μPLAN: Summarizing using a Content Plan as Cross-Lingual Bridge

Anonymous ACL submission

Abstract

Cross-lingual summarization aims to generate a summary in one language given input in a different language, allowing for the dissemination of relevant content among different language speaking populations. The task is challenging mainly due to the paucity of cross-lingual datasets and the compounded difficulty of summarizing *and* translating. This work presents μPLAN, an approach to cross-lingual summarization that uses an intermediate planning step as a cross-lingual bridge. We formulate the plan as a sequence of entities capturing the summary's content and the order in which it should be communicated. Importantly, our plans abstract from surface form: using a multilingual knowledge base, we align entities to their canonical designation across languages and generate the summary conditioned on this cross-lingual bridge and the input. Automatic and human evaluation on the XWikis dataset (across four language pairs) demonstrates that our planning objective achieves state-of-the-art performance in terms of informativeness and faithfulness. Moreover, μPLAN models improve the *zero-shot* transfer to new cross-lingual language pairs compared to baselines without a planning component.

1 Introduction

Given a document or multiple documents in a source language (e.g., English), cross-lingual summarization (Wang et al., 2022a) aims to generate a summary in a different target language (e.g., Czech or German). It enables the rapid dissemination of relevant content across speakers of other languages. For instance, providing summaries of English news articles to Czech or German speakers; or making available to English speakers the content of product and service descriptions in foreign languages.

Recent years have seen tremendous progress in abstractive summarization (Rush et al., 2015; Zhang et al., 2020) thanks to advances in neural network models and the availability of large-scale datasets (Sandhaus, 2008; Hermann et al., 2015; Grusky et al., 2018). While initial efforts have focused on English, more recently, with the advent of cross-lingual representations (Ruder et al., 2019) and large pre-trained models (Devlin et al., 2019; Liu et al., 2020), research on multilingual summarization (i.e., building monolingual summarization systems for different languages) has also gained momentum (Chi et al., 2020; Scialom et al., 2020; Aharoni et al., 2022).

Cross-lingual summarization faces the compounded challenge of having to tackle difficulties relating to both monolingual summarization (e.g., long inputs and outputs, hallucinations; Maynez et al. 2020) *and* machine translation (e.g., data imbalance, alignment across languages; Koehn and Knowles 2017). Recent work has shown that introducing an intermediate content planning step is helpful for summarization in English, resulting in higher quality summaries, especially in terms of faithfulness (Narayan et al., 2021, 2022; Huot et al., 2023). In this work, we argue that content planning also has the potential for producing higher quality outputs for cross-lingual summarization. In particular, it provides a way of sharing task-specific knowledge across languages, while formalizing important aspects of the summarization task: identifying salient content in the source documents, organizing this information in a meaningful order, and standardizing it across different source and target language pairs.

We present μPLAN, a cross-lingual summarization method that uses content planning as a cross-lingual bridge (Figure 1). Building upon previous work (Narayan et al., 2021), we express our content plans as entity chains, i.e., ordered sequences of salient entities. Although more elaborate plan representations have been proposed in the literature (Wang et al., 2022b; Puduppully et al., 2022; Narayan et al., 2022), entities are a natural choice

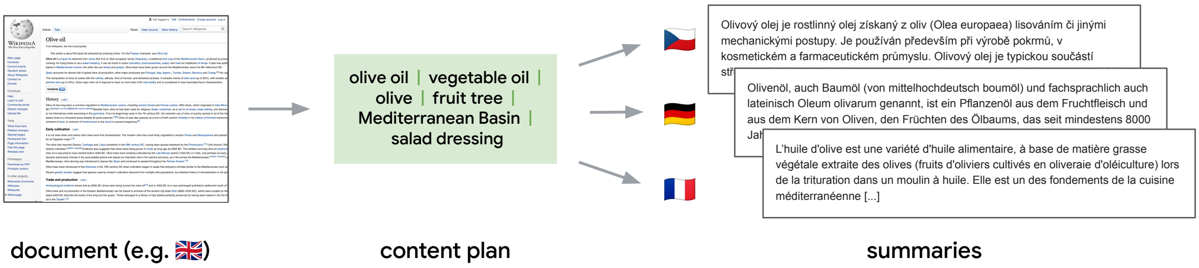

Figure 1: Source document and content plan in English; target summaries in Czech, German, and French.

for our task for two reasons. They can mitigate hallucinations in generated summaries which are commonly related to entities (Cao et al., 2022; Zhao et al., 2020; Maynez et al., 2020) and are well-suited as a bridge across languages, thanks to the availability of multilingual knowledge bases (e.g., DBpedia) which represent entities in different languages. An interesting question for our summarization task is which language to use for the content plan, given that the source document and target summary are in different languages. We employ a multilingual knowledge base to align the entities across languages, which allows us to canonically transpose the plan to different languages without the use of machine translation.

We use a Transformer-based encoder-decoder model (Vaswani et al., 2017) that first encodes the document in the source language and then decodes to generate an intermediate plan representation and the summary in the target language conditioned on the plan and the input. We evaluate our method on the XWikis dataset (Perez-Beltrachini and Lapata, 2021), a cross-lingual abstractive summarization dataset derived from Wikipedia[1] articles aligned across four different languages (English, Czech, French, and German). We augment the training data for fine-tuning by annotating each target summary with its corresponding content plan.

We investigate two distinct cross-lingual tasks, namely from English to other languages (EN → ALL) and from other languages to English (ALL → EN). We demonstrate that models fine-tuned with our planning objective outperform regular generated summaries both in terms of ROUGE and faithfulness on the XWikis dataset across all language pairs, in both settings. Given the scarcity of cross-lingual datasets, we also investigate zero-shot cross-lingual transfer to new language pairs and demonstrate that μPLAN models outperform comparison systems without planning components.

[1]https://www.wikipedia.org/

Our contributions can be summarized as follows: (a) we introduce a training objective for cross-lingual abstractive summarization that uses **entity planning as a bridge between languages**. Using automatic and human evaluation, we show that it yields better quality summaries and more effective zero-shot transfer to new language pairs than non-planning baselines; and (b) we leverage a multilingual knowledge base to annotate the training data with plans, thus **transposing entity names to their canonical designation** in all languages, avoiding errors induced by mistranslation altogether. This strategy enables the mapping of entities that do not have an equivalent name in the target language to fully-localized paraphrases.

2 Related Work

Cross-lingual Summarization A key challenge in cross-lingual summarization is the scarcity of training data. Indeed, while creating large-scale multilingual summarization datasets has proven feasible (Straka et al., 2018; Scialom et al., 2020), naturally occurring documents in a source language paired with summaries in different target languages are rare. For this reason, existing cross-lingual approaches create large-scale synthetic data using machine translation (Zhu et al., 2019; Cao et al., 2020; Ouyang et al., 2019).

Cross-lingual benchmarks include WikiLingua (Ladhak et al., 2020), a dataset derived from multilingual how-to guides, which are relatively short and their summaries limited to brief instructional sentences. CrossSum (Bhattacharjee et al., 2021) contains over a million article and summary samples, aligned from the multilingual XL-Sum (Hasan et al., 2021) dataset, but the summaries are limited to one or two sentences. We work with XWikis (Perez-Beltrachini and Lapata, 2021), a cross-lingual dataset derived from Wikipedia with long input documents and long target summaries. We compare these datasets in Appendix A.

