English and Spanish speaking world, but different objectives for the Arabic speaking world.

1 Introduction

RT (formerly Russia Today) is a media company owned by the Russian government that publishes articles promoting Russian foreign policy. RT publishes articles in six languages (Arabic, English, French, German, Russian, and Spanish). The goal of this paper is to discover how RT's coverage of different news topics varies by language. Our primary tool is the *Multilingual Universal Sentence Encoder* (MUSE) model (Yang et al., 2019), which can embed text written in all six languages into a common vector space.

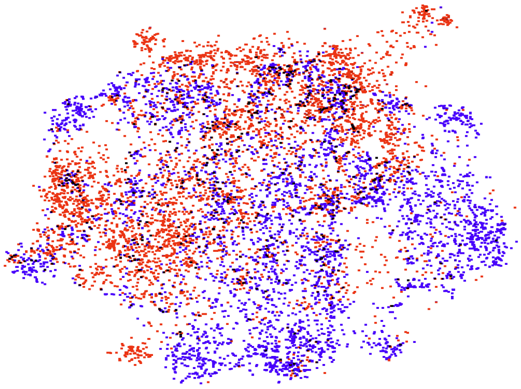

Figure 1: T-SNE projection of MUSE embeddings of RT's Arabic (blue) and Russian (red) articles. Each dot represents an individual article, and articles about similar topics are located close together. Some topics contain only Arabic documents, some contain only Russian documents, and some contain a mixture of documents in both languages.

Figure 1 shows a t-SNE projection (Maaten and Hinton, 2008) of the MUSE embeddings for articles written in Arabic and Russian. Using this embedding, we were able to discover that the Arabic edition's coverage of COVID-19 contains many articles about how Middle Eastern states are effectively combating the pandemic, the Russian edition's coverage contains articles about how former Soviet states are effectively combatting the pandemic, and both editions contain articles about how the United States is failing to combat the pandemic. This pattern of publication likely indicates that Russia is more interested in reducing American standing in the middle east than they are in improving their own standing. If the later situation were the case, they would also be publishing their articles about positive Russian responses to the COVID-19 pandemic in Arabic.

In the remainder of this paper, we attempt to answer the question: How can we quantify the publication bias between languages? In Section 2, we formally define the *Missing Content* task. This task generalizes the standard bitext retrieval task to the domain of missing content. In Section 3, we then apply the MUSE model to solve the missing content task. We use the standard `UNCorpus` dataset (Ziemski et al., 2016) to generate a synthetic missing content task with available ground truth to demonstrate that the MUSE model is able to effectively solve the missing content task. Then we introduce a novel dataset `RTArticles` containing all articles published by RT online, and we use the MUSE model to estimate which content is missing from which languages. Section 4 concludes with a discussion of how to extend this work to other state-run media like Voice of America and the BBC.

2 The Missing Content Task

In this section we introduce the *Missing Content* (MC) Task as a generalization of the *Bitext Retrieval* (BR) Task to the missing data regime. BR is a standard task in multilingual document retrieval that is used to measure the quality of multilingual document embeddings (Tiedemann, 2011). We begin by introducing our notation and formally describing the BR task. Then we formally define our MC generalization.

2.1 Bitext Retrieval (BR) Task

We are given a corpus of n text documents, and each document has a translation into into several different languages. Let $C_\ell = \{c_\ell^1, c_\ell^2, ..., c_\ell^n\}$ be the subcorpus for language ℓ, where c_ℓ^i is the ith document in the corpus translated into language ℓ. In particular, for any document index i and any two languages ℓ and ℓ', the documents c_ℓ^i and $c_{\ell'}^i$ contain the same text translated into different languages. In the *Bitext Retrieval* (BR) Task, we are given a source document $\gamma \in C_\ell$ and target language ℓ', and the goal is to find the document $\gamma' \in C_{\ell'}$ that is a translation of γ. This problem is difficult because the translation function cannot access the document indices and must operate only on the document text.

The standard solution to the BR Task uses document embeddings. Let $f : \texttt{text} \to \mathbb{R}^d$ be a function that embeds text into a d-dimensional vector space; that is, it converts any document c_ℓ^i into a vector. A good embedding function function should satisfy two properties. First, it should embed similar documents into similar vectors regardless of the documents' languages. Second, f should embed dissimilar documents into dissimilar vectors.

