Multi-Target Cross-Lingual Summarization: a novel task and a language-neutral approach

Anonymous ACL submission

Abstract

Cross-lingual summarization aims to bridge language barriers by summarizing documents in different languages. However, ensuring semantic coherence across languages is an overlooked challenge and can be critical in several contexts. To fill this gap, we introduce multi-target cross-lingual summarization as the task of summarizing a document into multiple target languages while ensuring that the produced summaries are semantically similar. We propose a principled re-ranking approach to this problem and a multi-criteria evaluation protocol to assess semantic coherence across target languages, marking a first step that will hopefully stimulate further research on this problem.

1 Introduction

Cross-lingual summarization refers to the task of producing a summary in a different language than the original document and has the potential to break language barriers by helping people to effectively capture the essence of documents written in foreign languages (Wang et al., 2022). This is a very challenging task, as it combines the difficulties of monolingual summarization, such as factual inconsistencies with respect to the source document (Maynez et al., 2020), with those of machine translation, such as translation of idiomatic expressions and cultural references (Fadaee et al., 2018).

The availability of large pre-trained multilingual transformers (Liu et al., 2020; Xue et al., 2021), followed by the widespread development and adoption of decoder-only language models (Radford et al., 2018; Touvron et al., 2023; Jiang et al., 2023; Team et al., 2024) has enabled a single model to perform cross-lingual summarization from multiple source languages to multiple target languages (many-to-many summarization, M2MS). Despite the increasing emphasis on this many-to-many paradigm, ensuring semantic coherence in summaries across different target languages has not been a primary focus of state-of-the-art methods, nor has it been systematically evaluated. Table 1 illustrates this issue by presenting an example where a state-of-the-art M2MS system based on mT5 (Xue et al., 2021) produces very different summaries, with one containing unfaithful content, depending on the chosen target language. Clearly, if information is not conveyed coherently across languages, the trustworthiness of the system is compromised. Users cannot rely on the summaries to be accurate and unbiased, regardless of the language in which they consume the content. In addition, in legal or regulatory contexts, ensuring that information is presented coherently across languages can be critical. This helps meet regulatory requirements and ensures that information is transmitted coherently across language boundaries.

To fill this gap, we introduce a novel variant of cross-lingual summarization, which we call *multi-target cross-lingual summarization* (MTXLS), where we specifically address the challenge of promoting semantic coherence across target languages. This framework represents an important step towards more comprehensive cross-lingual summarization techniques and evaluation. Our main contributions in this work are summarized as follows: First, we introduce MTXLS formally as a novel task (Section 3), motivated by the need of producing summaries coherently for multiple target languages. Second, we present a re-ranking-based approach to address this problem (Section 4), where the re-ranking phase selects a set of summaries that exhibit superior semantic coherence across target languages compared to treating each cross-lingual summarization task independently. Notably, our approach circumvents the need for a pivot language. This language-neutral strategy ensures a more robust and unbiased multilingual summarization process. Finally, we propose and conduct a multi-criteria evaluation protocol that goes beyond the simple evaluation of the similarity between gener-

Text (BBC)	en	Bitcoin uses more electricity annually than the whole of Argentina, analysis by Cambridge University suggests. "Mining" for the cryptocurrency is power-hungry, involving heavy computer calculations to verify transactions. Cambridge researchers say it consumes around 121.36 terawatt-hours (TWh) a year - and is unlikely to fall unless the value of the currency slumps. (...)
`M2MS` (mT5)	es	Bitcoin es la moneda más consumida del mundo, según un estudio de la Universidad de Cambridge.
	fr	Bitcoin consomme plus d'électricité chaque année, selon une analyse de l'université de Cambridge.
`Mistral 7B`	es	Los investigadores de la Universidad de Cambridge estiman que el Bitcoin consume alrededor de 121,36 TWh de energía eléctrica al año, lo que lo convierte en un proceso energético intenso y poco probable que disminuya a menos que la valoración de la moneda baje significativamente.
	fr	Des chercheurs de l'Université de Cambridge estiment que le Bitcoin consomme environ 121,36 TWh d'électricité par an, ce qui en fait un processus énergivore qui est peu probable de décroître si la valeur de la monnaie chute significativement.
`NeutralRR` (ours)	es	Bitcoin consume más electricidad que Argentina, según un estudio de la Universidad de Cambridge.
	fr	Bitcoin consomme plus d'électricité que l'Argentine, selon une analyse de l'université de Cambridge.

Table 1: An example of an en→{es, fr} summarization task solved by three different state-of-the-art systems, including ours. Text in red marks information that is present in a summary for one of the languages but not in the other summary.

ated summaries and references (Section 5). Specifically, we incorporate the important aspect of evaluating the coherence of the entire set of generated summaries across all target languages using quality estimation methods for machine translation. The code and data used in our experiments are publicly available.[1]

2 Related Work

2.1 Cross-Lingual Summarization

Research in cross-lingual summarization has recently gained traction, in part due to the increased availability of large datasets for this task (Ladhak et al., 2020; Perez-Beltrachini and Lapata, 2021). Among these, CrossSum (Bhattacharjee et al., 2023) stands out as the most resourceful. This news dataset contains document-summary pairs for 12 different languages and more than 1,500 language directions, and it was built by automatically pairing the data from the multilingual dataset XL-Sum (Hasan et al., 2021), which consists of news articles from BBC.

Earlier cross-lingual summarization models operated on a per-language-pair basis (Cao et al., 2020; Bai et al., 2021; Liang et al., 2022). However, with the emergence of large pre-trained multilingual transformers like mBART (Liu et al., 2020) and mT5 (Xue et al., 2021), alongside extensive cross-lingual summarization datasets covering multiple language directions, a shift to many-to-many approaches occurred (Bhattacharjee et al., 2023; Chen et al., 2023b; Wang et al., 2023b). Evaluation expanded to include large decoder-only language models, including in a zero-shot setting, with only GPT-4 showing competitive performance compared to fine-tuned mBART-50 (Wang et al., 2023a; Tang et al., 2021). The approaches most akin to our setting in the cross-lingual summarization literature either involve first generating a summary in the source language and then using it to guide the generation of the target language summary (Bai et al., 2021), or employing a content plan generation step to condition the decoding of the target summary (Huot et al., 2024). However, they do not explicitly enforce or evaluate semantic similarity across summaries in different target languages.

2.2 Quality Estimation for Machine Translation

In machine translation (MT), quality estimation methods aim to predict translation quality without access to gold standard outputs (Specia et al., 2013, 2018). Our focus is on using sentence-level MT quality estimation to evaluate semantic coherence in the generated summaries across target languages, by taking two system-generated summaries for different languages and evaluating how well one translates the other.

Quality estimation methods for MT can be performed at various levels: word-level, where binary labels (`OK` or `BAD`) are assigned to each machine-translated word, and sentence- or document-level, where a score is generated as an estimate of the quality of the whole translated sentence or document. Many quality estimation methods produce both word-level and sentence-level scores (Wang et al., 2018; Kepler et al., 2019a,b; Lee, 2020). A sentence-level quality estimation method can arise from training multilingual sentence encoders like LASER (Artetxe and Schwenk, 2019) or SONAR (Duquenne et al., 2023). These models align representations of translated sentences, allowing embedding similarity metrics in the common space to serve as quality estimation metrics for MT. BLASER (Chen et al., 2023a), an automatic text-free metric for evaluating speech translation,

[1] URL available upon acceptance.

refines this idea by using a regression model trained on the concatenation of the LASER embeddings of the source text and the reference and machine-generated translations. BLASER 2.0 (Communication et al., 2023) replaces LASER with SONAR embeddings, supports both speech and text modalities, and exists in both reference-dependent and reference-free (i.e., quality estimation) variants. Similarly, COMET (Rei et al., 2020) was initially introduced as a reference-dependent metric that cross-encodes the source text and the reference and machine-generated translations using an XLM-RoBERTa model (Conneau et al., 2020). Later, a similar idea was followed to build its reference-free version, called CometKiwi (Rei et al., 2022).

3 Multi-Target Cross-Lingual Summarization

3.1 Problem Formulation

This section formalizes the task of MTXLS. Let $\boldsymbol{x}_o \in \mathcal{X}$ represent a document in the source language o, and let $\mathcal{T} = \{t_1, t_2, \ldots, t_N\}$ denote a set of N target languages. Without loss of generality, we assume that $o \in \mathcal{T}$. The primary goal of MTXLS is to generate a set of N summaries, denoted as $\mathcal{S} = \{\boldsymbol{y}_{t_1}, \boldsymbol{y}_{t_2}, \ldots, \boldsymbol{y}_{t_N}\}$, where there is a summary $\boldsymbol{y}_{t_i} \in \mathcal{Y}$ for each language in $\mathcal{T}$.

It is evident that this task can be seen as a combination of a monolingual summarization task in language o and $N-1$ cross-lingual summarization tasks from o to each target language $t \in \mathcal{T} \setminus \{o\}$. While these tasks could be approached independently, we impose a constraint: all N summaries should convey identical information regardless of the language. This constraint ensures the alignment of information across different languages, thus promoting coherence in the resulting set of summaries.

3.2 Summarize-and-Translate

Consider a scenario where a summarization model is available for generating summaries from language o to a pivot language π. Additionally, there are models for translating from π to each language in $\mathcal{T}$. Common statistical approaches to these tasks involve modeling the summarization distribution $p(\boldsymbol{y}_\pi \mid \boldsymbol{x}_o, \pi)$ and the translation distributions $p(\boldsymbol{y}_t \mid \boldsymbol{y}_\pi, t)$ for each $t \in \mathcal{T}$.