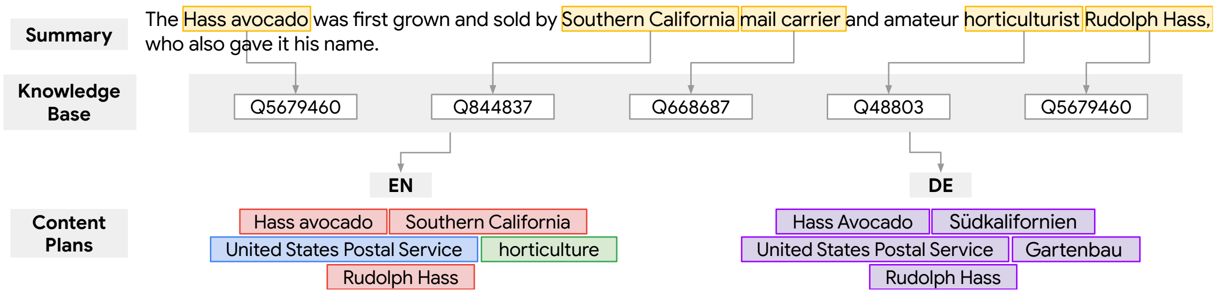

Figure 2: Plan annotation on an example summary (salient entities highlighted in yellow). After pivoting on the knowledge base, corresponding canonical entities in English are shown in the bottom left. Most times they match the surface form in the summary (in red), other times they have the same root (in green) but they could differ greatly when entities need disambiguation (in blue). The aligned German content plan is shown in the bottom right.

Content Plans for Summarization The idea of breaking down the generation task into smaller steps through a separate planning stage has proven helpful for data-to-text generation (Puduppully et al., 2019; Moryossef et al., 2019; Puduppully and Lapata, 2021; Liu and Chen, 2021) and lately for summarization and long-form question answering (Narayan et al., 2021, 2022). Our work is closest to Narayan et al. (2021) who show that an intermediate planning step conceptualized as a sequence of salient entities could yield more faithful and entity-specific summaries. Herein, we explore whether content plans can serve as a cross-lingual bridge and enable *task transfer* between languages.

Zero-shot Cross-lingual Transfer A substantial portion of the work on zero-shot cross-lingual transfer has focused on classification tasks (Hu et al., 2020), such as XNLI (Artetxe and Schwenk, 2019), part-of-speech tagging, dependency parsing, named entity recognition (Ansell et al., 2021), and question answering (Conneau et al., 2020). Some recent work has also investigated generative tasks in the zero-shot setting. Johnson et al. (2017) show that by prepending a special token to the input text to indicate the target language of the translation, models learn to perform implicit bridging between language pairs unseen during training. Chen et al. (2021) perform zero-shot cross-lingual machine translation, by using parallel data in only one language pair and leveraging a multilingual encoder to support inference in other languages. Vu et al. (2022) study how to fine-tune language models on only one language to perform zero-shot cross-lingual summarization in other languages, by adding unlabeled multilingual data. Whitehouse et al. (2022) use Wikidata to improve zero-shot cross-lingual transfer for code-switching in a number of entity-centric downstream tasks. We also resort to Wikidata to obtain a canonical designation of entities across languages, however, the use of plans as a cross-lingual bridge for summarization is new to our knowledge.

3 Plans as a Cross-Lingual Bridge

3.1 Problem Formulation

We formalize the cross-lingual abstractive summarization task as follows: Given an input document d in a source language SRC, generate a summary s in target language TGT. We model this as $p(s|d)$.

For the content planning objective, our goal is to teach the model to first generate a content plan c for the summary as $p(c|d)$, before generating the summary itself as $p(s|c,d)$. Following Narayan et al. (2021), instead of modeling $p(c|d)$ and $p(s|c,d)$ separately, we train the model to generate the concatenated plan and summary sequence $c; s$. As a result, the model first generates the content plan c and then continues to generates the summary s conditioned on both c and d. In the following section, we describe how we annotate the data with content plans for this planning objective.

3.2 Content Plans

Similarly to Narayan et al. (2021), we formulate the content plan as an ordered sequence of entities. Figure 2 illustrates our annotation process. We annotate each example with its corresponding content plan by extracting salient entities, i.e., entities that are important to mention when summarizing.

We extend this paradigm by linking each entity to its entry in a multilingual knowledge base. This way we obtain a canonical designation of each entity, removing morphology and selecting the most common designation out of multiple aliases.

<table>
<tr><th></th><th>Summary</th><th>Plan</th></tr>
<tr><td>EN → CS</td><td>Richard Dagobert Brauer byl německý matematik žijící v USA. Pracoval zejména v oblastech abstraktní algebry a teorie čísel. Je také zakladatelem modulární teorie reprezentací.</td><td>German Empire & Německé císařství | mathematician & matematik | United States of America & Spojené státy americké | algebra & algebra | number theory & teorie čísel</td></tr>
<tr><td>EN → FR</td><td>CALET est un observatoire spatial développé par le Japon et installé en 2015 à bord de la Station spatiale internationale. Cet instrument analyse les rayons cosmiques et le rayonnement gamma à haute énergie avec comme objectif principal l'identification des éventuelles signatures de la matière noire.</td><td>space observatory & télescope spatial | Japan & Japon | International Space Station & station spatiale internationale | cosmic radiation & rayonnement cosmique | gamma ray & rayon gamma | dark matter & matière noire</td></tr>
<tr><td>DE → EN</td><td>The TKS spacecraft ("Transport Supply Spacecraft", GRAU index 11F72) was a Soviet spacecraft conceived in the late 1960s for resupply flights to the military Almaz space station.</td><td>Hauptverwaltung für Raketen und Artillerie & GRAU | Sowjetunion & Soviet Union | Raumschiff & spacecraft | Almas & Almaz</td></tr>
</table>

Table 1: Summaries with annotated plans. Same color denotes alignment between entities in the plan and summary. Plans are entities in the language of the source document *and* the language of the target summary.

The knowledge base also provides disambiguation when it is needed. We use entity names in the content plans, instead of knowledge base indices, in order to leverage the natural language capabilities of pretrained language models.

We then use the inter-language information from the knowledge base to pivot content plans across languages. For each entity, we obtain its canonical designation in both the language of the source document and the language of the target summary. We provide an example of the multilingual mappings in our annotated content plans in Figure 2. This strategy enables the mapping of entities that do not have an equivalent name in the target language to fully-localized names. And the model learns to generate a content plan of localized entities, avoiding errors induced by translation.

Finally, we compose the content plan as a sequence of canonical entity names, each expressed in pairs in both the source and target language (Table 1). We designate the planning objective using these cross-lingual content plans as μPLAN.

3.3 Summarization Tasks

We next define the summarization tasks considered in this work, and our assumptions about the cross-lingual training data being available.

Cross-Lingual Tasks In what follows, let $\mathcal{L}$ be the set of all languages, SRC the language of the source document, and TGT the language of the target summary. We denote the cross-lingual data as $\mathcal{D}_{\text{SRC}\to\text{TGT}}$, i.e., $\mathcal{D}_{\text{EN}\to\text{CS}}$ for Czech summaries aligned with English inputs. Analogously, we denote the monolingual data as $\mathcal{D}_{\text{LANG}}$, i.e., $\mathcal{D}_{\text{CS}}$ for Czech summaries with Czech inputs.