To solve the BR Task using document embeddings, we first compute the distance matrix M defined by

$$M_{ij} = \|f(c_\ell^i) - f(c_{\ell'}^j)\|. \tag{1}$$

Then for each document i in language ℓ, we compute its translation by finding the document in language ℓ' with minimum distance. That is, we compute

$$\hat{\jmath}(i) = \arg\min_{j\in[n]} M_{ij} \tag{2}$$

and set the translation of c_ℓ^i to be $c_{\ell'}^{\hat{\jmath}(i)}$. If $\hat{\jmath}(i) = i$, then we say the translation is correct, otherwise we say the translation is incorrect. The overall accuracy of the BR Task is given by

$$\mathrm{acc}_{BR}(\ell, \ell') = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\left[\hat{\jmath}(i) = i\right]. \tag{3}$$

2.2 Missing Content (MC) Task

The *Missing Content* (MC) Task is the natural extension of the BR Task to the missing data setting. In particular, the MC Task allows each of the input c_ℓ^i variables to be either a translation of document i into language ℓ or the special value `None` if no such translation is provided. The goal of the MC Task is then the same as the goal of the BR Task: Take an input document $\gamma \in C_\ell$ and target language ℓ', and output the corresponding value $\gamma' \in C_{\ell'}$. In the BR Task, the output γ' is guaranteed to be a document, but in the MC Task, the output γ' may also be the special token `None` if no suitable translation is found.

We propose to solve the MC task using a natural generalization of the standard BR algorithm above. Calculate the M matrix as

$$M_{ij} = \begin{cases} \infty & \text{if } c_\ell^i \text{ or } c_{\ell'}^j \text{ is } \texttt{None} \\ \|f(c_\ell^i) - f(c_{\ell'}^j)\| & \text{otherwise} \end{cases}$$

and calculate

$$\hat{\jmath}(i) = \begin{cases} \arg\min_{j\in[n]} M_{ij} & \text{if } \min_{j\in[n]} M_{ij} < \tau \\ \texttt{None} & \text{otherwise} \end{cases},$$

where τ is a threshold hyperparameter that controls the trade-off between the precision versus recall of our algorithm.

The generalization of the accuracy to the MC Task is given by

$$\mathrm{acc}_{MC}(\ell,\ell') = \frac{1}{n}\sum_{i=1}^{n}\mathbb{I}\Big[\hat{j}(i) = i \\ \text{or } (\hat{j}(i) = \texttt{None} \text{ and } c_\ell^i = \texttt{None})\Big].$$

3 Experiments

We perform experiments on two datasets. First, we use the standard `UNCorpus` dataset (Ziemski et al., 2016) to demonstrate that the MUSE model (Yang et al., 2019) effectively solves the missing content task. Then, we introduce the `RTArticles` dataset which, unlike the `UNCorpus` dataset (Ziemski et al., 2016), has no ground truth annotations. We apply the MUSE model to generate these annotations.

3.1 The `UNCorpus` Dataset

The United Nations Parallel Corpus (`UNCorpus`) is a standard, high quality corpus commonly used to evaluate bitext retrieval models (Ziemski et al., 2016). For example, it was the primary dataset used to evaluate the MUSE model we use in this paper (Yang et al., 2019). The dataset contains 68.2 million sentences that have been professionally translated into the six working languages of the United Nations (Arabic, Chinese, English, French, Russian, and Spanish). All of these languages are supported by the MUSE model, and this is a similar set of languages to the languages that RT publishes in.

In our first experiment, we measure how good the MUSE model is at separating equivalent sentences from non-equivalent sentences. Figure 2 plots the empirical distribution of the diagonal $M_{i,i}$ and off-diagonal $M_{i,j\neq i}$ entries of the M matrix for the Arabic-Russian language pair. Recall that $M_{i,j}$ represents the distance between document i in the Arabic language and document j in the Russian language, so the $M_{i,i}$ values indicate the distance between two translated documents, and the $M_{i,i\neq j}$ distances represent the distances between all non-translations. We can see a clear separation in these two distributions, indicating that the model is able to identify whether a document is a translation or not.