To enforce the desired coherence constraint across target languages, a simple strategy is to assume that the target summaries are conditionally independent of the source document given the pivot summary, expressed as $(\boldsymbol{y}_t \perp\!\!\!\perp \boldsymbol{x}_o) \mid \boldsymbol{y}_\pi, \forall t \in \mathcal{T}$ and entailed by the Bayesian network in Figure 1a. This implies that, for each target language t, the information utilized to generate $\boldsymbol{y}_t$ from $\boldsymbol{x}_o$ comes solely from $\boldsymbol{y}_\pi$. Notably, since translation is a more deterministic task than summarization, this assumption serves to mitigate the potential variability of $\boldsymbol{y}_t$ across different target languages.

The previous assumption allows us to write the cross-lingual summarization distributions that use π as the pivot language as:

$$\begin{aligned} p(\boldsymbol{y}_t \mid \boldsymbol{x}_o, t, \pi) &= \sum_{\boldsymbol{y}_\pi} p(\boldsymbol{y}_\pi \mid \boldsymbol{x}_o, \pi) p(\boldsymbol{y}_t \mid \boldsymbol{y}_\pi, t) \\ &= \mathbb{E}_{\boldsymbol{y}_\pi \mid \boldsymbol{x}_o, \pi} p(\boldsymbol{y}_t \mid \boldsymbol{y}_\pi, t), \end{aligned} \quad (1)$$

for each $t \in \mathcal{T}$. Approximating this expectation with a single sample and using the source language as the pivot language yields the conventional summarize-and-translate approach to cross-lingual summarization. While this baseline ensures coherence across multiple target languages by deriving summaries from the translation of the same pivot summary, it has inherent drawbacks. In particular, it involves two successive phases of decoding: first generating the pivot summary, and then generating summaries for each target language, thus potentially suffering from error accumulation from both decoding phases. Moreover, it is likely to degrade the similarity to the reference summaries in the target languages because it is biased towards the pivot language. Thus, all resulting summaries will reflect any biases introduced during the summarization from language o to language π.

4 Methodology

4.1 Beyond Summarize-and-Translate

We now relax the conditional independence assumption made previously by explicitly conditioning $\boldsymbol{y}_t$ on $\boldsymbol{x}_o$, as shown in Figure 1b. Notably, this approach does not involve decoding $\boldsymbol{y}_t$ after $\boldsymbol{y}_\pi$, but rather allows the two processes to run in parallel, and explicitly promotes semantic similarity between $\boldsymbol{y}_\pi$ and each $\boldsymbol{y}_t$, as required to satisfy our constraint. We now have:

$$p(\boldsymbol{y}_t \mid \boldsymbol{x}_o, t, \pi) = \mathbb{E}_{\boldsymbol{y}_\pi \mid \boldsymbol{x}_o, \pi} p(\boldsymbol{y}_t \mid \boldsymbol{x}_o, \boldsymbol{y}_\pi, t). \quad (2)$$

Let us impose that:

$$p(\boldsymbol{y}_t \mid \boldsymbol{x}_o, \boldsymbol{y}_\pi, t) = \frac{1}{Z} \phi(\boldsymbol{y}_t, \boldsymbol{y}_\pi) q(\boldsymbol{y}_t \mid \boldsymbol{x}_o, t), \quad (3)$$

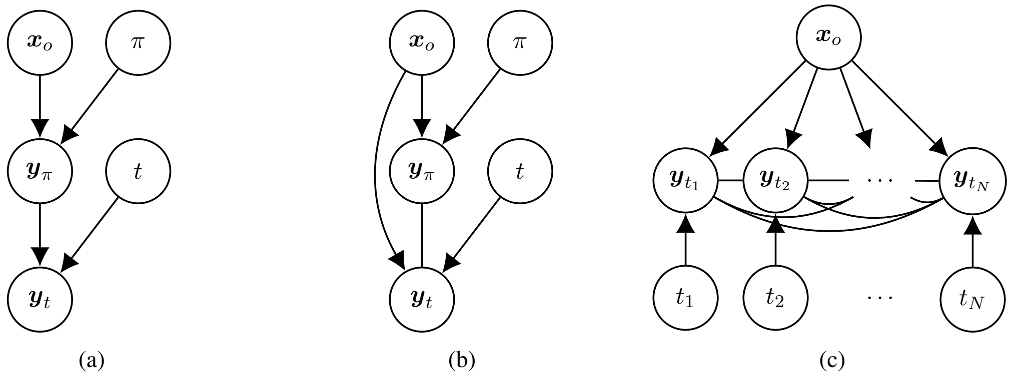


Figure 1: Graphical models representing summarize-and-translate (a), our method with a pivot language (b), and our language-neutral approach (c). Here, $\boldsymbol{x}_o$ denotes the document in the source language o, $\boldsymbol{y}_\pi$ denotes the summary in the pivot language π, and $\boldsymbol{y}_{t_i}$ denotes the summary in the target language t_i, $i \in \{1, 2, \dots, N\}$.

where Z is a normalizing function independent of $\boldsymbol{y}_t$, $\phi : \mathcal{Y}^2 \mapsto \mathbb{R}^+$ is a symmetric function measuring the semantic similarity between two texts in different languages and satisfies $\sum_{\boldsymbol{y}_t} \phi(\boldsymbol{y}_t, \cdot) < \infty$, and $q(\boldsymbol{y}_t \mid \boldsymbol{x}_o, t)$ is modeled by a cross-lingual summarization system from language o to language t. This formulation explicitly addresses both of our goals: to produce a text $\boldsymbol{y}_t$ that serves as a good summary of $\boldsymbol{x}_o$ in language t and has a high similarity to the pivot $\boldsymbol{y}_\pi$. Finally, we get:

$$
\begin{aligned}
p(\boldsymbol{y}_t \mid \boldsymbol{x}_o, t, \pi) &= \mathbb{E}_{\boldsymbol{y}_\pi \mid \boldsymbol{x}_o, \pi} \frac{1}{Z} \phi(\boldsymbol{y}_t, \boldsymbol{y}_\pi) q(\boldsymbol{y}_t \mid \boldsymbol{x}_o, t) \\
&\approx \frac{1}{Z} \phi(\boldsymbol{y}_t, \boldsymbol{y}_\pi) q(\boldsymbol{y}_t \mid \boldsymbol{x}_o, t) \\
&\propto \phi(\boldsymbol{y}_t, \boldsymbol{y}_\pi) q(\boldsymbol{y}_t \mid \boldsymbol{x}_o, t), \qquad (4)
\end{aligned}
$$

where $\boldsymbol{y}_\pi \sim p(\boldsymbol{y}_\pi \mid \boldsymbol{x}_o, t)$. This framework unveils diverse avenues for MTXLS. One is to directly train $p(\boldsymbol{y}_t \mid \boldsymbol{x}_o, t, \pi)$ by jointly learning ϕ and q from data, which requires cross-lingual document-summary pairs for all target languages and parallel data between the pivot and each target language. Alternatively, ϕ could be used as a re-scoring function at each decoding step from q, but this would introduce a significant computational burden.

In our work, we adopt a simpler re-ranking approach. We use q to generate k candidate summaries for each target language t, and then use ϕ to select the optimal candidate. Notably, this allows simultaneous generation of candidate and pivot summaries, and enhances the semantic coherence of generated summaries while maintaining similarity to the reference cross-lingual distribution used to train the summarizer, which were not possible in the summarize-and-translate approach. As shown in Section 4.3, our approach has a deep connection with rejection sampling.

4.2 A Language-Neutral Formulation

Despite not using translation to obtain summaries for the target languages, the approach we have described in Section 4.1 still relies in a pivot language. However, following the same formulation, we can circumvent this issue by defining a joint distribution for the summaries in all the target languages:

$$
p(\mathcal{S} \mid \boldsymbol{x}_o, \mathcal{T}) \propto \varphi(\mathcal{S}) \prod_{i=1}^{N} q(\boldsymbol{y}_{t_i} \mid \boldsymbol{x}_o, t_i), \qquad (5)
$$

where

$$
\varphi(\mathcal{S}) = \frac{1}{\binom{N}{2}} \sum_{i,j:\, j>i} \phi(\boldsymbol{y}_{t_i}, \boldsymbol{y}_{t_j}) \qquad (6)
$$

measures the semantic similarity of the set of summaries $\mathcal{S}$ by averaging all the pairwise similarities between each pair of summaries in $\mathcal{S}$. This model is represented graphically in Figure 1c. Note that the formulation in Section 4.1 is a particular case of this one where $\mathcal{S} = \{\boldsymbol{y}_t, \boldsymbol{y}_\pi\}$ and $p(\mathcal{S} \mid \boldsymbol{x}_o, \mathcal{T}) = p(\boldsymbol{y}_t \mid \boldsymbol{x}_o, t, \pi) q(\boldsymbol{y}_\pi \mid \boldsymbol{x}_o, \pi)$.

4.3 Summary Sampling

Our primary goal is now to conceive a method that allows us to sample summaries from:

$$
p(\mathcal{S} \mid \boldsymbol{x}_o, \mathcal{T}) = \frac{\varphi(\mathcal{S})}{Z'} \prod_{i=1}^{N} q(\boldsymbol{y}_{t_i} \mid \boldsymbol{x}_o, t_i). \qquad (7)
$$

We demonstrate we can achieve this goal through rejection sampling, which works as follows. Given a distribution $f(\boldsymbol{x})$ from which we aim to sample and a proposal distribution $g(\boldsymbol{x})$ satisfying $\sup_{\boldsymbol{x}} \frac{f(\boldsymbol{x})}{g(\boldsymbol{x})} \leq M$, we start by generating a sample $\boldsymbol{x}$ from g and a sample u uniformly in $[0, 1]$. Subsequently, we accept $\boldsymbol{x}$ if $\frac{f(\boldsymbol{x})}{Mg(\boldsymbol{x})} \geq u$ and reject it otherwise.