Herein, we investigate two specific cross-lingual tasks: (a) from English to other languages and (b) from other languages to English, which we denote as EN → ALL and ALL → EN, respectively. The EN → ALL task is the main focus of our work. The task is particularly interesting because it would make a large amount of English information available to speakers of other languages but also challenging since it involves a cross-lingual summarization model that can generate fluent text in many languages. We define the data for the EN → ALL task as:

$$\mathcal{D}_{\text{EN}\to\text{ALL}} = \mathcal{D}_{\text{EN}} \cup \bigcup_{\text{TGT}\in\mathcal{L}-\{\text{EN}\}} \mathcal{D}_{\text{EN}\to\text{TGT}},$$

and for the ALL → EN, task as:

$$\mathcal{D}_{\text{ALL}\to\text{EN}} = \mathcal{D}_{\text{EN}} \cup \bigcup_{\text{SRC}\in\mathcal{L}-\{\text{EN}\}} \mathcal{D}_{\text{SRC}\to\text{EN}}.$$

Note that both tasks have access to monolingual EN data. For models that do not use an intermediate planning step, each data example is a document and summary pair (d, s). For μPLAN models, each data example also includes a content plan, $(d, c; s)$.

Zero-Shot Cross-Lingual Tasks Given the scarcity of cross-lingual datasets, we investigate whether μPLAN can help with the zero-shot cross-lingual transfer to new language pairs. For each target language TGT, we perform zero-shot transfer experiments on the EN → ALL task by holding out the EN → TGT cross-lingual data during fine-tuning. We then evaluate performance on the EN → TGT test data. To ensure that the model

	Train	Validation	Test
EN	624,178	8,194	7,000
EN → CS	134,996	250†	6,855†
EN → DE	409,012	250†	9,750†
EN → FR	451,964	250†	9,727†
CS → EN	48,519	2,549	6,999
DE → EN	344,438	18,160	6,999
FR → EN	283,182	14,899	6,992

Table 2: Number of data samples in the XWikis dataset and splits considered in this work. New splits for the EN → ALL language pairs are marked by †.

maps the language token to the correct language and to prevent catastrophic forgetting of the TGT language during fine-tuning (Vu et al., 2022), we include TGT monolingual summarization data in the fine-tuning data mixture, under the assumption that monolingual data is easier to come by than cross-lingual data. We denote this zero-shot cross-lingual transfer task as EN → TGT_{ZS} and define as:

$$\mathcal{D}_{\text{EN}\rightarrow\text{TGT}_{\text{ZS}}} = \mathcal{D}_{\text{EN}} \cup \mathcal{D}_{\text{TGT}} \cup \bigcup_{\text{L}\in\mathcal{L}-\{\text{EN},\text{TGT}\}} \mathcal{D}_{\text{EN}\rightarrow\text{L}}.$$

For greater generalization, we could use unlabeled monolingual data (without summaries), however, we leave this to future work.

4 Experimental Setup

4.1 Dataset

The XWikis dataset (Perez-Beltrachini and Lapata, 2021) was created from Wikipedia articles under the assumption that the body and lead paragraph constitute a document-summary pair. Cross-lingual document-summary instances were derived by combining lead paragraphs and articles' bodies from language-aligned Wikipedia titles. Although XWikis covers only four languages, English (EN), Czech (CS), German (DE), and French (FR), the dataset creation procedure is general and applicable to any languages represented in Wikipedia.

Table 2 shows the number of data samples for each language pair. Note that the EN → TGT language pairs are not parallel between all languages. Cross-lingual language pairs in the ALL → EN setting have separate training, validation and test splits, but in the EN → ALL setting there are only training and validation splits. Therefore, for all the EN → ALL cross-lingual language pairs, we separate the validation split into two, taking the first 250 examples for validation and the rest for testing.

The XWikis dataset provides the input documents as a list of section titles and paragraphs that constitute the body of the Wikipedia article to summarize. We format the input documents by concatenating the titles and paragraphs, marking each title with an end-of-title token EOT and each paragraph with an end-of-paragraph token EOP. We prepend the source language code and target language code to the input document for each cross-lingual document and summary pair.

Since the XWikis dataset is derived from Wikipedia, we annotate the plans by extracting the entities from the summaries using the embedded hyperlinks that point to Wikidata[2] knowledge base entries. Not all the links correspond to entities, and we exclude those pointing to phonetic pronunciations. The XWikis dataset was generated from a 2016 Wikipedia data dump and we used one from 2023 for extracting the hyperlinks from the summaries. Therefore, for articles that went through significant changes between 2016 and 2023, the pages were not aligned and we did not annotate these examples with content plans. This problem affects about 4.5% of the training data. We create a *filtered* version of the training data that excludes these examples with missing content plans.

4.2 Comparison Models

We demonstrate µPLAN on both the EN → ALL and ALL → EN tasks and compare it with a number of different modeling approaches.

Machine Translation A common approach is to adopt a machine translation-based pipeline which can be used in two ways: (a) first translate the original document into the target language and then summarize the translated document or (b) first summarize the original document and then translate the summary (Ouyang et al., 2019; Wan et al., 2010; Ladhak et al., 2020). We denote the former approach as Translate-train (TR_{train}) and the latter as Translate-test (TR_{test}). We perform machine translation with Google Translate.

Previous work (Kramchaninova and Defauw, 2022; Vu et al., 2022) has highlighted various limitations with these approaches such as dependence on the quality of available machine translation systems in a given language and in turn the availability of high-quality parallel data, a potential misalignment of the data after translation, and translationese artifacts (Clark et al., 2020).

[2]https://www.wikidata.org/

Plan Type	Predicted Plan	Gold Plan
SRC[EN]	Dutch | fortification | Banda Neira | Maluku Islands | Netherlands | Dutch East Indies	Banda Neira | Banda Islands | Maluku Islands | Indonesia | Maluku | nutmeg
TGT[DE]	Estland | Folk Metal | Band | Tallinn | Markus Lõhmus	Estland | Folk Metal | Euphemismus | Wolf
SRC[EN]_TGT[FR]	county seat & siège de comté | Crawford County & comté de Crawford | Arkansas & Arkansas | United States of America & États-Unis	Arkansas & Arkansas | United States of America & États-Unis

Table 3: Examples of generated and gold content plans for different source and target languages.

End-to-end Summarization This approach, which we denote as E2E, directly fine-tunes a multilingual pretrained model on the cross-lingual data (Perez-Beltrachini and Lapata, 2021). It does not incorporate a planning component, but avoids the potential error propagation problem of machine translation pipeline systems.

µPLAN Variants We experiment with different plan formulations to establish which type of plan performs well as a cross-lingual bridge. The language of the source document being different from the language of the target summary raises the question of which language to use for the content plans. In the default µPLAN setup, entities in the plan are expressed in pairs, with their canonical name in both the language of the source document and the language of the target summary. In addition, we explore two alternatives: (a) entity names only in the source language and (b) entity names only in the target language. Table 3 presents examples of different language plans. Moreover, we experiment with the internal constitution of the plans: we provide the length of the gold plan during training [LENGTH], and shuffle entities to investigate the importance of the sequence order [SHUFFLE].

