Chinese-to-English Cross-Lingual Summarisation: Datasets, Models, and Automatic Content Assessment via Natural Language Inference

Huajian Zhang

Master of Science
Artificial Intelligence
School of Informatics
University of Edinburgh
2022

Abstract

In the Cross-lingual Summarisation (CLS) task a summary in one language is generated from a corresponding document in a different language. There are two main problems limiting the research on CLS, one is the absence of large-scale high-quality parallel corpus to serve as supervised CLS datasets, and the other is the lack of good automatic evaluation metric. To make available CLS datasets, Perez-Beltrachini and Lapata [52] proposed XWikis, a CLS corpus including four European languages. This project extends XWikis with Chinese, a distant language, and evaluates a wide range of models on the cross-lingual summarisation task (Chinese-English direction), including translate-then-summarise, supervised, and zero/few-shot scenarios. For evaluation metrics, we propose three strategies to leverage existing NLI models to evaluate cross-lingual document-summary pairs at the sentence and sub-sentence levels. The conducted experiments reveal that predictions from NLI models correlate with human judgements on content adequacy. Furthermore, we find that sub-sentence segmentation of summary sentences can further help the model to capture entailment relations between document and summary sentences.

Research Ethics Approval

This project was planned in accordance with the Informatics Research Ethics policy. It did not involve any aspects that required approval from the Informatics Research Ethics committee.

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

(*Huajian Zhang*)

Acknowledgements

I would like to express my gratitude to my supervisor, Laura Perez-Beltrachini, for the invaluable advice and feedback throughout the dissertation. I really benefit a lot from her thinking, troubleshooting, tens of emails and the meeting every week.

I would like to thank two members of EdinburghNLP group, one for his advice about mBART, and the other one for her careful annotation.

I would like to thank my friends, Xinxi Lyu, Ang Li, Chenfeng Shan, Shutong Feng, Guangyu Li, and Xiaoyu Jiang. Their voices accompanied me in all the hard times this year.

I would like to thank my father. His optimism and humour across half the earth, encourage me as ever.

Most importantly, I save the last part for my mother. I have not heard your voice for 37 months and I do not know when I can see your face again. However, your face, your voice, and your unconditional love, always lead me and support me, in this ever-changing world.

I dedicate this dissertation to you.

Table of Contents

Chapter 1

Introduction

Cross-lingual summarisation (CLS) aims at generating a summary from a document where the summary and the document come from different languages. With the rapid globalisation, individuals are more exposed to information in different languages, including news, product descriptions, advertisements, encyclopedia, and so on. This trend suggests the significant practical values of CLS. However, the research about CLS receives limited attention due to the absence of large-scale high-quality parallel corpus [33] and a faithful automatic evaluation metric [43].

Motivated by the lack of data resources for CLS, Perez-Beltrachini and Lapata [52] propose the XWikis corpus, which consists of four European languages: Czech, English, French, and German. They collect data from Wikipedia and create (document, summary) pairs. Given any two Wikipedia titles describing the same entity but in different languages, e.g., Olive Oil (English) and Huile d'Olive (French), the leading section in one language stands as a summary and the main section in another language as the corresponding document. For further research on CLS, they set up a benchmark based on XWikis and provide baseline performance with pre-trained language models on the All-to-En direction for different learning scenarios. The All-to-En direction means that the input document of the model is in any of three languages and the output summary in English. The different scenarios include supervised, zero-shot, and few-shot CLS, where the supervised model is trained on the full training set, and zero/few-shot model only receives no/a few data as the training set.

Despite not being members of the same language family, the four European languages chosen by XWikis still have cognates because of geographic circumstances [42], which could facilitate transfer, especially in the zero/few-shot setting. To enlarge the scope of the XWikis benchmark, this project extends XWikis with Chinese, a more

distant language than existing languages in XWikis. We create (document, summary) pairs for Chinese-to-all and all-to-Chinese directions. Following [52], we conduct experiments on the CLS task on Chinese-to-English pairs in the supervised, few-shot, and zero-shot scenarios.