In our context, we may use $\prod_{i=1}^{N} q(\boldsymbol{y}_{t_i} \mid \boldsymbol{x}_o, t)$ as the proposal distribution and assume without loss of generality that ϕ is bounded in $[0,1]$, so φ is also bounded in $[0,1]$ and therefore:

$$\sup_{\mathcal{S}} \frac{p(\mathcal{S} \mid \boldsymbol{x}_o, \mathcal{T})}{\prod_{i=1}^{N} q(\boldsymbol{y}_{t_i} \mid \boldsymbol{x}_o, t_i)} = \sup_{\mathcal{S}} \frac{\varphi(\mathcal{S})}{Z'} \leq \frac{1}{Z'}. \quad (8)$$

Thus, $M = \frac{1}{Z'}$ satisfies the condition above. The rejection sampling procedure for sampling from $p(\mathcal{S} \mid \boldsymbol{x}_o, \mathcal{T})$ is then:

1. Sample $\mathcal{S}$ by sampling $\boldsymbol{y}_t \sim q(\boldsymbol{y}_t \mid \boldsymbol{x}_o, t)$ independently for each $t \in \mathcal{T}$.
2. Sample $u \sim U(0,1)$.
3. Accept $\mathcal{S}$ if $\varphi(\mathcal{S}) \geq u$; otherwise, reject it.

In step 1, summaries can be sampled independently and in parallel for each target language because of the factorized form of the proposal distribution.

4.4 A Mode-Seeking Heuristic

The procedure presented in Section 4.3 offers a systematic means to sample sets of summaries from the distribution $p(\mathcal{S} \mid \boldsymbol{x}_o, \mathcal{T})$. However, in many practical scenarios, the objective is to obtain a single set of high-quality summaries, i.e. a set with high probability under this distribution. This goal motivates the approach we present here.

Let us assume we can generate k candidate summaries for each target language using diverse beam search (Vijayakumar et al., 2018) or a sampling algorithm. In this setup, there are k^N different sets of summaries resulting from the different combinations of selecting a candidate from each target language. Among these sets, we wish to choose the set $\mathcal{S}^*$ that maximizes $\varphi(\mathcal{S})$, in order to achieve our goal of having a maximally semantically coherent set of summaries. Interestingly, this criterion corresponds to choosing the set $\mathcal{S}^*$ with maximum probability of being accepted in the rejection sampling procedure described in Section 4.3.

However, finding $\mathcal{S}^*$ among the k^N candidate sets is an instance of the generalized maximum clique problem, which is NP-hard (Feremans et al., 2003), and therefore we must resort to a heuristic search. For this purpose, we introduce a random permutation σ of the target languages $\mathcal{T}$, e.g. $\sigma(\mathcal{T}) = (t_N, t_{N-1}, \ldots, t_1)$, and define the proxy similarity function as follows:

$$\hat{\varphi}(\mathcal{S}; \sigma) = \frac{1}{N-1} \sum_{i=1}^{N-1} \phi(\boldsymbol{y}_{\sigma(\mathcal{T})_i}, \boldsymbol{y}_{\sigma(\mathcal{T})_{i+1}}). \quad (9)$$

Algorithm 1 Language-neutral multi-target cross-lingual summarization

Require: Input document ($\boldsymbol{x}_o$); Set of target languages ($\mathcal{T}$, with size N); Number of candidates per language (k); Number of random permutations (m).

```
for each t ∈ T do                                    ▷ Generate candidates
    for i ← 1 to k do
        Sample y_t^(i) ~ q(y_t | x_o, t).
    end for
end for
for i ← 1 to m do                         ▷ Find set with high similarity
    Build a weighted directed graph G = (V, E), where V
    has Nk + 2 nodes, one for each candidate summary plus a
    source and a sink node, and E ← ∅.
    Sample a random permutation σ(T) = (t'_1, t'_2, ..., t'_N).
    E ← E ∪ {(source → y_{t'_1}^(i), 0)}_{i=1}^k
    E ← E ∪ {(y_{t'_N}^(i) → sink, 0)}_{i=1}^k
    for l ← 1 to N − 1 do
        E ← E ∪ {(y_{t'_l}^(i) → y_{t'_{l+1}}^(j), 1 − φ(y_{t'_l}^(i), y_{t'_{l+1}}^(j)))}_{i,j=1}^k
    end for
    Ŝ_i^* ← shortest path(G, source, sink)
end for
return Ŝ^* ← arg max_{S ∈ {Ŝ_1^*, ..., Ŝ_m^*}} φ(S)          ▷ eq. (6)
```

This proxy represents a sparsification of the clique in the graphical model shown in Figure 1c, where only the edges connecting adjacent target summaries according to the permutation σ are retained. This sparsification embodies the assumption of transitivity in semantic similarity: For any three languages a, b, and c, if the summary $\boldsymbol{y}_a$ is similar to $\boldsymbol{y}_b$, and $\boldsymbol{y}_b$ is similar to $\boldsymbol{y}_c$, then it follows that $\boldsymbol{y}_a$ should also share a significant degree of similarity with $\boldsymbol{y}_c$. Notably, the set that maximizes $\hat{\varphi}(\mathcal{S}; \sigma)$ can be found in $O(Nk^2)$ time using dynamic programming. This observation motivates Algorithm 1, where we consider k candidate summaries per target language and $m \ll N!$ random permutations of the target languages. Then, for each permutation, we find the candidate set $\hat{\mathcal{S}}_i^*$ that maximizes $\hat{\varphi}(\mathcal{S}; \sigma_i)$ using dynamic programming. Finally, we choose the set among $\hat{\mathcal{S}}_1^*, \hat{\mathcal{S}}_2^*, \ldots, \hat{\mathcal{S}}_m^*$ that has the highest score according to φ.

4.5 Choice of ϕ

So far, we have presented our methodology in a formal manner, but have not yet provided specifics on implementing a function ϕ capable of measuring the semantic similarity between two summaries in different languages. In practice, any quality estimation model for MT (Section 2.2) could be used. In our experiments, we leverage the cosine similarity of SONAR embeddings (Duquenne et al., 2023) as the similarity metric, reserving BLASER 2.0 (Chen et al., 2023a) and CometKiwi (Rei et al., 2022) for evaluation. Our selection of the cosine similarity of SONAR embeddings is motivated by its symmetry, unlike the remaining options, and the fact that the SONAR encoder is relatively lightweight. Specifically, we define the similarity function as:

$$\phi(\boldsymbol{y}_a, \boldsymbol{y}_b) = \frac{1 + \boldsymbol{s}_a^\top \boldsymbol{s}_b}{2}, \tag{10}$$

where $\boldsymbol{s}_a$ and $\boldsymbol{s}_b$ represent the L_2-normalized SONAR embeddings of summaries $\boldsymbol{y}_a$ and $\boldsymbol{y}_b$.

5 Experiments

5.1 Dataset

We use data from the CrossSum dataset, which contains documents and summaries in seven languages: Arabic, Chinese (simplified), English, French, Portuguese, Russian, and Spanish. CrossSum pairs documents in one language with summaries from documents in another language, using automatic similarity metrics. However, mispairings are frequent due to this automated process. Additionally, the dataset is designed for single-target cross-lingual summarization and does not perfectly fit our multi-target setting. To adapt the dataset to our needs, we restructured the dataset into clusters. This process is explained in Appendix A. Each resulting cluster consists of up to seven multilingual document-summary pairs, with one such pair for each language. This allows us to select any document within the cluster as a source for summarization, with all summaries within the cluster serving as references for each of the languages. Statistics about the clustered data and an analysis of the semantic coherence of the dataset summaries are also provided in Appendix A.

5.2 Methods

Our pivot-free re-ranking method (`NeutralRR`) proposed in Algorithm 1 was tested using $k = 8$ candidates per target language for re-ranking and $m = 6$ language permutations, unless otherwise specified. We study the effects of varying k and m in Section 5.5 and Appendix D.2, respectively. We compare our method with four other approaches, namely: a many-to-many summarizer with beam search decoding (`M2MS`) with a beam size of 8; the summarize-and-translate approach (`S&T`), where summaries are obtained in the source language and then translated to each of the target languages using beam search with a beam size of 8 in both decoding steps; a `Mistral 7B` (Jiang et al., 2023) large language model (LLM) used in a zero-shot setting and instructed to write summaries with identical information for all the target languages (see Appendix C); our pivot-dependent re-ranking approach (`PivotRR`) as described in Section 4.1, where we use the source language as the pivot.

All summaries except those of `Mistral 7B` were decoded from the same mT5 base model (Xue et al., 2021) fine-tuned in CrossSum. In the `S&T` approach, translations were performed using the NLLB 1.3B model (Costa-jussà et al., 2022). `NeutralRR` and `PivotRR` used beam search multinomial sampling using with 5 beams and a temperature of 1.0 for candidate generation.[2] The pivot summary in `PivotRR` was decoded using beam search with 8 beams. For `Mistral 7B`, we used multinomial sampling with a temperature of 0.1. Further implementation details are provided in Appendix B.