All baselines and µPLAN variants are based on the mT5 model (Xue et al. 2021; XL 3.7B parameters) which we finetune with maximum input and output sequence lengths of 2,048 and 256 tokens, respectively. Our models are finetuned with a learning rate of 0.002, a batch size of 128, up to 80,000 steps, evaluating every 1,000 steps. We select the best checkpoints by measuring ROUGE-L (see Section 5.1 for details) on 250 examples of the validation split for each language pair and take the best unweighted average across all language pairs.

Note on LLMs We performed few-shot experiments with LLMs, however, these were consistently inferior to our fine-tuned systems confirming the observations of Maynez et al. (2023). It is particularly challenging to learn to plan and summarize simply from a few examples. We report LLM experiments (1-shot, no planning) in Appendix E.

5 Results

5.1 Automatic Evaluation

We automatically evaluate system output along the dimensions of summary relevance, summary faithfulness, and content plan relevance. For *summary relevance*, we use ROUGE (Lin, 2004) to compare system-generated summaries with gold-standard ones. Since the availability of word tokenizers differs for non-English languages, we follow Aharoni et al. (2022) and compute ROUGE with a SentencePiece tokenizer (Kudo and Richardson, 2018) trained on mC4 (Xue et al., 2021).

In terms of *summary faithfulness*, following Honovich et al. (2022), we employ an entailment classifier that predicts whether the input document supports the output summary. In line with previous work (Narayan et al., 2022; Schuster et al., 2022), we split the summary into sentences for a more fine-grained evaluation. We predict the entailment of each sentence and average the entailment scores. We use an mT5-XXL model (Xue et al., 2021) trained on XNLI (Conneau et al., 2018), a multilingual NLI dataset. There are currently no cross-lingual datasets for NLI, however our preliminary analysis reported in Appendix B shows that an XNLI-trained mT5 model works well in predicting cross-lingual entailment. It has the added benefit of avoiding potential error propagation from introducing a machine translation step in the evaluation process (e.g., translating the document or the summary in English). Finally, we evaluate *plan relevance*, by comparing generated content plans against gold-standard ones. Specifically, we compute F1 scores on the entities in the predicted summaries against the corresponding reference entities.

Planning outperforms translation-based approaches Table 4 presents an overview of our results for the EN $\rightarrow$ ALL and ALL $\rightarrow$ EN tasks.

	ROUGE-L				XNLI			
	TR$_{train}$	TR$_{test}$	E2E	μPLAN	TR$_{train}$	TR$_{test}$	E2E	μPLAN
EN → EN	37.42	37.38	37.57	**39.53**	53.99	47.50	53.54	**56.16**
EN → CS	32.81	26.26	32.74	**33.18**	34.32	36.90	33.79	**37.70**
EN → DE	38.28	28.47	38.58	**38.94**	39.52	38.19	38.92	**42.98**
EN → FR	41.19	31.59	41.36	**41.57**	41.45	40.75	40.83	**52.72**
EN → ALL	37.42	30.93	37.56	**38.30**	42.32	40.84	41.77	**47.39**

	ROUGE-L				XNLI			
	TR$_{train}$	TR$_{test}$	E2E	μPLAN	TR$_{train}$	TR$_{test}$	E2E	μPLAN
EN → EN	33.15	34.43	35.47	**36.09**	63.29	**66.46**	51.79	60.71
CS → EN	29.47	31.93	**33.30**	32.82	**45.39**	30.39	30.14	30.81
DE → EN	29.89	32.48	33.70	**34.32**	**45.20**	42.17	35.22	41.16
FR → EN	29.60	32.35	33.22	**34.20**	**41.63**	39.81	32.58	39.34
ALL → EN	30.53	32.80	33.92	**34.36**	**48.88**	44.71	37.43	43.00

Table 4: ROUGE-L and XNLI results per language pair and overall for the EN → ALL and ALL → EN tasks. Systems significantly different from μPLAN are underlined (using paired bootstrap resampling; $p < 0.05$).

	ROUGE-L	XNLI	F1
μPLAN	38.30	47.39	0.40
μPLAN$_{\text{SRC}}$	38.14	47.72	0.41
μPLAN$_{\text{TGT}}$	37.97	47.37	0.40
μPLAN$_{\text{LENGTH}}$	37.09	45.71	0.37
μPLAN$_{\text{SHUFFLE}}$	38.01	46.25	0.40
μPLAN^{oracle}	48.28	40.83	1.00
μPLAN$^{oracle}_{\text{SRC}}$	47.96	41.22	1.00
μPLAN$^{oracle}_{\text{TGT}}$	48.13	40.84	1.00

Table 5: Comparison of different μPLAN plan formulations (including oracles) on the EN → ALL task.

We report results on the filtered data, as we observed little difference overall between filtered and non-filtered training samples (results with non-filtered training data are provided in Appendix D). Moreover, for the sake of brevity, we only present ROUGE-L results, however see Appendix C for additional metrics. We see that μPLAN consistently outperforms both the translation-based approaches and the non-planning baseline (E2E) in terms of ROUGE-L and XNLI scores on both EN → ALL and ALL → EN tasks. Note that TR$_{train}$ is the overall winner according to XNLI in the ALL → EN task. We hypothesize XNLI just works better in this setting as it is faced with a simpler monolingual task (both the input document and summary are in English). Previous work (Perez-Beltrachini and Lapata, 2021) has focused on ALL → EN tasks using mBART50 (Tang et al., 2020) and E2E models; they report an average ROUGE-L of 32.76 for the same language pairs shown in Table 4 (last row).

Best plans include entities in source and target language We compare different types of plan formulations on the EN → ALL task and report our results in Table 5. Mixed language plans that contain entities in both the source and target language, which is the default μPLAN setting, deliver better results than plans with entities in only one language (marked here as SRC and TGT). Table 3 shows some plans generated by μPLAN under these different settings and compares them to the gold ones.

Predicted and gold plans have similar length, measured by the number of entities in the plan (6 on average). We also find that gold and predicted plans have overlapping but not identical entities (the F1 score is around 0.4; see Tables 5 and 3). However, we do not expect perfect overlap; gold summaries in XWikis are derived from lead paragraphs in Wikipedia articles, and as a result some of the entities in the gold plans might not even appear in the source document. This is corroborated by XNLI scores which are lower for oracle summaries compared to machine-generated ones. Providing information about the length of the gold plan during training, reported as LENGTH, does not affect the results very much and actually yields slightly lower metrics than the default μPLAN setup. The SHUFFLE metrics, for which the entity order is shuffled, are similar to the default setup. This result indicates that the order of the entities does not matter much for planning the summary generation.

Oracle plans show there is room for improvement For comparison, we report results when models have access to oracle content plans, which

	ROUGE-L		XNLI	
	E2E	μPLAN	E2E	μPLAN
EN → CS$_{ZS}$	15.10	**18.64**	34.95	**39.04**
EN → DE$_{ZS}$	17.50	**19.18**	45.51	**48.80**
EN → FR$_{ZS}$	18.54	**23.61**	45.51	**45.96**

Table 6: Zero-shot cross-lingual transfer results.