Another challenge in summarisation is the automatic evaluation of model outputs. Intuitively, the generated summaries should be factually consistent with the input document. However, ROUGE score [38], the most widely used metric for text summarisation, only focuses on the lexical n-gram overlap between generated and referenced summaries. This leads to two problems (1) Summaries can use synonyms and paraphrasing. Both convey the same meaning but with different forms, which cannot be captured by ROUGE. (2) Summaries can have a high token overlap with reference but do not convey the same meaning. Some studies [39] [24] also find ROUGE score does not coincide with human judgements.

Previous works [18][43][27] have proved that natural language inference (NLI) models can be leveraged to automatically detect inconsistency between summaries and documents. Given two sentences, premise and hypothesis, the NLI model is aimed to classify the hypothesis as either entailed by, neutral, or contradicting a premise sentence. Laban et al. [27] have shown that the NLI model can correlate positively with human judgements in English-English document-summary pairs.

In this project, we adopt two state-of-art multi-lingual (which can support many languages but the premise and hypothesis are in the same language, e.g. French-French) NLI models and adapt them for the CLS task in which the premise is from a document in one language and the hypothesis is from a summary in a different language. We investigate their effectiveness aiming to automatically check the content adequacy (1) of the extracted cross-lingual document-summary pairs in XWikis, and (2) model generated summaries. In addition, previous work only focuses on one-to-one relationships between one summary sentence and one document sentence. Naturally, one summary sentence can aggregate information from several document sentences. To access the alignment, we split the discussion into one-to-one and many-to-one relationships. We propose three strategies to model many-to-one relationships and further apply sub-sentence segmentation on summary sentences to gain a finer granularity view. The experiments show that predictions from NLI models correlates with human judgement on our XWikis domain. Experiments also show that the sub-sentence segmentation can help the NLI model to better capture information and thus obtain a better correlations.

The structure of the dissertation is as follows: Chapter 2 presents the background of

CLS and the context of the project. Chapter 3 displays how to collect Chinese data from Wikipedia, specific considerations for Chinese texts, and conducts experiments to verify the quality of created dataset (faithfulness and abstractiveness) by automatic metrics and human evaluation. Chapter 4 introduces the neural CLS models we assess as well as baselines and shows their performance in supervised, few-shot, and zero-shot scenarios. We compare and analyze results obtained for Chinese-to-English with other language pairs. Chapter 5 explores how to leverage out-of-box NLI models for the evaluation of cross-lingual document-summary pairs. Finally, Chapter 6 provides conclusions and a discussion of future work.

Chapter 2

Background: Cross-lingual Summarisation

This chapter will display the context of this project, cross-lingual summarisation. We will first briefly introduce the task formulation of CLS, then go through three aspects of this field, including methods, datasets, and evaluation metrics. In each aspect, we also provide the existing shortages or problems, which also motivate our project.

2.1 Task Formulation

Automatic text summarisation, is aimed to create a summary S from input text D, where S is shorter but still conveys the key points of D. In general, there are two categories for text summarisation: extractive summarisation and abstractive summarisation [46]. Extractive summarisation seeks to construct a summary S by extracting a portion of the input text D while abstractive summarisation generates words that might not be found in D [9]. In this project, we only focus on abstractive summarisation.

The goal of cross-lingual summarisation is similar to normal text summarisation, except that the summary and document are from two different languages [62]. Formally, given a document D^{src} from a source language, the goal of CLS is to produce the corresponding summary S^{tgt} as sequence of tokens $S^{tgt} = (S_1^{tgt} \cdots S_n^{tgt})$ in a target language. The conditional distribution of CLS models is:

$$p_\theta(S|D) = \prod_{t=1}^{n} p_\theta(S_t^{tgt}|D^{src}, S_{1:t-1}^{tgt}) \tag{2.1}$$

where θ represents model parameters.

Dataset	Lang	Pairs	SumL	DocL
Global Voices	15	229	51	359
WikiLingua	18	45,783	39	391
XWikis (comp.)	4	213,911	77	945
XWikis (para.)	4	7,000	76	972

Table 2.1: Number of languages (Lang), average number of document-summary pairs (Pairs), average summary (SumL) and document (DocL) length in terms of number of tokens [52]. comp. refers to the training/validation set while para. refers to the test set.