5.3 Evaluation Metrics

Throughout this work, we emphasize the importance of evaluating MTXLS not only by comparing the generated summaries for each target language with their respective references, but also by evaluating the semantic coherence across different target languages. To evaluate the former, we present the ROUGE-2 scores (Lin, 2004) for each generated summary against its corresponding reference in the same target language. In addition, we calculate the BLASER 2.0 score (Communication et al., 2023) by treating the generated summary as the translation and the reference summary for the source language as the source text. This evaluation metric is justified due to mismatched articles in CrossSum, as explained in Section 5.1, which reduces the reliability of reference summaries in languages other than the source.

To assess semantic coherence across various target languages, we evaluate how well each generated summary translates the generated summaries for the remaining target languages. For this purpose, we use two quality estimation models for MT, namely CometKiwi (Rei et al., 2022) and BLASER 2.0. Here, for each target language, we use the generated summary as the translation and the summaries generated for all the other target languages as the source texts and then report the average across those languages.

[2] https://huggingface.co/docs/transformers/generation_strategies#beam-search-multinomial-sampling

Source	Method	ROUGE-2 (R)			BLASER 2.0 (R)			CometKiwi (C)			BLASER 2.0 (C)			T	#P
		en	zh	rest	en	zh	rest	en	zh	rest	en	zh	rest		
en	M2MS	**17.88**	**18.63**	**13.20**	3.52	3.04	3.25	59.28	61.00	60.31	3.48	3.26	3.61	0.52	582
	S&T	**17.88**	7.51	11.77	3.52	2.73	3.26	**85.00**	**79.24**	**86.40**	**4.67**	**3.87**	**4.79**	0.53	1,953
	Mistral 7B	6.52	3.18	4.57	2.45	2.13	2.31	69.77	65.95	71.09	3.09	3.24	3.16	4.64	7,241
	PivotRR (ours)	**17.88**	17.54	13.12	3.52	**3.09**	3.28	63.72	64.42	63.35	3.71	3.45	3.81	0.96	1,348
	NeutralRR (ours)	17.59	17.87	12.90	**3.53**	3.08	**3.29**	64.34	65.43	64.76	3.76	3.49	3.89	0.99	1,348
zh	M2MS	17.95	**24.13**	16.32	3.58	3.14	3.31	61.95	60.23	60.56	3.40	3.20	3.39	0.40	582
	S&T	13.51	**24.13**	12.11	3.48	3.14	3.25	**83.61**	**82.50**	**82.09**	**4.26**	**4.10**	**4.29**	0.52	1,953
	Mistral 7B	4.58	3.93	3.68	2.47	2.02	2.39	67.28	66.40	66.98	3.19	2.98	3.15	11.48	7,241
	PivotRR (ours)	18.32	**24.13**	16.36	3.60	3.14	**3.36**	64.73	62.99	61.90	3.54	3.37	3.54	0.89	1,348
	NeutralRR (ours)	**18.34**	23.72	**16.37**	**3.61**	**3.18**	3.35	66.94	63.74	63.23	3.63	3.43	3.62	0.90	1,348
rest	M2MS	**16.73**	**23.83**	**13.83**	3.48	3.07	3.23	60.50	60.33	61.13	3.55	3.15	3.54	0.56	582
	S&T	11.88	7.63	11.41	3.38	2.72	3.20	**85.63**	**80.38**	**85.67**	**4.71**	**3.88**	**4.75**	0.59	1,953
	PivotRR (ours)	16.32	23.56	13.66	3.50	**3.12**	3.25	63.63	62.04	63.28	3.72	3.33	3.73	0.98	1,348
	NeutralRR (ours)	16.48	23.01	13.75	**3.51**	**3.12**	**3.27**	65.37	63.30	64.62	3.83	3.39	3.82	1.02	1.348

Table 2: Results of evaluated methods in CrossSum for multi-target cross-lingual summarization using different languages as the source language. The language in each column is the target, with "rest" indicating the average for the remaining target languages. Metrics with (R) evaluate similarity to reference summaries, while those with (C) evaluate semantic coherence across languages. ROUGE-2 and CometKiwi range from 0 to 100, while BLASER 2.0 ranges from 1 to 5 (higher values are better). Best results are bold, second best results are underlined. Columns T and #P indicate the average computation time per generated summary in seconds and the number of model parameters in millions, respectively.

5.4 Main Results

In this section, we present results on MTXLS considering all the seven languages mentioned in Section 5.1 as targets. To perform this task, we took each of the seven languages as the source in turn and discarded the clusters that lacked a document in the source language. Then, we iterated through the remaining clusters taking the document in the source language as the input for summarization and we generated summaries for all the languages in the cluster, including the source language, using each of the methods mentioned in Section 5.2.

The results are in Table 2 and are presented per language pair. Due to space limitations, we present detailed results only for English (en) and Chinese (zh), and show the averages for the remaining source and target languages (rest). An extended version of this table, including detailed results for more languages, confidence intervals, and the accuracy of each approach on following the target language is shown in Appendix D.1. When the source and target languages are the same, `S&T` and `PivotRR` reduce to `M2MS` because we use the source language as the pivot. Consequently, the results of these three methods for ROUGE-2 and BLASER 2.0 (R) coincide for en→en and zh→zh.

We begin by discussing the results of `Mistral 7B`, as these deserve special attention. Interestingly, the model always performs worst in terms of similarity to the reference summaries (ROUGE-2 and BLASER 2.0 (R)), even though it was instructed that the articles were obtained from the BBC and that the summaries should follow the BBC style (see Appendix C). Regarding the coherence across target languages, we observe that the model has a very decent performance, as illustrated in Table 1, ranking second in CometKiwi scores, only surpassed by `S&T`. However, the model often failed to produce the output in the requested format, in which case we had to repeat the request, or did not produce text in the specified target language (see Appendix D.1). For these reasons, we did not extend its evaluation to other source languages beyond English and Chinese.

The method `M2MS` conducts cross-lingual summarization for each target language independently, disregarding semantic coherence across languages. Consequently, it consistently achieves the highest ROUGE-2 scores but ranks lowest in coherence metrics (CometKiwi and BLASER 2.0 (C)). Conversely, `S&T` ensures the best semantic coherence across target languages by directly translating the source language summary for each target language. However, this often results in significant degradation in similarity with the references for each target language, as measured by ROUGE-2, and, in many cases, even diminishes similarity to the reference summary for the source language, as measured by

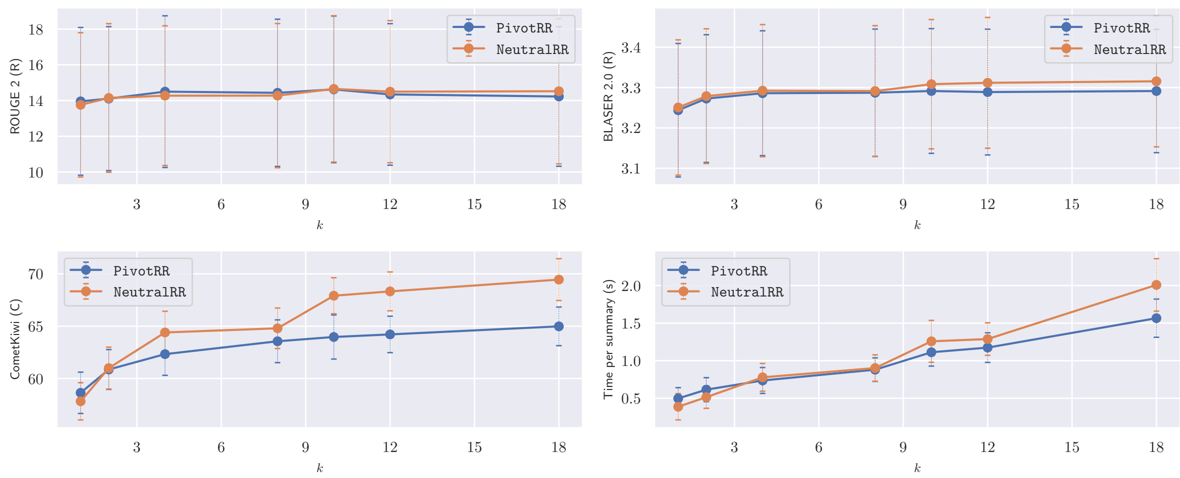


Figure 2: Results of `PivotRR` and `NeutralRR` as a function of the number of candidates per target language for re-ranking (k in Algorithm 1). Error bars indicate the standard deviations across the target languages.

BLASER 2.0 (R). This indicates that the MT model introduced errors compromising summary quality.

Our approaches (`PivotRR` and `NeutralRR`) do not significantly degrade ROUGE-2 scores compared to `M2MS` and notably achieve the highest similarity to the the reference summary for the source language. As expected, our methods also significantly improve semantic coherence across different target languages compared to `M2MS`. `NeutralRR` performs comparably to `PivotRR` in terms of similarity to the reference summaries, and consistently outperforms it in terms of semantic coherence across target languages. This was expected because `NeutralRR` treats all languages equally and aims for a set of summaries with high similarity. Conversely, `PivotRR` utilizes a fixed pivot summary and seeks candidates in each target language that closely resemble the pivot.

5.5 Effect of Varying the Number of Candidates

In this experiment, we investigate how the performance of our methods changes as we vary the number of candidates for re-ranking, using English as the source language. To vary the number of candidates generated by beam search multinomial sampling, we kept the number of beams per output sequence constant and equal to 5 and varied the number of output sequences. The results are in Figure 2, where we show the averages and standard deviations across the seven target languages.