In our second experiment, we modify the `UNCorpus` dataset to make it suitable for the missing content task. To generate missing values, we use a sampling strategy. Specifically, given a language pair ℓ and ℓ', we randomly delete a fraction α sentences from the data for language ℓ'. Therefore, all data contained in ℓ' will be contained in ℓ, but not all data contained in ℓ will be contained in the sampled ℓ'. Our goal is then to predict which documents in ℓ are not contained in the sampled ℓ'. The MUSE model performs exceptionally well at this task. Figure 3 shows the ROC curve for the Arabic-Russian language pair. We achieve the excellent area under the curve value of 0.94. Surprisingly, the model's performance is essentially unaffected by the fraction of missing data α. This indicates that it should continue to perform well on non-synthetic data that can have an arbitrary fraction of missing data.

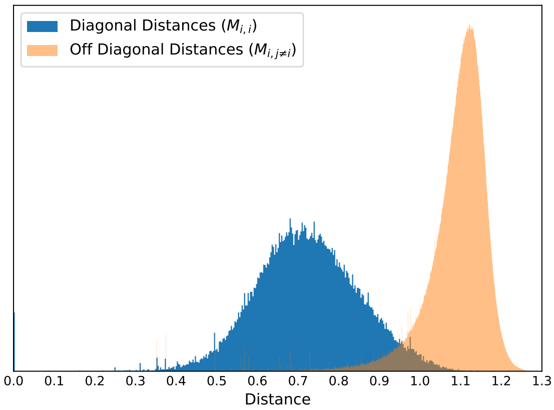

Figure 2: The distributions of diagonal ($M_{i,i}$) and off-diagonal ($M_{i,j\neq i}$) distances for the Arabic-Russian language pair. Other language pairs are similar. These distributions are well separated, indicating the MUSE model (Yang et al., 2019) is able to distinguish between translations in different languages. We recommend using the location where these distributions cross (0.9) to set the threshold τ for the missing content task.

Based on the results of these experiments, we recommend using a τ value of 0.9. This τ value is the distance where the $M_{i,i}$ and $M_{i,j\neq i}$ distributions cross, and provides a relatively good trade-off between true and false positives. On the Arabic-Russian dataset, for example, this τ achieves a true positive rate 0.980 and a false positive rate of 0.302. We optimize for a high true positive rate to ensure that all matching documents are found.

3.2 The `RTArticles` Dataset

We introduce the `RTArticles` dataset for the missing content task.This dataset contains newspaper article headlines scraped from RT's web-

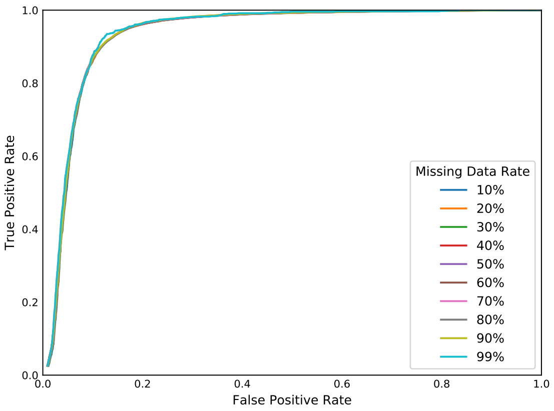

Figure 3: ROC plot for the Arabic-Russian language pair in the UNCorpus dataset with various missing data rates (α). We achieve the exceptional area under the curve value of 0.94. The model has excellent, nearly identical, performance across the full range of possible α values, which indicates that it will perform well on non-synthetic tasks like the RTArticles dataset where the true α is unknown.

sites. RT publishes articles in six languages (Arabic, English, French, German, Russian, and Spanish), and articles in each language are published on a different domain. For example, Arabic language articles are published under arabic.rt.com, and Russian language articles are published under russian.rt.com. To gather the data, we performed an exhaustive crawl of all subdomains of rt.com. Table 1 shows the size of the dataset in each language. Because the number of articles in each language is drastically different, clearly not all articles are being published in every language. We have no ground-truth annotations, however, about which articles are published in multiple languages. We therefore will use the MUSE model to predict which articles are published in multiple languages.