	μPLAN vs. E2E			μPLAN vs. Reference		
	Win	Lose	Tie	Win	Lose	Tie
Coherence	6.3	**7.0**	86.7	**10.7**	7.6	81.7
Accuracy	**13.3**	7.0	79.7	**15.7**	13.6	70.7
Inform	**20.0**	11.7	68.3	14.0	**16.7**	69.3
Overall	**41.0**	24.7	34.3	33.0	**35.7**	31.3

Table 7: Human evaluation results aggregated over three language pairs (EN → CS, EN → DE, EN → FR); statistically significant differences are underlined.

we denote as *oracle*. At inference time, the encoder first encodes the source document, while the decoder gets the gold plan as a forced prompt before generating the summary. These oracle experiments provide an upper bound of how μPLAN models would perform in a best case scenario. In Table 5, we see that the oracle metrics are higher by a wide margin, of around 10 ROUGE-L points, from the best predicted results. This behavior is expected and shows that models can correctly generate summaries from plans in the target language but also from aligned English plans. Moreover, these results confirm that μPLAN's mixed language plans provide additional information that models can leverage effectively.

Planning enables zero-shot transfer Table 6 shows the results of our zero-shot cross-lingual transfer experiments. We observe that μPLAN delivers higher ROUGE-L when evaluated on an unseen language pair. This indicates that an intermediate planning step helps transfer task knowledge to new language pairs.

5.2 Human Evaluation

In addition to automatic metrics, we also conducted a judgment elicitation study. Specifically, we compared μPLAN, against the E2E system, and reference summaries. Raters were shown a document, alongside two summaries and were asked to provide pairwise references along the following dimensions: *Coherence* (is the summary easy to understand and grammatically correct?), *Accuracy* (is all the information in the summary attributable to the original text?), and *Informativeness* (does the summary capture important information from the original text?). We recruited 178 annotators (all native speakers) and elicited preferences for 100 summaries (test set) per language pair (EN → CS, EN → DE, EN → FR). Appendix F showcases our instructions and examples of summaries our annotators rated.

We present aggregate results in Table 7 (see Appendix F for detailed analysis). μPLAN summaries are as coherent as E2E summaries but significantly more accurate and informative ($p < 0.05$ using a Wilcoxon signed-rank test). Interestingly, our raters find μPLAN summaries on par with gold summaries across all dimensions (differences between them are *not* significant).

6 Conclusion

In this work we present μPLAN, an approach to cross-lingual summarization that uses an intermediate planning step as a cross-lingual bridge. Since hallucinations and mistranslations in cross-lingual summarization are often tied to incorrect entities, we formulate the content plan as a sequence of entities expressing salient content and how it should be presented. Evaluation on the XWikis dataset demonstrates that this planning objective achieves state-of-the-art performance in EN → ALL and ALL → EN settings and enables zero-shot cross-lingual transfer to new language pairs.

In this work, we use the embedded hyperlinks in Wikipedia articles to extract salient entities and align them on the Wikidata knowledge base. With recent entity annotation systems such as REFINED (Ayoola et al., 2022), the same operation can be applied on out-of-domain data, including the multilingual alignment of the entity names. Unlike latent variable-based intermediate representations, our content plans are interpretable (they are expressed in natural language) and can be easily edited, e.g., by filtering the entities at inference time or with a human in the loop (Narayan et al., 2021, 2022; Huot et al., 2023). Using forced prompting methods as described in the oracle experiments, would also allow us to localize entity names at inference time from a knowledge base. In the future, we plan to explore the task transfer capabilities of μPLAN in low-resource settings as we cannot realistically expect to have large-scale cross-lingual data on all possible language pairs.

Limitations

An ethical consideration with generative language models is the problem of misinformation. While the work we present here makes a step towards improving the faithfulness and factual consistency of text generation systems, it is important to note that current systems are still far from being perfect in this respect, and thus should be used with caution.

References

Roee Aharoni, Shashi Narayan, Joshua Maynez, Jonathan Herzig, Elizabeth Clark, and Mirella Lapata. 2022. mface: Multilingual summarization with factual consistency evaluation. *arXiv preprint arXiv:2212.10622*.

Rohan Anil, Andrew M Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, et al. 2023. Palm 2 technical report. *arXiv preprint arXiv:2305.10403*.

Alan Ansell, Edoardo Maria Ponti, Jonas Pfeiffer, Sebastian Ruder, Goran Glavaš, Ivan Vulić, and Anna Korhonen. 2021. MAD-G: Multilingual adapter generation for efficient cross-lingual transfer. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4762–4781, Punta Cana, Dominican Republic. Association for Computational Linguistics.

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.

Tom Ayoola, Shubhi Tyagi, Joseph Fisher, Christos Christodoulopoulos, and Andrea Pierleoni. 2022. ReFinED: An efficient zero-shot-capable approach to end-to-end entity linking. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Industry Track*, pages 209–220, Hybrid: Seattle, Washington + Online. Association for Computational Linguistics.

Abhik Bhattacharjee, Tahmid Hasan, Wasi Uddin Ahmad, Yuan-Fang Li, Yong-Bin Kang, and Rifat Shahriyar. 2021. Crosssum: Beyond english-centric cross-lingual abstractive text summarization for 1500+ language pairs. *arXiv preprint arXiv:2112.08804*.

Meng Cao, Yue Dong, and Jackie Cheung. 2022. Hallucinated but factual! inspecting the factuality of hallucinations in abstractive summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3340–3354, Dublin, Ireland. 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.

Guanhua Chen, Shuming Ma, Yun Chen, Li Dong, Dongdong Zhang, Jia Pan, Wenping Wang, and Furu Wei. 2021. Zero-shot cross-lingual transfer of neural machine translation with multilingual pretrained encoders. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 15–26, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Zewen Chi, Li Dong, Furu Wei, Wenhui Wang, Xian-Ling Mao, and Heyan Huang. 2020. Cross-lingual natural language generation via pre-training. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence*, pages 7570–7577. AAAI Press.

Jonathan H. Clark, Eunsol Choi, Michael Collins, Dan Garrette, Tom Kwiatkowski, Vitaly Nikolaev, and Jennimaria Palomaki. 2020. TyDi QA: A benchmark for information-seeking question answering in typologically diverse languages. *Transactions of the Association for Computational Linguistics*, 8:454–470.

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.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Max Grusky, Mor Naaman, and Yoav Artzi. 2018. Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies. In *Proceedings of the 2018 Conference of the North American Chapter of* the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 708–719, New Orleans, Louisiana. Association for Computational Linguistics.

Tahmid Hasan, Abhik Bhattacharjee, Md Saiful Islam, Kazi Samin, Yuan-Fang Li, Yong-Bin Kang, M Sohel Rahman, and Rifat Shahriyar. 2021. Xl-sum: Large-scale multilingual abstractive summarization for 44 languages. *arXiv preprint arXiv:2106.13822*.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *Advances in Neural Information Processing Systems 28*, pages 1693–1701. Curran Associates, Inc.

Or Honovich, Roee Aharoni, Jonathan Herzig, Hagai Taitelbaum, Doron Kukliansy, Vered Cohen, Thomas Scialom, Idan Szpektor, Avinatan Hassidim, and Yossi Matias. 2022. TRUE: Re-evaluating factual consistency evaluation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3905–3920, Seattle, United States. Association for Computational Linguistics.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. XTREME: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 4411–4421. PMLR.