Interestingly, increasing the number of candidates does not affect the similarity between the selected summaries and their respective references, as evaluated by ROUGE-2. In addition, it has a positive effect on the similarity between the selected summaries and the reference in the source language, as measured by BLASER 2.0 (R). We justify this observation by the hypothesis that a set of summaries with high similarity can serve as a reliable indicator of summary quality, since it is unlikely that the model generates the same false information in multiple languages. This was illustrated in the example in Table 1 for `NeutralRR`. Finally, as more candidates are considered, computation time increases, yet so does the similarity of selected summaries, as evaluated by CometKiwi. Notably, this similarity increase is more significant for `NeutralRR`, which is not limited by maximizing similarity to a fixed pivot summary.

6 Conclusion

This work introduces multi-target cross-lingual summarization to address the challenge of achieving coherent summaries across multiple target languages. We propose two re-ranking approaches tailored to this task, which improve semantic coherence across languages compared to conventional beam search decoding, while still preserving similarity to the reference summaries. In particular, one of these methods eliminates the need for a pivot language, thus treating all languages equally and eliminating potential biases arising from pivot language selection. Furthermore, we extended the evaluation framework for cross-lingual summarization by including the assessment of semantic coherence across different target languages.

Limitations

While we believe that our approach has merit, it is equally important to recognize its inherent limitations. First, we anticipate that as large language models continue to improve and become fluent in more languages, instructing the model to produce summaries with identical information for all target languages will eventually be sufficient to satisfy our semantic coherence constraint. Second, the success of our re-ranking approaches depends on the quality of the sampled candidates. If all candidates are of low quality, or if they have poor semantic coherence across target languages, our approaches will inevitably fail. Investigating computationally efficient ways to incorporate the semantic coherence constraint directly at decoding time is an interesting research direction. Finally, our method introduces increased computational complexity compared to the usual beam search decoding.

References

Mikel Artetxe and Holger Schwenk. 2019. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610.

Yu Bai, Yang Gao, and Heyan Huang. 2021. Cross-lingual abstractive summarization with limited parallel resources. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6910–6924, Online. Association for Computational Linguistics.

Abhik Bhattacharjee, Tahmid Hasan, Wasi Uddin Ahmad, Yuan-Fang Li, Yong-Bin Kang, and Rifat Shahriyar. 2023. CrossSum: Beyond English-centric cross-lingual summarization for 1,500+ language pairs. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2541–2564, Toronto, Canada. Association for Computational Linguistics.

Yue Cao, Hui Liu, and Xiaojun Wan. 2020. Jointly learning to align and summarize for neural cross-lingual summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6220–6231, Online. Association for Computational Linguistics.

Mingda Chen, Paul-Ambroise Duquenne, Pierre Andrews, Justine Kao, Alexandre Mourachko, Holger Schwenk, and Marta R. Costa-jussà. 2023a. BLASER: A text-free speech-to-speech translation evaluation metric. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9064–9079, Toronto, Canada. Association for Computational Linguistics.

Yulong Chen, Huajian Zhang, Yijie Zhou, Xuefeng Bai, Yueguan Wang, Ming Zhong, Jianhao Yan, Yafu Li, Judy Li, Xianchao Zhu, and Yue Zhang. 2023b. Revisiting cross-lingual summarization: A corpus-based study and a new benchmark with improved annotation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9332–9351, Toronto, Canada. Association for Computational Linguistics.

Seamless Communication, Loïc Barrault, Yu-An Chung, Mariano Cora Meglioli, David Dale, Ning Dong, Paul-Ambroise Duquenne, Hady Elsahar, Hongyu Gong, Kevin Heffernan, John Hoffman, Christopher Klaiber, Pengwei Li, Daniel Licht, Jean Maillard, Alice Rakotoarison, Kaushik Ram Sadagopan, Guillaume Wenzek, Ethan Ye, Bapi Akula, Peng-Jen Chen, Naji El Hachem, Brian Ellis, Gabriel Mejia Gonzalez, Justin Haaheim, Prangthip Hansanti, Russ Howes, Bernie Huang, Min-Jae Hwang, Hirofumi Inaguma, Somya Jain, Elahe Kalbassi, Amanda Kallet, Ilia Kulikov, Janice Lam, Daniel Li, Xutai Ma, Ruslan Mavlyutov, Benjamin Peloquin, Mohamed Ramadan, Abinesh Ramakrishnan, Anna Sun, Kevin Tran, Tuan Tran, Igor Tufanov, Vish Vogeti, Carleigh Wood, Yilin Yang, Bokai Yu, Pierre Andrews, Can Balioglu, Marta R. Costa-jussà, Onur Celebi, Maha Elbayad, Cynthia Gao, Francisco Guzmán, Justine Kao, Ann Lee, Alexandre Mourachko, Juan Pino, Sravya Popuri, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, Paden Tomasello, Changhan Wang, Jeff Wang, and Skyler Wang. 2023. SeamlessM4T: Massively multilingual & multimodal machine translation. *arXiv preprint arXiv:2308.11596*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Marta R Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Paul-Ambroise Duquenne, Holger Schwenk, and Benoît Sagot. 2023. SONAR: Sentence-level multimodal and language-agnostic representations. *arXiv preprint arXiv:2308.11466*.

Marzieh Fadaee, Arianna Bisazza, and Christof Monz. 2018. Examining the tip of the iceberg: A data set for idiom translation. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Corinne Feremans, Martine Labbé, and Gilbert Laporte. 2003. Generalized network design problems. *European Journal of Operational Research*, 148(1):1–13.

Tahmid Hasan, Abhik Bhattacharjee, Md. Saiful Islam, Kazi Mubasshir, Yuan-Fang Li, Yong-Bin Kang, M. Sohel Rahman, and Rifat Shahriyar. 2021. XL-sum: Large-scale multilingual abstractive summarization for 44 languages. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4693–4703, Online. Association for Computational Linguistics.

Fantine Huot, Joshua Maynez, Chris Alberti, Reinald Kim Amplayo, Priyanka Agrawal, Constanza Fierro, Shashi Narayan, and Mirella Lapata. 2024. μPLAN: Summarizing using a content plan as cross-lingual bridge. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2146–2163, St. Julian's, Malta. Association for Computational Linguistics.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hérve Jégou, and Tomas Mikolov. 2016a. Fasttext.zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651*.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016b. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*.

Fabio Kepler, Jonay Trénous, Marcos Treviso, Miguel Vera, António Góis, M. Amin Farajian, António V. Lopes, and André F. T. Martins. 2019a. Unbabel's participation in the WMT19 translation quality estimation shared task. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pages 78–84, Florence, Italy. Association for Computational Linguistics.

Fabio Kepler, Jonay Trénous, Marcos Treviso, Miguel Vera, and André F. T. Martins. 2019b. OpenKiwi: An open source framework for quality estimation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 117–122, Florence, Italy. Association for Computational Linguistics.

Faisal Ladhak, Esin Durmus, Claire Cardie, and Kathleen McKeown. 2020. WikiLingua: A new benchmark dataset for cross-lingual abstractive summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4034–4048, Online. Association for Computational Linguistics.

Dongjun Lee. 2020. Two-phase cross-lingual language model fine-tuning for machine translation quality estimation. In *Proceedings of the Fifth Conference on Machine Translation*, pages 1024–1028, Online. Association for Computational Linguistics.

Yunlong Liang, Fandong Meng, Chulun Zhou, Jinan Xu, Yufeng Chen, Jinsong Su, and Jie Zhou. 2022. A variational hierarchical model for neural cross-lingual summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2088–2099, Dublin, Ireland. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pretraining for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, Online. Association for Computational Linguistics.

Laura Perez-Beltrachini and Mirella Lapata. 2021. Models and datasets for cross-lingual summarisation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9408–9423, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training. *Technical report, OpenAI*.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Ricardo Rei, Marcos Treviso, Nuno M. Guerreiro, Chrysoula Zerva, Ana C Farinha, Christine Maroti, José G. C. de Souza, Taisiya Glushkova, Duarte Alves, Luisa Coheur, Alon Lavie, and André F. T. Martins. 2022. CometKiwi: IST-unbabel 2022 submission for the quality estimation shared task. In

Proceedings of the Seventh Conference on Machine Translation (WMT), pages 634–645, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Lucia Specia, Carolina Scarton, and Gustavo Henrique Paetzold. 2018. *Quality estimation for machine translation*, volume 11. Springer.

Lucia Specia, Kashif Shah, Jose G.C. de Souza, and Trevor Cohn. 2013. QuEst - a translation quality estimation framework. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 79–84, Sofia, Bulgaria. Association for Computational Linguistics.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2021. Multilingual translation from denoising pre-training. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3450–3466, Online. Association for Computational Linguistics.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, Pouya Tafti, Léonard Hussenot, Aakanksha Chowdhery, Adam Roberts, Aditya Barua, Alex Botev, Alex Castro-Ros, Ambrose Slone, Amélie Héliou, Andrea Tacchetti, Anna Bulanova, Antonia Paterson, Beth Tsai, Bobak Shahriari, Charline Le Lan, Christopher A. Choquette-Choo, Clément Crepy, Daniel Cer, Daphne Ippolito, David Reid, Elena Buchatskaya, Eric Ni, Eric Noland, Geng Yan, George Tucker, George-Christian Muraru, Grigory Rozhdestvenskiy, Henryk Michalewski, Ian Tenney, Ivan Grishchenko, Jacob Austin, James Keeling, Jane Labanowski, Jean-Baptiste Lespiau, Jeff Stanway, Jenny Brennan, Jeremy Chen, Johan Ferret, Justin Chiu, Justin Mao-Jones, Katherine Lee, Kathy Yu, Katie Millican, Lars Lowe Sjoesund, Lisa Lee, Lucas Dixon, Machel Reid, Maciej Mikuła, Mateo Wirth, Michael Sharman, Nikolai Chinaev, Nithum Thain, Olivier Bachem, Oscar Chang, Oscar Wahltinez, Paige Bailey, Paul Michel, Petko Yotov, Pier Giuseppe Sessa, Rahma Chaabouni, Ramona Comanescu, Reena Jana, Rohan Anil, Ross McIlroy, Ruibo Liu, Ryan Mullins, Samuel L Smith, Sebastian Borgeaud, Sertan Girgin, Sholto Douglas, Shree Pandya, Siamak Shakeri, Soham De, Ted Klimenko, Tom Hennigan, Vlad Feinberg, Wojciech Stokowiec, Yu hui Chen, Zafarali Ahmed, Zhitao Gong, Tris Warkentin, Ludovic Peran, Minh Giang, Clément Farabet, Oriol Vinyals, Jeff Dean, Koray Kavukcuoglu, Demis Hassabis, Zoubin Ghahramani, Douglas Eck, Joelle Barral, Fernando Pereira, Eli Collins, Armand Joulin, Noah Fiedel, Evan Senter, Alek Andreev, and Kathleen Kenealy. 2024. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Ashwin Vijayakumar, Michael Cogswell, Ramprasaath Selvaraju, Qing Sun, Stefan Lee, David Crandall, and Dhruv Batra. 2018. Diverse beam search for improved description of complex scenes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Jiaan Wang, Yunlong Liang, Fandong Meng, Beiqi Zou, Zhixu Li, Jianfeng Qu, and Jie Zhou. 2023a. Zero-shot cross-lingual summarization via large language models. In *Proceedings of the 4th New Frontiers in Summarization Workshop*, pages 12–23, Singapore. Association for Computational Linguistics.