2.2 Datasets

Different with most NLP tasks, it might be expensive and unrealistic to manually create CLS data. Since for N languages, there are $\frac{N!}{(N-2)!}$ possible directions for $D_{src} \rightarrow S_{tgt}$. Thus, most existing datasets for the CLS task rely on off-the-shelf resources. They can be classified into two types, synthetic dataset and website dataset [63].

A synthetic dataset is constructed using an existing machine translation (MT) dataset resource. The data pairs are built by pairing the original document and translated summaries. As mentioned above, information might be lost or tampered with during translation. Some researchers attempt to employ several methods to post-select high-quality samples.

Website datasets usually benefit from websites supplied with multi-lingual versions. Such websites, provide texts in different languages which are translated by professional translators. As a result, it is convenient to align them to form a CLS dataset.

In this project, we focus on website datasets since they usually contain higher quality data than synthetic datasets and enable more directions for CLS task naturally. Specifically, we consider three key points of a website dataset: (1) the content of (summary, document) pairs, (2) the authors (annotators) of the contents, and (3) the quality of data (diversity, abstractness and faithful). We introduce two CLS website datasets, Global Voices and Wiki-lingual

Global Voices [49]: Nguyen and Daume III construct Global Voices, a multilingual dataset for CLS task. This dataset collect documents from the same name website [1], which aims to translate news about voices in the social network. Translators in the website are volunteers. Then, researchers collect the descriptions of news from the social network (e.g. Twitter and Facebook) as the corresponding summaries. However,

[1] https://globalvoices.org/

the motivation of these descriptions is to not summarize but to draw user clicks, thus they further crowd-source a set of summaries in English. Since the summaries is manually created, they are high quality but rare. The English-only summaries also forbid comparison among other languages.

Wiki-lingual [28]: Ladhak et al. extract CLS data pairs from WikiHow, which includes 18 languages. This site contains a high-quality resource written by human authors in the domain of how-to guides. Each document provides different methods with multiple steps to complete a procedural task, and the corresponding summary describes each step with one-sentence instruction. The dataset includes a diverse set of topics but the summaries are relatively short. Each summary sentence is a short version of several corresponding document sentences. As for the human authors, they are fluent in English and the target language. After human translation, their works also need to be checked by WikiHow's team.

XWikis differs from two existing datasets in terms of size and summarisation task. From table 2.1, XWikis consists of more document summary pairs. Specifically, the input documents are twice more as the other dataset. Such long texts create a challenge for current neural summarisation methods since most of them are trained on relatively short texts and thus struggle to represent multi-paragraph texts [52]. As for the number of languages, existing XWikis contains four European languages. However, the data creation approach potentially enables for the extension to additional languages. In this project, we extend XWikis with Chinese, a more distant language.

2.3 Methods

There are two paradigms of the current models on CLS problem, pipeline [32][37] and end-to-end (E2E) [68]. Most early-stage researchers concentrate on pipeline approaches [63]. Intuitively, they decompose the CLS task into the machine translation (MT) task and machine summarisation (MS) tasks.

One issue with pipeline approaches, according to Wan et al.[62], is that the performance of such methods is strictly constrained by each subtask. In pipeline approaches, the output of the first subtask is the input of the second subtask. The error from the previous subtask, either MT or MS, will catastrophically propagate to the next subtask, resulting in more severe error. Owing in part to the popularity of neural network models, end-to-end methods are proposed to alleviate this issue.

The E2E methods can be further divided into several paradigms. In this project,

we only focus on pre-trained multi-lingual language models, which provide language modeling in different languages and cross-lingual shared representation, such as Multilingual BERT(mBERT) [13], Multilingual BART (mBART)[41], mBART50 [60] and Multilingual T5 [65]. Such models are usually first trained on a language modeling task on a large multi-lingual corpus. They are expected to capture the implicit linguistic features or common sense knowledge [11, 67] and then applied to downstream tasks. This project applies mBART and mBart50 to the CLS task for comparison.

2.4 Evaluation Metrics

2.4.1 Lexical Overlap Text Evaluation

CLS shares the same evaluation metrics with the text summarisation task. For text summarisation, manual evaluation of models is costly and intractable [47]. Many researchers make effort to develop automatic metrics, which support fast and cheap evaluation for summarisation models.