Following the results in Section 3.1, we set $\tau = 0.9$ and run the missing content algorithm. Table 2 shows the estimated fraction of articles that have translations into each language pair. The results show that some language pairs like English-Spanish language pair have a large fraction of articles in common. Since RT published similar content in both languages, Russia likely has similar foreign policy objectives for both the English and Spanish speaking regions of the world. Other language pairs like English-Arabic have relatively little content overlap, indicating that Russia has very different foreign policy objectives for the English and Arabic speaking world. Analysts can also use the t-SNE technique we demonstrated in the introduction to "zoom in" on the different topics that are shared and not shared between all language pairs.

Arabic	English	Spanish	French	German	Russian
54 492	144 409	88 104	32 744	35 244	301 624

Table 1: The number of documents in each language in the RTArticles dataset.

	ar	en	es	fr	de	ru
ar	1.000	0.439	0.374	0.218	0.254	0.404
en	0.331	1.000	0.532	0.392	0.377	0.463
es	0.326	0.589	1.000	0.385	0.385	0.426
fr	0.264	0.577	0.501	1.000	0.428	0.383
de	0.292	0.510	0.444	0.379	1.000	0.392
ru	0.213	0.377	0.258	0.157	0.178	1.000

Table 2: Estimated fraction of articles in the row language that are also present in the column language for the RTArticles dataset. The table is not symmetric primarily due to differences in dataset size.

4 Discussion

We introduced the missing content task for data retrieval, and we used the MUSE model (Yang et al., 2019) to solve this task for the RT newspaper. RT has been widely criticized for being a source of "Russian propaganda" (e.g. Yablokov, 2015; Van Herpen, 2015; Wright et al., 2020), but we emphasize that we chose to analyze RT primarily for technical and not political reasons. Western state-run newspapers such as Voice of America (VOA) and the British Broadcasting Corporation (BBC) are also frequently criticized for their propagandistic character (e.g. Herman and Chomsky, 1988; Rawnsley, 2016). From a technical perspective, however, analyzing the RT dataset was simpler. RT is published in only 6 languages, whereas VOA is published in 46 languages and the BBC is published in 44 languages. Since the MUSE model only supports 16 languages, it cannot be used for analyzing VOA or the BBC. We also ran our experiments using the multilingual BERT model (Devlin et al., 2018), which has support for 102 languages, and so would be able to work on the VOA and BBC datasets, but the BERT model achieved significantly worse accuracy than the MUSE model to the point of being unusable. As these multilingual models improve, however, we expect to be able to run full analyses on these western media corporations as well.

References

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Edward S Herman and Noam Chomsky. 1988. *Manufacturing consent: The political economy of the mass media*. Random House.

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605.

Gary D Rawnsley. 2016. *Radio diplomacy and propaganda: The BBC and VoA in international politics, 1956–64*. Springer.

Jörg Tiedemann. 2011. Bitext alignment. *Synthesis Lectures on Human Language Technologies*, 4(2):1–165.

Marcel H Van Herpen. 2015. *Putin's Propaganda Machine: Soft Power and Russian Foreign Policy*. Rowman & Littlefield.

Kate Wright, Martin Scott, and Mel Bunce. 2020. Soft power, hard news: How journalists at state-funded transnational media legitimize their work. *The International Journal of Press/Politics*, page 1940161220922832.

Ilya Yablokov. 2015. Conspiracy theories as a russian public diplomacy tool: The case of russia today (rt). *Politics*, 35(3-4):301–315.

Yinfei Yang, Daniel Cer, Amin Ahmad, Mandy Guo, Jax Law, Noah Constant, Gustavo Hernandez Abrego, Steve Yuan, Chris Tar, Yun-Hsuan Sung, et al. 2019. Multilingual universal sentence encoder for semantic retrieval. *arXiv preprint arXiv:1907.04307*.

Michał Ziemski, Marcin Junczys-Dowmunt, and Bruno Pouliquen. 2016. The united nations parallel corpus v1. 0. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3530–3534.