Jiaan Wang, Fandong Meng, Duo Zheng, Yunlong Liang, Zhixu Li, Jianfeng Qu, and Jie Zhou. 2022. A survey on cross-lingual summarization. *Transactions of the Association for Computational Linguistics*, 10:1304–1323.

Jiaan Wang, Fandong Meng, Duo Zheng, Yunlong Liang, Zhixu Li, Jianfeng Qu, and Jie Zhou. 2023b. Towards unifying multi-lingual and cross-lingual summarization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15127–15143, Toronto, Canada. Association for Computational Linguistics.

Jiayi Wang, Kai Fan, Bo Li, Fengming Zhou, Boxing Chen, Yangbin Shi, and Luo Si. 2018. Alibaba submission for WMT18 quality estimation task. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 809–815, Belgium, Brussels. Association for Computational Linguistics.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

A Dataset Clustering and Analysis

As mentioned in Section 5.1, the original CrossSum dataset presents documents in one language paired with summaries in another language, a format that does not serve our multi-target setting. Therefore, we clustered the dataset to obtain clusters of multilingual document-summary pairs about the same story. To achieve this, we aggregated all documents across the mentioned languages and constructed an undirected graph representing their pairwise connections. In this graph, two documents in different languages are connected if they are paired in CrossSum. We then built clusters by extracting all maximal cliques from this graph and we discarded all singleton cliques. Consequently, each maximal clique is a cluster of up to seven multilingual documents pertaining to the same story, where each document is accompanied by a summary in its respective language.

This clustering procedure was applied separately to the CrossSum validation and test splits. The resulting validation set consisted of 4,525 clusters and 10,479 documents, while the test set consisted of 4,560 clusters and 10,535 documents. Table 3 provides a breakdown of cluster sizes in the test set, as well as the distribution of documents for each language and cluster size. Notably, none of the clusters in the test set are complete, indicating that no cluster includes a document for all seven languages considered. In addition, we conducted an analysis of the co-occurrence of different language pairs within the clusters to verify whether a robust evaluation of cross-lingual summarization was possible across all language directions. Figure 3 illustrates the distribution of clusters containing examples of each language pair. While certain language pairs have higher representation than others, it is noteworthy that even the least represented pair (fr, zh) is found in 35 clusters, indicating a diverse linguistic coverage across the dataset.

Since one of our goals is to assess the semantic coherence of the generated summaries in different target languages, it is crucial to evaluate the coherence of reference summary clusters in this regard. This evaluation helps to determine the level of coherence that can be achieved in the generated summaries without degrading similarity to the reference summaries. To achieve this, we computed BLASER 2.0 and CometKiwi scores between reference summaries within the same cluster for each language pair. The results are shown in Figure 4.

Language	Cluster Size						
	2	3	4	5	6	7	All
ar	1,022	455	153	34	6	0	1,670
en	1,780	598	176	37	7	0	2,598
es	1,271	367	130	31	7	0	1,806
fr	224	84	46	14	5	0	373
pt	1,027	280	118	33	6	0	1,464
ru	1,077	482	140	38	5	0	1,742
zh	531	224	93	28	6	0	882
All	3,466	830	214	43	7	0	4,560

Table 3: Number of clusters in the test set containing a document of each language, organized by cluster size.

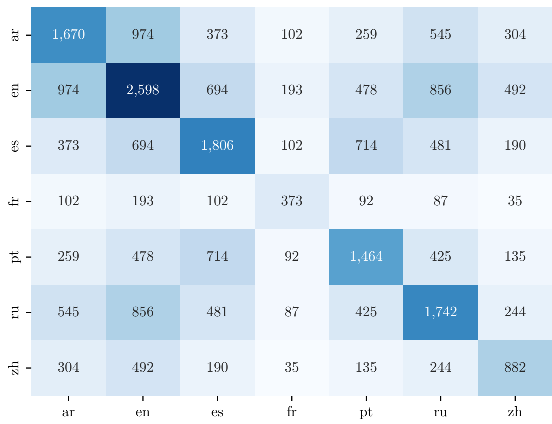


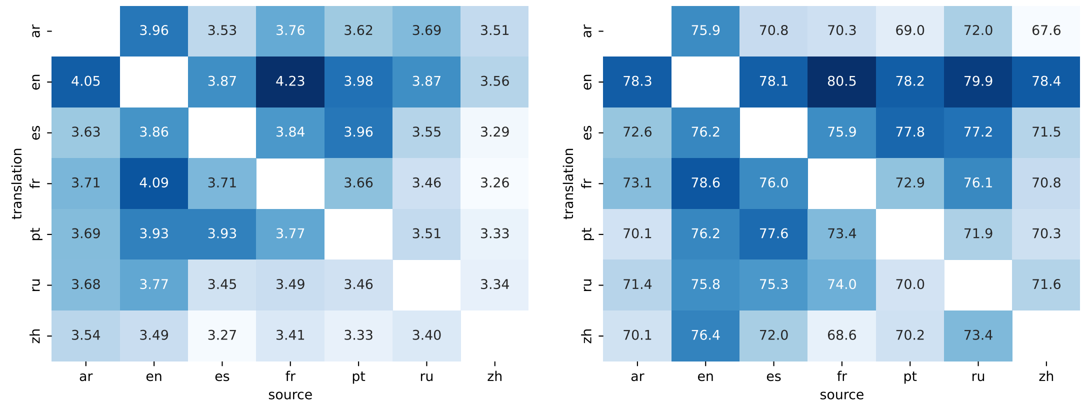
Figure 3: Number of clusters in the test set containing documents of each language pair.

It is important to note that the matrices are non-symmetric due to the nature of BLASER 2.0 and CometKiwi metrics. Firstly, we note a significant agreement between the two metrics, as anticipated. Additionally, coherence tends to be higher among languages using the Latin script. However, for most language pairs, coherence remains above 3.40 BLASER 2.0 points and 70.0 CometKiwi points. This suggests room for improvement compared to the results outlined in Table 2.

B Implementation Details

To represent the cross-lingual summarization distribution $q(\boldsymbol{y}_t \mid \boldsymbol{x}_o, t)$, we use an mT5 model (Xue et al., 2021) for all the methods except `Mistral 7B`. mT5 allows us to perform summarization across all language directions by conditioning the decoder on a unique start-of-sequence token that specifies the intended target language.

We used the publicly available SONAR checkpoint `text_sonar_basic_encoder` to implement ϕ, the mT5 checkpoint `csebuetnlp/mT5_m2m_crossSum_enhanced`, which was fine-tuned in the CrossSum



(a) (b)

Figure 4: Average BLASER 2.0 (a) and CometKiwi (b) scores between reference summaries within the same cluster for each language pair in the test set.

dataset, and the Mistral 7B checkpoint `mistralai/Mistral-7B-Instruct-v0.2`. All of these checkpoints are available at the Hugging Face model hub.[3]

The optimal beam size and sampling temperature for beam search multinomial sampling were determined through a grid search. We explored beam sizes of 1, 3, and 5, and temperatures of 0.1, 0.3, 0.5, 1.0, 1.5, and 2.0 in order to maximize the ROUGE-2 score on the validation set of English-to-all summarization. We also tried with other decoding strategies, namely (single-beam) multinomial sampling and diverse beam search (Vijayakumar et al., 2018), but these degraded ROUGE scores considerably. The number of random language permutations (m in Algorithm 1) used by `NeutralRR` was set to 6 when the number of target languages was at least three and was set to 2 if there were only two target languages, since there are only two possible permutations of two languages.

Regarding the evaluation metrics, we used the multilingual implementation of ROUGE by Hasan et al. (2021).[4] For CometKiwi and BLASER 2.0, we used the `Unbabel/wmt22-cometkiwi-da` and `blaser_2_0_qe` checkpoints, respectively.

All experiments were run on an 80-core Intel Xeon Gold 5218R CPU @ 2.10GHz with 800GB of RAM and an NVIDIA A100 GPU with 80GB of memory.