The most widely used automatic matrice is the ROUGE package (Recall-Oriented Understudy for Gisting Evaluation) [38]. It provides a set of metrics based on token overlap between generated and reference summaries. Such overlap can be measured based on n-grams (ROUGE-N), skip-grams (ROUGE-S), or the longest common subsequence (ROUGE-L) of tokens. These metrics assume that the more shared tokens indicate the more similarity between sequences. However, this assumption is too strong. Since the score is totally computing by token match, ROUGE cannot support synonyms and paraphrasing. This leads to arguments about its effectiveness. Graham's experiments [20] show ROUGE is of statistical correlations with human judgment, where [24] ROUGE is only moderate correlation with human evaluation for extractive models and weak correlation for abstractive models.

2.4.2 Model-based Text Evaluation

With the emergence of large pre-trained language models such as BERT [14], researchers investigate how to leverage them to measure similarity between two summaries based on similarity distribution. BERTScore [66] leverages word embeddings from BERT and attempts to align the tokens from generated summaries and references pairwise. BLEURT [58] further fine-tunes BERT to predict human judgment scores on a large

synthetic dataset. Clark et al. [7] explore sentence-level alignment with contextualized sentence embeddings.

In the field of summarisation, model-based metrics can be divided into three approaches [2]: model-based approaches, question answering (QA) based approaches and natural language inference (NLI) based approaches.

Model-based approach, as its name suggests, is to train a model to detect factual errors in the summary [12]. Kryscinski et al. [25] propose a weakly-supervised, model-based approach for verifying consistency in summarisation. They generate the training samples by applying rule-based transformations on documents.

QA-based approaches contain three steps: question generation, question answering using the document and the summary, and matching answers from documents and summaries. The key difference among these methods lies in the first step. Researchers generate questions using the summary [16], document [56], or combined both together [55]. A summary is considered consistent if few or no questions have differing answers from the document.

Given two sentences, premise and hypothesis, NLI (also referred to as textual entailment), aims to classify the hypothesis as either entailed by, neutral, or contradicting a premise sentence. When applying on NLI model to evaluate text summarisation, a document serves as the premise and a summary serves as the hypothesis. A summary is considered factual if all its sentences are entailed by the document. Falke et al. [18] first attempt to re-rank the generated summaries by the out-of-box NLI-based model. They score each summary sentence, based on its entailment probability given the document, and then select the sentences with top scores to form a new summary. Maynez et al. [43] find NLI-based models have a better correlation with human evaluation than standard metrics. Recently, Laban et al. [27] show that NLI-based models can successfully be used for summarisation evaluation, if at a suitable granularity. They find that past work suffered from misuse of the NLI-based model. The input of the original NLI model is at the sentence level. However, when the model is applied to evaluating the summary, its input is at the paragraph level.

The above works only focus on the English domain and the datasets they evaluated are rather extractive but not abstractive. In this project, we investigate how to apply the NLI model in the cross-lingual setting and then evaluate the model on XWikis, a more complicated and abstractive dataset.

Chapter 3

Extending the XWikis corpus with Chinese

In this chapter, we describe the dataset creation pipeline we follow to extend the XWikis corpus with the Chinese. It consists of two parts, data collection and dataset evaluation.

3.1 Dataset Creation

A Wikipedia article typically contains two sections: the lead section, which offers a succinct introduction, and the body section, which offers the pertinent details. We can consider the lead section and the body section of an article to be a document-summary pair. Wikipedia provides mutli-language versions of a given article's title, for instance, Huile d'Olive (French) and Olive Oil (English), since they are written collaboratively by a large number of anonymous volunteers from different language communities. An article available in one language can be aligned with articles which describe the same title but in other languages. Based on these two observations, Perez-Beltrachini and Lapata [52] proposed the XWikis corpus, which contains cross-lingual document-summary pairs. Concretely, the lead section of an article in language X will serve as the summary and the body section of a corresponding article in language Y as the document, creating document summary pairs (Doc_X, Sum_Y) in the dataset $\mathcal{D}_{\mathcal{X}\rightarrow\mathcal{Y}}$. The entire data creation pipeline is shown in Figure 3.1.