Fantine Huot, Joshua Maynez, Shashi Narayan, Reinald Kim Amplayo, Kuzman Ganchev, Annie Priyadarshini Louis, Anders Sandholm, Dipanjan Das, and Mirella Lapata. 2023. Text-blueprint: An interactive platform for plan-based conditional generation. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 105–116, Dubrovnik, Croatia. Association for Computational Linguistics.

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Philipp Koehn and Rebecca Knowles. 2017. Six challenges for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 28–39, Vancouver. Association for Computational Linguistics.

Alina Kramchaninova and Arne Defauw. 2022. Synthetic data generation for multilingual domain-adaptable question answering systems. In *Proceedings of the 23rd Annual Conference of the European Association for Machine Translation*, pages 151–160, Ghent, Belgium. European Association for Machine Translation.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. 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.

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 pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Zhengyuan Liu and Nancy Chen. 2021. Controllable neural dialogue summarization with personal named entity planning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 92–106, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Joshua Maynez, Priyanka Agrawal, and Sebastian Gehrmann. 2023. Benchmarking large language model capabilities for conditional generation. *arXiv preprint arXiv:2306.16793*.

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.

Amit Moryossef, Yoav Goldberg, and Ido Dagan. 2019. Step-by-step: Separating planning from realization in neural data-to-text generation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2267–2277, Minneapolis, Minnesota. Association for Computational Linguistics.

Shashi Narayan, Joshua Maynez, Reinald Kim Amplayo, Kuzman Ganchev, Annie Louis, Fantine

Huot, Dipanjan Das, and Mirella Lapata. 2022. Conditional generation with a question-answering blueprint. *arXiv preprint arXiv:2207.00397*.

Shashi Narayan, Yao Zhao, Joshua Maynez, Gonçalo Simões, Vitaly Nikolaev, and Ryan McDonald. 2021. Planning with learned entity prompts for abstractive summarization. *Transactions of the Association for Computational Linguistics*, 9:1475–1492.

Jessica Ouyang, Boya Song, and Kathy McKeown. 2019. A robust abstractive system for cross-lingual summarization. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2025–2031, Minneapolis, Minnesota. 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.

Ratish Puduppully, Li Dong, and Mirella Lapata. 2019. Data-to-text generation with content selection and planning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 6908–6915.

Ratish Puduppully, Yao Fu, and Mirella Lapata. 2022. Data-to-text generation with variational sequential planning. *Transactions of the Association for Computational Linguistics*, 10:697–715.

Ratish Puduppully and Mirella Lapata. 2021. Data-to-text generation with macro planning. *Transactions of the Association for Computational Linguistics*, 9:510–527.

Sebastian Ruder, Ivan Vulić, and Anders Søgaard. 2019. A survey of cross-lingual word embedding models. *Journal of Artificial Intelligence Research*, 65(1):569–630.

Alexander M. Rush, Sumit Chopra, and Jason Weston. 2015. A neural attention model for abstractive sentence summarization. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 379–389, Lisbon, Portugal. Association for Computational Linguistics.

Evan Sandhaus. 2008. The New York Times Annotated Corpus. *Linguistic Data Consortium, Philadelphia*, 6(12).

Tal Schuster, Sihao Chen, Senaka Buthpitiya, Alex Fabrikant, and Donald Metzler. 2022. Stretching sentence-pair NLI models to reason over long documents and clusters. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 394–412, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Thomas Scialom, Paul-Alexis Dray, Sylvain Lamprier, Benjamin Piwowarski, and Jacopo Staiano. 2020. MLSUM: The multilingual summarization corpus. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8051–8067, Online. Association for Computational Linguistics.

Milan Straka, Nikita Mediankin, Tom Kocmi, Zdeněk Žabokrtský, Vojtěch Hudeček, and Jan Hajič. 2018. SumeCzech: Large Czech news-based summarization dataset. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2020. Multilingual translation with extensible multilingual pretraining and finetuning. *arXiv preprint arXiv:2008.00401*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6000–6010, Red Hook, NY, USA. Curran Associates Inc.

Tu Vu, Aditya Barua, Brian Lester, Daniel Cer, Mohit Iyyer, and Noah Constant. 2022. Overcoming catastrophic forgetting in zero-shot cross-lingual generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9279–9300, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Xiaojun Wan, Huiying Li, and Jianguo Xiao. 2010. Cross-language document summarization based on machine translation quality prediction. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 917–926, Uppsala, Sweden. Association for Computational Linguistics.

Jiaan Wang, Yunlong Liang, Fandong Meng, Zhixu Li, Jianfeng Qu, and Jie Zhou. 2023. Cross-lingual summarization via chatgpt. *arXiv preprint arXiv:2302.14229*.

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

Ye Wang, Xiaojun Wan, and Zhiping Cai. 2022b. Guiding abstractive dialogue summarization with content planning. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 3408–3413, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Chenxi Whitehouse, Fenia Christopoulou, and Ignacio Iacobacci. 2022. EntityCS: Improving zero-shot cross-lingual transfer with entity-centric code switching. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 6698–6714, Abu Dhabi, United Arab Emirates. 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.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. 2020. PEGASUS: Pre-training with extracted gap-sentences for abstractive summarization. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 11328–11339. PMLR.

Zheng Zhao, Shay B. Cohen, and Bonnie Webber. 2020. Reducing quantity hallucinations in abstractive summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2237–2249, Online. Association for Computational Linguistics.

Junnan Zhu, Qian Wang, Yining Wang, Yu Zhou, Jiajun Zhang, Shaonan Wang, and Chengqing Zong. 2019. NCLS: Neural cross-lingual summarization. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3054–3064, Hong Kong, China. Association for Computational Linguistics.

	Lang	Pairs	SumL	DocL
MultiLing'13	40	30	185	4,111
MultiLing'15	38	30	233	4,946
Global Voices	15	229	51	359
WikiLingua	18	45,783	39	391
XWikis	4	213,911	77	945
CrossSum	45	22,727	23	431

Table 8: Number of languages (Lang), average number of document-summary pairs (Pairs), average summary (SumL) and document (DocL) length in terms of number of tokens for different cross-lingual datasets.

A Cross-lingual Summarization Datasets

Table 8 summarizes existing cross-lingual datasets. We see that the XWikis dataset (Perez-Beltrachini and Lapata, 2021) features longer input documents and target summaries.

B Cross-lingual NLI

Table 9 compares different ways of computing NLI. It is computed on the summaries generated by the baseline E2E model on the EN → ALL and ALL → EN tasks. The first setting, denoted as ANLI, is the English setting, for which we translate the non-English document (ALL → EN) or summary (EN → ALL) to English and apply an NLI model trained on an English corpus. The second one is the multilingual NLI setting, which we denote as XNLI-m. For the cross-lingual language pairs, we translate the English document or summary such that both document and summary are in the same language (which is either the source or target language, depending on whether it is the EN → ALL or ALL → EN task). We then apply a multilingual NLI model. The last setting is the cross-lingual setting, which we denote as XNLI-x. In this setting, we do not use translation, and directly apply the multilingual NLI model to the cross-lingual data.

C Experimental Results

In Table 10 we present the full set of ROUGE scores for the EN → ALL and ALL → EN tasks.

D Effects of Filtered Training Data

Table 11 compares the results obtained with the *filtered* and non-filtered training data. Overall, the results are similar, which is expected since the difference in the number of training samples is relatively small.