[3] https://huggingface.co/models

[4] https://github.com/csebuetnlp/xl-sum/tree/master/multilingual_rouge_scoring

C LLM Prompt

The following prompt was used on the experiments with `Mistral 7B`:

```
For the <source_lang> news article from BBC written below, provide a summary in <target_lang_1>, a summary in <target_lang_2>, ... and a summary in <target_lang_N>. All summaries should be one or two sentences long and follow the style of BBC. All summaries must contain the same information. Present the answer in the format of a JSON object where the keys are the language codes and the values are the summaries.
Text:
<source_document>
```

D Further Experimental Results

D.1 Main Results Extended

An extended version of the results presented in Table 2 is shown in Tables 4 and 5. In addition to English and Chinese, we also show results for Spanish and French. Spanish is the second most represented language in the dataset, surpassed only by English, while French is the least represented (see Table 3). All the results are accompanied by 95% bootstrap confidence intervals with 1,000 resamples. Apart from the metrics mentioned in Section 5.3, we also include the target language accuracy in Table 4. This metric corresponds to the percentage of times a method generated text in the specified target language, and is calculated by comparing the specified language with the dominant language identified in the generated text by

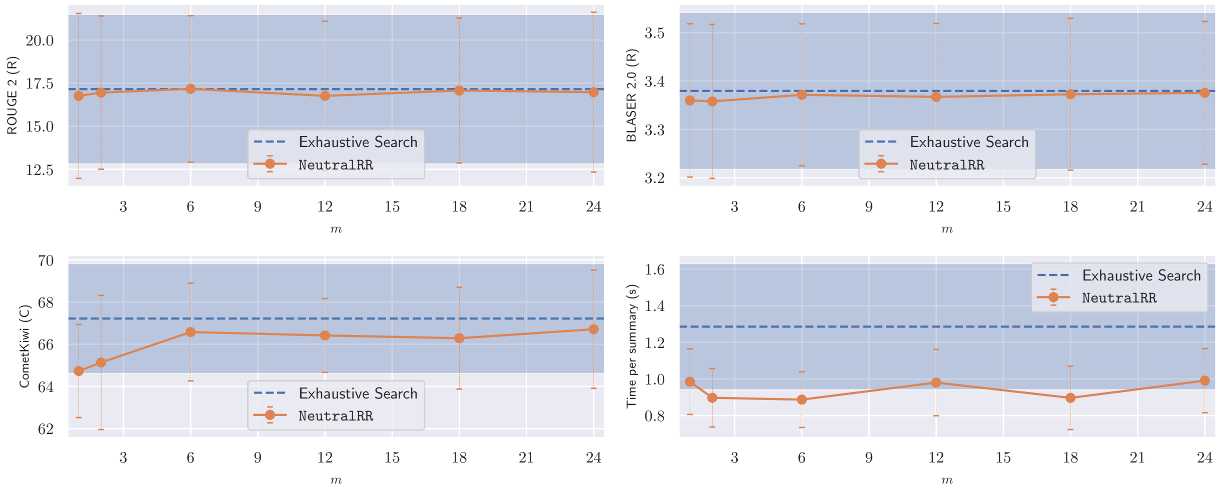


Figure 5: Effect of varying the number of language permutations (m in Algorithm 1) on the results of `NeutralRR`. The results of doing an exhaustive search for the most coherent set are also shown for comparison. The error bars and the shaded area indicate the standard deviations across the target languages.

the fastText model (Joulin et al., 2016a,b). We observe that the mT5-based methods generate text in the correct target language in the vast majority (if not all) of the cases. `Mistral 7B` sometimes struggles to generate text in the correct target language, especially for Arabic.

D.2 Effect of the Heuristic Search

In this experiment, we investigate the effect of the number of language permutations (m in Algorithm 1) on the performance of `NeutralRR`. In this experiment, we always use English as the source language and only consider clusters of documents with 4 languages, allowing up to 24 language permutations. The number of candidate summaries per language is kept fixed at 8. For this cluster size and number of candidates, maximizing φ (equation (6)) directly with an exhaustive search is feasible since there are only $8^4 = 4096$ possible sets of summaries. Therefore, we also compare the results of our approach with the exhaustive search. The results are shown in Figure 5.

The first observation is that changing m or performing an exhaustive search does not significantly affect the similarity to the reference summaries. Changing m also has no significant effect on the computation time, which is natural since the time required by the dynamic programming optimization is much smaller than the decoding time of the summarization model. However, an exhaustive search obviously increases the computation time, and the difference would only become larger for larger cluster sizes or more candidate summaries per language. Regarding the semantic coherence of the resulting set of summaries, an exhaustive search yields the best results as expected, but they are only slightly better than our heuristic search with a sufficiently large number of language permutations.

Source	Method	R2 (R) en	R2 (R) es	R2 (R) fr	R2 (R) zh	R2 (R) rest	BLASER 2.0 (R) en	BLASER 2.0 (R) es	BLASER 2.0 (R) fr	BLASER 2.0 (R) zh	BLASER 2.0 (R) rest	Target Lang. Acc. en	Target Lang. Acc. es	Target Lang. Acc. fr	Target Lang. Acc. zh	Target Lang. Acc. rest
en	M2MS	17.88 ± 0.59	13.91 ± 0.97	21.55 ± 3.14	18.63 ± 1.82	10.19 ± 0.83	3.52 ± 0.02	3.03 ± 0.03	3.31 ± 0.08	3.04 ± 0.05	3.31 ± 0.03	100.0	99.7	100.0	97.8	99.6
	S&T	17.88 ± 0.59	11.91 ± 0.84	18.98 ± 2.52	7.51 ± 0.78	9.33 ± 0.77	3.52 ± 0.02	3.07 ± 0.04	3.30 ± 0.07	2.73 ± 0.04	3.31 ± 0.04	100.0	99.9	100.0	99.8	99.9
	Mistral 7B	6.52 ± 0.28	6.68 ± 0.58	7.86 ± 1.57	3.18 ± 0.37	2.78 ± 0.29	2.45 ± 0.02	2.17 ± 0.03	2.46 ± 0.08	2.13 ± 0.03	2.31 ± 0.03	99.7	99.8	100.0	99.2	82.2
	PivotRR	17.88 ± 0.59	13.58 ± 0.96	20.77 ± 3.08	17.54 ± 1.69	10.42 ± 0.85	3.52 ± 0.02	3.06 ± 0.03	3.32 ± 0.08	3.09 ± 0.04	3.34 ± 0.03	100.0	99.4	100.0	98.8	99.6
	NeutralRR	17.59 ± 0.58	13.70 ± 0.94	20.08 ± 3.24	17.87 ± 1.75	10.23 ± 0.86	3.53 ± 0.02	3.06 ± 0.03	3.34 ± 0.08	3.08 ± 0.04	3.34 ± 0.03	100.0	99.7	99.5	99.4	99.5
es	M2MS	15.88 ± 1.26	14.37 ± 0.66	20.48 ± 4.77	20.01 ± 3.38	9.12 ± 1.04	3.40 ± 0.04	2.99 ± 0.02	3.26 ± 0.15	2.92 ± 0.10	3.20 ± 0.04	99.7	99.8	100.0	97.9	99.9
	S&T	10.71 ± 0.80	14.37 ± 0.66	14.28 ± 2.91	7.99 ± 1.31	8.01 ± 0.85	3.29 ± 0.04	2.99 ± 0.02	3.16 ± 0.12	2.60 ± 0.07	3.15 ± 0.04	100.0	99.8	100.0	99.5	99.8
	PivotRR	15.35 ± 1.20	14.37 ± 0.66	18.70 ± 4.55	19.72 ± 3.03	9.37 ± 1.02	3.42 ± 0.04	2.99 ± 0.02	3.27 ± 0.14	2.95 ± 0.09	3.23 ± 0.04	99.7	99.8	99.0	99.5	99.7
	NeutralRR	15.59 ± 1.25	14.68 ± 0.65	20.68 ± 4.61	19.69 ± 3.07	9.54 ± 1.07	3.43 ± 0.04	3.05 ± 0.02	3.33 ± 0.14	2.98 ± 0.08	3.23 ± 0.04	99.7	99.9	99.0	100.0	99.7
fr	M2MS	21.25 ± 3.19	15.49 ± 3.49	23.78 ± 2.46	38.04 ± 10.21	13.38 ± 3.65	3.57 ± 0.08	3.05 ± 0.10	3.33 ± 0.07	3.24 ± 0.24	3.29 ± 0.11	100.0	100.0	100.0	100.0	100.0
	S&T	16.14 ± 2.08	13.19 ± 2.56	23.78 ± 2.46	9.87 ± 3.11	11.01 ± 2.54	3.43 ± 0.08	3.11 ± 0.11	3.33 ± 0.07	2.78 ± 0.18	3.28 ± 0.10	100.0	100.0	100.0	100.0	100.0
	PivotRR	20.11 ± 2.98	15.31 ± 3.00	23.78 ± 2.46	38.21 ± 9.44	13.16 ± 3.33	3.54 ± 0.08	3.10 ± 0.11	3.33 ± 0.07	3.35 ± 0.21	3.34 ± 0.11	100.0	100.0	100.0	100.0	100.0
	NeutralRR	20.95 ± 2.79	14.42 ± 3.13	22.76 ± 2.41	36.28 ± 10.27	13.18 ± 3.29	3.59 ± 0.08	3.08 ± 0.11	3.33 ± 0.06	3.26 ± 0.21	3.36 ± 0.11	100.0	100.0	100.0	100.0	100.0
zh	M2MS	17.95 ± 1.52	15.54 ± 2.24	30.91 ± 10.19	24.13 ± 1.54	11.72 ± 1.77	3.58 ± 0.05	3.04 ± 0.07	3.50 ± 0.27	3.14 ± 0.04	3.34 ± 0.06	98.0	100.0	100.0	99.8	99.6
	S&T	13.51 ± 1.17	11.36 ± 1.42	21.79 ± 5.13	24.13 ± 1.54	9.13 ± 1.41	3.48 ± 0.04	3.00 ± 0.08	3.32 ± 0.16	3.14 ± 0.04	3.31 ± 0.07	100.0	99.5	100.0	99.8	99.3
	Mistral 7B	4.58 ± 0.43	4.39 ± 0.65	8.22 ± 3.00	3.93 ± 0.30	1.92 ± 0.45	2.47 ± 0.04	2.26 ± 0.05	2.44 ± 0.15	2.02 ± 0.03	2.43 ± 0.05	99.8	91.6	100.0	97.4	63.1
	PivotRR	18.32 ± 1.46	15.73 ± 2.09	30.67 ± 8.23	24.13 ± 1.54	11.80 ± 1.65	3.60 ± 0.05	3.09 ± 0.08	3.51 ± 0.18	3.14 ± 0.04	3.39 ± 0.06	98.4	100.0	100.0	99.8	99.9
	NeutralRR	18.34 ± 1.46	15.63 ± 2.11	29.25 ± 7.83	23.72 ± 1.47	12.32 ± 1.73	3.61 ± 0.05	3.09 ± 0.07	3.50 ± 0.20	3.18 ± 0.04	3.39 ± 0.06	98.6	100.0	100.0	99.9	99.9
rest	M2MS	15.50 ± 1.15	12.82 ± 1.10	22.16 ± 5.59	20.36 ± 3.16	11.02 ± 0.99	3.48 ± 0.04	2.99 ± 0.04	3.29 ± 0.14	3.06 ± 0.08	3.30 ± 0.04	99.9	99.4	99.7	99.6	99.9
	S&T	10.84 ± 0.81	10.95 ± 0.90	14.69 ± 2.68	6.76 ± 1.15	9.52 ± 0.80	3.38 ± 0.03	3.01 ± 0.04	3.19 ± 0.10	2.74 ± 0.06	3.28 ± 0.04	100.0	99.8	100.0	100.0	99.8
	PivotRR	15.38 ± 1.09	12.93 ± 1.08	20.91 ± 4.98	19.96 ± 3.08	11.13 ± 0.98	3.51 ± 0.04	3.03 ± 0.04	3.32 ± 0.13	3.10 ± 0.08	3.32 ± 0.04	99.9	99.4	99.6	99.8	99.9
	NeutralRR	15.29 ± 1.13	13.04 ± 1.09	21.14 ± 4.78	19.70 ± 2.94	11.18 ± 0.99	3.50 ± 0.04	3.03 ± 0.04	3.34 ± 0.13	3.11 ± 0.08	3.34 ± 0.04	99.8	99.5	100.0	99.6	100.0