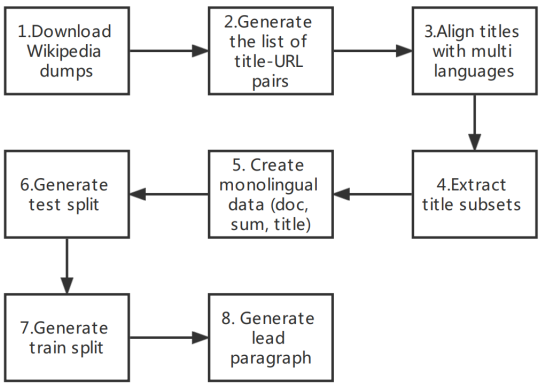

Figure 3.1: Workflow for dataset setup

Steps 1-4. We download [1] Wikipedia dumps English (en), German (de), French (fr), Czech (cs) and Chinese (zh) from Wikimedia Downloads [2]. Wikipedia also provides the Wikipedia Inter-language Links table [3]. In the following steps, we use the table to align Wikipedia titles across different languages. Then, we generate a list of title-URL pairs for each language (step2) and align such titles in all languages (step3), as shown in the Figure 3.1.[4] We revise the file containing aligned titles to obtain the desired cross-lingual (T_X, T_Y) or mono-lingual (T_X) titles subsets (step4). That is, to create the French-English dataset $\mathcal{D}_{fr \to en}$, we only select the titles that appear in French and English Wikipedia at the same time.

Steps 5. Based on the subsets of titles from the previous step, we scan all pages from Wikipedia dump to obtain zh texts (document, summary, title) (step5). The next step is to obtain the cross-lingual document and summary (Doc_X, Sum_Y) data pairs. We pair zh texts with the other four languages by using the aligned titles subsets (T_X, T_Y). Here we retain document summary pairs (Doc_X, Sum_Y) whose documents and summaries match the length constraints. That is, the collected pairs should satisfy a) the length of the document should be limited to be between 250 and 5,000 words and b) the length of the summary should be between 20 and 400 words. Such constraints aim to keep each document-summary not too long or short. Another implicit reason is to assure certain content overlap between documents and summaries; the content for a Wikipedia title in two different languages might not be fully aligned since the editors from different language communities might hold different interests on the same title.

[1] To keep consistent with previous research, we select all files collected on 20/June/2020
[2] https://dumps.wikimedia.org/
[3] https://en.wikipedia.org/wiki/Help:Interlanguage_links
[4] These steps, including the Chinese language, were already available in the XWikis corpus.

	en	de	fr	cs	zh
en		425,279	468,670	148,519	135,674
de	376,803		252,026	109,467	103,044
fr	312,408	213,425		91,175	99,301
cs	64,310	53,275	51,578		32,588
zh	75,524	73,969	81,847	43,281	

Table 3.1: Total number of document-summary pairs in the XWikis corpus considering all language pairs and directions. Each table cell corresponds to a cross-lingual dataset $\mathsf{D}_{\mathcal{X}\to\mathcal{Y}}$.

Chinese texts do not use space as separator between words/characters. There are some Chinese word segmentation models [22] but they are not perfect, segment at different granularities, and most importantly, the are slow to run on a large corpus as Wikipedia. Therefore, we take the length constraints for Chinese texts based on character level. Specially, to approximate the difference between Chinese characters and words, the above constraints are enlarged by a multiply factor of 1.2 for all Chinese texts. The reason for introducing this factor and how to deal with Chinese text is discussed in Appendix A.1. Table 3.1 shows the number of samples in each set $\mathcal{D}_{X\to Y}$ created following above procedure. We add all zh language subsets (columns/rows) in this project, the other language subsets were available in the XWikis corpus [52].

Steps 6-8. For better comparison in the evaluation stage, we align titles across five languages to obtain the test set (step 6). In other words, each Wikipedia title that is part of the test set can be found in each of the five languages in Wikipedia. All titles in each dataset $\mathcal{D}_{X\to Y}$ minus those selected for test are split into training (95%) and validation (5%) sets (step 7). We call the test set as XWikis-parallel and training/validation split as XWikis-comparable. We further refine the extracted summaries to take only the lead (first) paragraph (adding more sentences if necessary to complete a minimum length). The idea is to take the discussion of the most general content related to the Wikipedia title (step 8).