		ANLI	XNLI-m	XNLI-x
EN → ALL	EN	54.04	–	53.63
	EN → CS	32.09	31.15	35.88
	EN → DE	38.47	39.89	40.15
	EN → FR	43.09	35.74	41.32
ALL → EN	EN	57.91	–	53.05
	CS → EN	34.73	32.95	29.74
	DE → EN	40.28	38.64	35.12
	FR → EN	37.28	35.71	32.40

Table 9: Entailment metrics on English, multilingual, and cross-lingual settings.

E Few-shot Prompting of LLMs

LLMs have demonstrated promising results in few-shot settings for cross-lingual summarization (Wang et al., 2023). In Table 12, we report 1-shot results obtained using PaLM 2 (Anil et al., 2023), a 340B parameter LLM. We perform 1-shot experiments for all language pairs in the EN → ALL and ALL → EN tasks. For each language pair, the prompt is formulated as follows:

```
From a document in [source language],
write a summary in [target language].

(1)
Document: [example document]
Summary: [example summary]

(2)
Document: [document]
Summary:
```

The example document and summary are taken from the training splits. We truncate the input documents at 2000 tokens to fit within the model's maximum sequence input length. We limit the experiments to the 1-shot setting, since more than one data example exceeds the maximum sequence length.

These 1-shot LLM experiments underperformed overall compared to our finetuned baselines. The ROUGE-L scores are lower than both the E2E and µPLAN models and the NLI scores are much lower than all models. In the EN → CS task, the model often generated outputs in English instead of Czech. These results highlight some of the challenges of learning cross-lingual summarization from just a few examples.

While the few-shot setting has its limitations, fine-tuning large language models (LLMs) is com-

	ROUGE-1				ROUGE-2			
	TR$_{train}$	TR$_{test}$	E2E	μPLAN	TR$_{train}$	TR$_{test}$	E2E	μPLAN
EN → EN	45.38	**47.95**	45.47	47.43	28.61	30.26	28.73	**30.61**
EN → CS	40.74	35.12	40.72	**41.02**	23.86	17.08	23.70	**24.43**
EN → DE	44.51	37.49	44.58	**45.34**	28.99	18.27	29.26	**29.35**
EN → FR	48.69	42.15	48.73	**49.23**	32.81	22.00	32.89	**33.20**
EN → ALL	44.83	40.68	44.87	**45.75**	28.56	21.90	28.65	**29.40**

	ROUGE-1				ROUGE-2			
	TR$_{train}$	TR$_{test}$	E2E	μPLAN	TR$_{train}$	TR$_{test}$	E2E	μPLAN
EN → EN	40.61	42.87	44.57	**44.65**	21.12	25.24	25.61	**26.52**
CS → EN	36.80	41.46	**43.48**	43.18	16.85	20.53	**22.46**	22.06
DE → EN	37.47	40.18	43.15	**43.22**	17.32	21.93	23.38	**24.21**
FR → EN	36.82	40.83	42.85	**43.19**	17.17	21.85	22.75	**23.98**
ALL → EN	37.93	41.34	43.51	**43.56**	18.11	22.39	23.55	**24.19**

Table 10: ROUGE-1 and ROUGE-2 results per language pair and overall for the EN → ALL and ALL → EN tasks.

	EN → ALL		**ALL → EN**	
	ROUGE-L	XNLI	ROUGE-1 / 2 / L	XNLI
E2E	44.54 / 28.57 / 37.40	42.75	43.54 / 23.44 / 33.79	37.58
filtered	44.87 / 28.65 / 37.56	41.77	43.51 / 23.55 / 33.92	37.87

Table 11: Comparison of cross-lingual summarization results obtained with *filtered* and non-filtered training data.

	ROUGE-L	XNLI
EN → EN	36.37	36.87
EN → CS	28.64	31.90
EN → DE	32.83	31.68
EN → FR	39.93	34.40
EN → ALL	34.44	33.71

	ROUGE-L	XNLI
EN → EN	36.37	36.87
CS → EN	26.27	29.00
DE → EN	34.97	32.68
FR → EN	30.39	24.44
ALL → EN	32.00	30.75

Table 12: One-shot prompting results with PaLM 2 per language pair and overall for the EN → ALL and ALL → EN tasks.

putationally expensive, and not suited for studies with many experiments.

F Human Evaluation Study

Figure 3 presents the template and instructions that were used in the human elicitation study. To recruit our participants, we screened their language skills to determine whether they are native speakers, their education level and country of residence as well as origin. In addition, we created a screener test to determine the raters' suitability for the task. In total, we recruited 178 annotators across four languages. Our annotators were paid adequately by our suppliers adhering to the supplier code of conduct.

Tables 14 and 15 show examples of the summaries rated by our participants (gold-standard references or output generated by μPLAN and the E2E systems).

Hill of Tara (https://en.wikipedia.org/wiki/Hill_of_Tara)	
E2E	La colline de Tara (en irlandais : "Cnoc na Teamhrach", « colline des rois ») est une colline du comté de Dublin en République d'Irlande.
μPLAN	La colline de Tara (en irlandais "Cnoc na Teamhrach", « colline des rois ») est une colline située à l'est de Dublin, en Irlande. C'était autrefois le haut lieu des rois d'Irlande. Dans la mythologie celtique irlandaise, elle était la capitale des Tuatha Dé Danann.
Reference	Tara est un site archéologique d'Irlande dans le comté de Meath. Dans la mythologie celtique irlandaise, Tara est la capitale mythique de l'Irlande, située dans la cinquième province de Mide, dans le centre du pays : c'est la « colline des rois » ().

Richelsley (https://en.wikipedia.org/wiki/Richelsley)	
E2E	Le Richelsley est un rocher de l'Eifel, situé dans la commune de Waimes, près de Reichenstein/Monschau, en Belgique.
μPLAN	Le Richelsley est une formation rocheuse située dans la commune de Waimes, dans la province de Liège, en Région wallonne, en Belgique, près de la frontière allemande, à proximité de l'abbaye de Reichenstein/Monschau. Le Richelsley est surtout connu pour sa grande croix de six mètres de haut, qui a été érigée en 1890 par le prêtre Gerhard Joseph Arnoldy, qui travaillait de 1869 à 1914 à Kalterherberg et était le bâtisseur de l'actuelle cathédrale d'Eifel. Le roman "Das Kreuz im Venn" de Clara Viebig a fait du Richelsley un lieu de pèlerinage.
Reference	Le rocher de Richelsley est un important rocher isolé situé à l'est de la Belgique dans les Hautes Fagnes et faisant partie de la commune de Waimes. Sur ce rocher, se dresse une croix appelée Kreuz im Venn.

Table 13: Examples of system output (E2E, μPLAN) and gold-standard reference for the EN → FR setting. Only title and url are shown for input Wikipedia article, for the sake of brevity.

Carduus (https://en.wikipedia.org/wiki/Carduus	
E2E	Die Carduonen ("Carduus") sind eine Pflanzengattung in der Familie der Korbblütler (Asteraceae). Die etwa 90 bis 127 Arten sind fast weltweit verbreitet.
μPLAN	Die Stiele ("Carduus") sind eine Pflanzengattung in der Unterfamilie Carduoideae innerhalb der Familie der Korbblütler (Asteraceae). Die etwa 90 bis 127 Arten sind in den gemäßigten Gebieten der Nordhal.
Reference	Die Ringdisteln ("Carduus") sind eine Pflanzengattung in der Familie der Korbblütler (Asteraceae). Die etwa 90 Arten sind ursprünglich in Eurasien und Afrika verbreitet.