Table 4: Extended results of multi-target cross-lingual summarization in CrossSum for the metrics evaluating similarity to the reference summaries. Target language accuracies are also shown.

Source	Method	CometKiwi (C)					BLASER 2.0 (C)				
		en	es	fr	zh	rest	en	es	fr	zh	rest
en	M2MS	59.28 ± 0.56	61.79 ± 1.16	58.21 ± 2.39	61.00 ± 1.40	60.52 ± 1.06	3.48 ± 0.02	3.52 ± 0.04	3.63 ± 0.08	3.26 ± 0.01	3.6 ± 0.01
	S&T	85.00 ± 0.20	87.33 ± 0.33	87.93 ± 0.39	79.24 ± 1.15	85.58 ± 0.39	4.67 ± 0.01	4.80 ± 0.02	4.81 ± 0.03	3.87 ± 0.05	4.78 ± 0.03
	Mistral 7B	69.77 ± 0.66	79.40 ± 0.75	77.97 ± 2.01	65.95 ± 1.04	66.04 ± 0.96	3.09 ± 0.04	3.44 ± 0.08	3.19 ± 0.15	3.24 ± 0.07	3.06 ± 0.06
	PivotRR	63.72 ± 0.52	64.79 ± 1.15	61.51 ± 2.48	64.42 ± 1.28	63.49 ± 1.03	3.71 ± 0.02	3.73 ± 0.03	3.83 ± 0.06	3.45 ± 0.04	3.83 ± 0.03
	NeutralRR	64.34 ± 0.54	66.12 ± 1.21	63.20 ± 2.33	65.43 ± 1.25	64.83 ± 1.04	3.76 ± 0.02	3.81 ± 0.03	3.91 ± 0.05	3.49 ± 0.04	3.91 ± 0.03
es	M2MS	61.25 ± 1.06	61.61 ± 0.72	60.40 ± 3.15	59.23 ± 2.23	61.65 ± 1.39	3.52 ± 0.04	3.44 ± 0.03	3.54 ± 0.11	3.07 ± 0.10	3.5 ± 0.01
	S&T	86.31 ± 0.33	86.36 ± 0.29	86.81 ± 1.05	79.53 ± 1.72	85.61 ± 0.59	4.69 ± 0.02	4.72 ± 0.02	4.78 ± 0.05	3.82 ± 0.07	4.76 ± 0.03
	PivotRR	64.42 ± 1.08	65.18 ± 0.74	61.44 ± 3.10	60.56 ± 2.12	64.26 ± 1.39	3.70 ± 0.04	3.68 ± 0.03	3.65 ± 0.10	3.24 ± 0.07	3.76 ± 0.04
	NeutralRR	65.82 ± 1.07	65.59 ± 0.75	63.79 ± 3.25	61.90 ± 2.23	64.92 ± 1.39	3.80 ± 0.03	3.78 ± 0.02	3.78 ± 0.09	3.36 ± 0.06	3.84 ± 0.04
fr	M2MS	60.63 ± 2.11	63.81 ± 2.92	61.62 ± 1.63	59.83 ± 5.47	59.98 ± 3.05	3.57 ± 0.07	3.47 ± 0.09	3.60 ± 0.05	3.04 ± 0.20	3.5 ± 0.10
	S&T	86.70 ± 0.38	87.27 ± 0.81	86.51 ± 0.57	81.17 ± 3.12	85.42 ± 1.22	4.68 ± 0.03	4.70 ± 0.05	4.76 ± 0.03	3.72 ± 0.18	4.66 ± 0.06
	PivotRR	64.67 ± 2.02	66.29 ± 2.93	64.62 ± 1.53	62.68 ± 4.89	62.06 ± 3.14	3.74 ± 0.06	3.67 ± 0.09	3.81 ± 0.05	3.27 ± 0.14	3.66 ± 0.09
	NeutralRR	66.34 ± 1.99	66.42 ± 3.29	65.32 ± 1.60	63.04 ± 5.42	64.02 ± 2.98	3.86 ± 0.05	3.76 ± 0.08	3.89 ± 0.05	3.25 ± 0.14	3.75 ± 0.07
zh	M2MS	61.95 ± 1.16	59.45 ± 2.13	61.08 ± 5.42	60.23 ± 1.05	60.76 ± 2.04	3.40 ± 0.04	3.21 ± 0.07	3.41 ± 0.16	3.20 ± 0.01	3.4 ± 0.10
	S&T	83.61 ± 0.53	82.72 ± 1.54	83.48 ± 3.32	82.50 ± 0.57	81.42 ± 1.40	4.26 ± 0.04	4.22 ± 0.06	4.30 ± 0.13	4.10 ± 0.02	4.31 ± 0.06
	Mistral 7B	67.28 ± 1.27	68.27 ± 1.66	70.39 ± 4.04	66.40 ± 0.86	65.42 ± 1.67	3.19 ± 0.06	3.07 ± 0.09	3.44 ± 0.23	2.98 ± 0.05	3.08 ± 0.09
	PivotRR	64.73 ± 1.13	60.92 ± 2.15	61.59 ± 5.66	62.99 ± 1.01	62.33 ± 2.04	3.54 ± 0.04	3.37 ± 0.06	3.56 ± 0.15	3.37 ± 0.03	3.60 ± 0.05
	NeutralRR	66.94 ± 0.96	62.77 ± 2.20	62.83 ± 6.14	63.74 ± 1.03	63.51 ± 1.88	3.63 ± 0.03	3.47 ± 0.06	3.64 ± 0.14	3.43 ± 0.03	3.66 ± 0.05
rest	M2MS	60.21 ± 0.98	62.63 ± 1.33	60.04 ± 3.37	60.86 ± 1.94	60.87 ± 1.22	3.56 ± 0.03	3.45 ± 0.04	3.51 ± 0.10	3.22 ± 0.06	3.59 ± 0.04
	S&T	85.05 ± 0.26	85.82 ± 0.54	85.72 ± 1.17	80.39 ± 1.35	85.22 ± 0.53	4.72 ± 0.02	4.75 ± 0.03	4.72 ± 0.05	3.96 ± 0.06	4.78 ± 0.03
	PivotRR	63.02 ± 0.98	64.63 ± 1.34	61.47 ± 3.15	62.33 ± 2.03	63.02 ± 1.20	3.73 ± 0.03	3.69 ± 0.04	3.68 ± 0.09	3.37 ± 0.06	3.77 ± 0.04
	NeutralRR	64.90 ± 0.94	66.30 ± 1.27	62.98 ± 3.15	63.86 ± 1.98	64.42 ± 1.21	3.83 ± 0.03	3.78 ± 0.04	3.78 ± 0.08	3.45 ± 0.05	3.86 ± 0.04

Table 5: Extended results of multi-target cross-lingual summarization in CrossSum for the metrics evaluating semantic coherence across target languages.