	XWikis (comp)				XWikis (para)	
	de	fr	cs	zh	para	zh
Words/Doc	906	1040	890	769	972	890
Sents/Doc	41	38	42	36	42	40
Sections/Doc	5	7	6	6	6	6
Words/Sum	56	59	65	61	61	70
Sents/Sum	3	2	3	3	3	3
Aspects	253,425	248,561	65,151	75,796	9,283	11,722
Coverage	65.53	72.23	55.97	55.05	65.41	62.21
Density	1.23	1.51	0.99	0.94	1.23	1.14
Compression	17.44	20.16	15.12	15.17	18.35	10.89
% new unigrams	33.30	26.77	42.29	39.77	33.25	33.38
bigrams	80.70	73.19	85.17	84.37	79.51	80.01
trigrams	93.60	90.25	95.19	95.32	93.17	93.44
4-grams	97.98	95.68	97.98	98.13	97.11	97.17
LEAD	19.09	23.51	20.21	12.24	20.88	13.19
EXT-ORACLE	24.59	28.38	24.25	16.48	25.95	17.43

Table 3.2: XWikis statistics (number of words and sentences per document (/Doc) and summary (/Sum)) and task characterisation metrics. Statistics for de/fr/cs-en are taken from [52] and zh-en are computed in this project.

3.2 Dataset Evaluation

In this section, we evaluate the created document-summary pairs to access their characteristics and compare them with the existing document-summary pairs across the XWikis corpus. We follow [52] to carry out the evaluation based on both automatic metrics and human judgement. Table 3.2 provides statistics for the XWikis-comparable (training/validation) and XWikis-parallel (test set) for all language subsets.

3.2.1 Automatic Metrics

Size. The top 5 rows in the table report the size of document-summary pairs for zh document-summary pairs. The documents contain in average 769 words, 36 sentences and 6 sections. The summaries contain in average 61 words and 3 sentences. We count the number of words for Chinese texts after word segmentation using *Hanlp*

[22] package. This number might be slightly changed due to segmentation tools with different granularity. Generally, Chinese documents are relatively shorter than those in the other four languages but summaries are closer.

Diversity. To approximately gauge the diversity of the corpus, we report the number of sections per document in row 3, which are computed by counting the number of HTML tags such as h2 and h3. In row 6, we also provide aspects, the number of **distinct** section titles in the whole set. For zh these numbers are closer to cs, much lower than de and fr, due to the fewer number of document summary pairs.

Generally, zh documents indeed contain multi-topic and these two numbers (section per doc and aspects) match other four languages. This feature keeps XWikis challenging because a summarisation model needs to learn to decide which sections are more summary-worthy and to extract content from multiple sections in the documents.

Level of Abstraction. Another important factor for an abstractive summarisation dataset is the level of abstractness. Here we use two sets of automatic metrics [21, 45] to measure the abstractiveness of the dataset.

The first set of metrics proposed by Grusky et al. [21] includes: coverage, density and compression. The first two metrics are designed to quantify the extent to which content overlaps between the document and the summary. They greedily search all shared token sequences and store the longest overlap at each search step. They use **coverage** to measure the ratio of shared sequence over the whole summary. **Density** describes in average the length of the shared token sequence in summary. The third metric **compression**, provides the ratio between the content from the article and the summary. As shown in Table 3.2, the coverage is high but the density is low. This suggests that more than half of the token sequences in the summary can be found in the document, however, the average length of such shared sequences is quite small, which means the shared sequences are likely to be common-used phrases.

The second set of metrics proposed by Narayan et al. [45], measures the percentage of n-grams occurring in the summary but not in the document. Similarly, the percentage of unigram is low, matching the high number of coverage, and it increases steadily for higher n-grams, matching the low density.

We also evaluate the performance of two extractive summarisation models, **LEAD** and **EXT-ORACLE**, on the validation set. We assume that these two extractive models should obtain good results if the dataset is extractive in nature. LEAD generates a

Dataset	de $\rightarrow$ *en*	fr $\rightarrow$ *en*	cs $\rightarrow$ *en*	zh $\rightarrow$ *en*
Overall	71.7 %	96.6 %	73.3 %	85.7%
Sentence	66.2 %	77.4 %	60.5 %	71.5%

Table 3.3: Proportion of **YES** answers given to questions of overall summary and sentence adequacy. Results for de/fr/cs-en are taken from [52].

summary by simply copying the first N tokens from the document, where N equals the length of the reference summaries. EXT-ORACLE is to generate a summary by selecting the portion of sentences to maximize ROUGE-2 with respect to the reference [5]. Intuitively, when the salient information concentrates on the first several sentences of the document, or so-called lead bias, LEAD performs well. And when the summarisation task is extractive, EXT-ORACLE should have a good performance [52]. Rouge-L is used as the performance indicator.