Francesco Satolli (https://en.wikipedia.org/wiki/Francesco_Satolli)	
E2E	Francesco Kardinal Satolli (* 21. Juli 1839 in Marsciano, Provinz Perugia, Italien; † 8. Januar 1910 in Rom) war ein Kardinal der Römischen Kirche.
μPLAN	Francesco Kardinal Satolli (* 21. Juli 1839 in Marsciano; † 8. Januar 1910 in Rom) war ein italienischer Geistlicher und Kurienkardinal.
Reference	Francesco di Paola Kardinal Satolli (* 21. Juli 1839 im Marsciano in der Nähe von Perugia; † 8. Januar 1910 in Rom) war ein italienischer Theologe, Kardinal und der erste Apostolische Delegat in die Vereinigten Staaten.

Table 14: Examples of system output (E2E, μPLAN) and gold-standard reference for the EN → DE setting. Only title and url are shown for input Wikipedia article, for the sake of brevity.

Legendre's formula (https://en.wikipedia.org/wiki/Legendre%27s_formula)

E2E	Legendreova věta je matematické tvrzení z oblasti teorie čísel, které umožňuje počítat součty čísel a rozdělení jejich mocnin: Tedy například pokud "n" je kladné a "p" je velké číslo, pak
μPLAN	Legendrova věta je matematické tvrzení z oblasti teorie čísel, které říká, že pro každé prvočíslo "p" a každé přirozené číslo "n" existuje největší mocninná mocnina "p", která dělí "n". Tedy
Reference	Legendreův vzorec (také De Polignacův vzorec) dovoluje vypočítat nejvyšší exponent u prvočísla formula_1, kde formula_1 umocněné na tento exponent ještě dělí číslo formula_3 (faktoriál přirozeného čísla formula_4). Jedná se v podstatě o výpočet p-adické valuace čísla formula_3.

Stack overflow (https://en.wikipedia.org/wiki/Stack_overflow)

E2E	Přetečení na zásobníku () je v informatice označení pro speciální případ přetečení na zásobníku. Přetečení na zásobníku nastává v důsledku umělého vyčerpání úložného prostoru programem. V důsledku toho dojde k přepsání obsahu na zásobník, což způsobí vznik segmentation fault.
μPLAN	Přetečení na zásobníku () je v informatice situace, kdy je v počítači program, který používá zásobník operační paměti, více než je v ní uložen. Přetečení na zásobníku vzniká většinou v důsledku dlouhých procesů, které konají nadbytečné údaje. Nejčastější příčinou přetečení na zásobníku je nekonečná rekurze. V programech s více vlákny je přetečení na zásobníku běžné.
Reference	Přetečení zásobníku je situace v programování, kdy dojde k pokusu uložit na zásobník volání více dat, než kolik se tam vejde. Velikost tohoto zásobníku je obvykle předem dána při startu programu v závislosti na architektuře systému, překladači, množství volné paměti atp. Když se program pokusí posunout vrchol zásobníku mimo vymezenou paměť, mluvíme o přetečení zásobníku. To má obvykle za následek pád programu.

Table 15: Examples of system output (E2E, μPLAN) and gold-standard reference for the EN → CZ setting. Only title and url are shown for input Wikipedia article, for the sake of brevity.

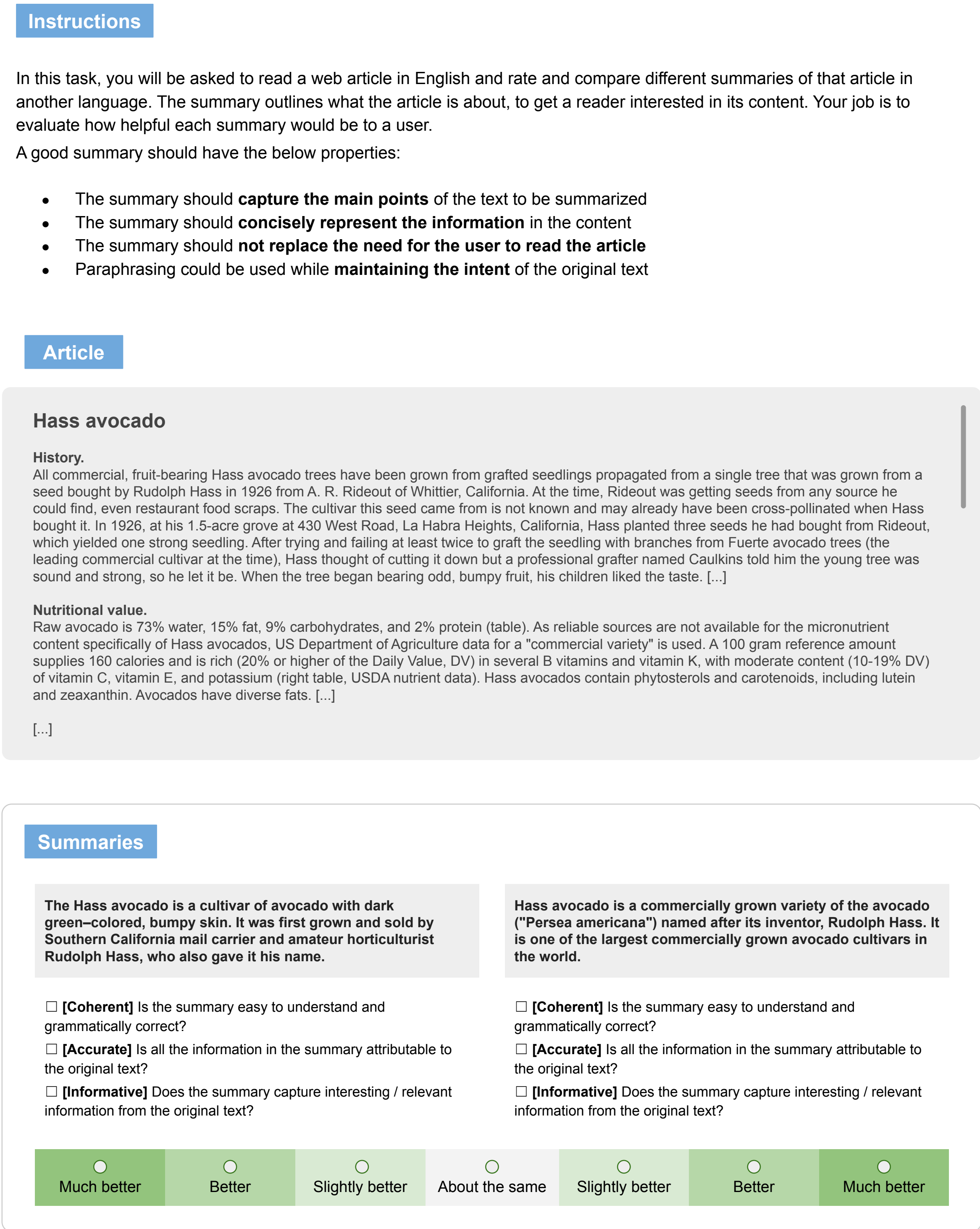

Figure 3: The template and instructions that were used in the human elicitation study.