The last two rows of Table 3.2 provide the results for the two extractive models. LEAD is below EXT-ORACLE by around 4 ROUGE-L points, suggesting no lead bias in the summaries. The performance of EXT-ORACLE is also not good. One reason can be that in the reference summaries, each sentence aggregates information from multiple sentences across the document. Another reason could be related to the high ratio of new paraphrases appearing in summary, which cannot be captured by an extractive model. The performance of these two models is extremely weak in Chinese. We further explore this phenomenon by translating target summaries and generated summaries (by LEAD/EXT-ORACLE) on the test set to English. They achieve 17.14 and 24.57, respectively, which can be comparable with the other three languages. One explanation is that evaluating the ROUGE score on the other three languages is after stemming, which does not support Chinese, leading to the performance drop. There potentially exists other reasons such as noise introduced by the Chinese word segmentation tool in the data pre-processing stage. We leave this for future work.

3.2.2 Human Evaluation

In addition to automatic evaluation, we conduct a human evaluation study to rate the quality of cross-lingual pairs (Doc_X, Sum_Y). Specifically, we examine the assumption that given a pair of aligned titles (T_X, T_Y) , a lead paragraph in language Y can serve as the summary and be supported by the document in language X.

We recruit three bilingual (Chinese and English) judges and selected 20 data in-

stances from the validation set[5]. We prepare two types of questions for participants. The first is about the overall judgment: **Does the summary provide a general overview of the Wikipedia title?**. We further ask questions for each sentence in the summary: **Does the sentence contain facts that are supported by the document?** The participants need to answer with yes/no to the questions. In the Appendix B we provide a screenshot of the interface used to elicit such annotations.

Table 3.3 shows the proportion of yes answers given by three judges for the zh-en direction. Generally, three judges view the summary as an acceptable overview of the document. As for the more fine-grained sentence-based judgements, 71.5% of the summary sentences are as supported by the document in zh-en, which is close to the other three directions. We used Fleiss's Kappa to measure inter-annotator agreement across three judges. This was 0.48 for German-English speakers, 0.55 for French-English, and 0.59 for Czech-English, and 0.47 for Chinese-English.

[5]The bilingual speakers were the author of the thesis, a college student, and a member of the EdinburghNLP group.

Chapter 4

Benchmarking Models on the Chinese-to-English Summarisation Task

In this chapter, we benchmark different cross-lingual abstractive summarisation models on the task of Chinese-to-English summarisation. We benchmark all model variants proposed in [52] and analyze the results comparing them with those obtained for other language pairs (e.g., French-to-English).

4.1 Comparison Models

In this section, we introduce the cross-lingual abstractive summarisation models we evaluate. We also evaluate a range of extractive summarisation models for comparison [52]. [1]

4.1.1 Extractive Methods

LEAD. This method holds the assumption that the first several sentences of the document can serve as a good summary. In this project, we utilise this method to directly copy the first K tokens of the input document to form the summary, where K equals the length of the reference.

[1] This chapter is constructed using part of the descriptions of the models provided in the student's Informatics Project Proposal.

EXT-ORACLE [44]. This method view CLS as a classification task for sequence. The model will go through each sentence in the document and then determine whether the current sentence should be included in the summary or not. The such binary decision will be made by maximizing the ROUGE-2 score with respect to the reference summary. We limited the upper bound of generated text following Nallapati et al.'s [5] procedure.

LEXRANK [17]. This method represents text as a graph. Each sentence is a vertex in the graph, and they are connected based on the similarity(TF-IDF) and ranked by the number of neighbors. This method assumes that a sentence with the greater number of neighbors contains more salient information. The generated summary is extracted from the first K tokens of the top sentences, where K equals the length of the reference.

4.1.2 Abstractive Models

Multilingual BART (mBART)[41] is an encoder-decoder model extended from Bart [34] with multiple languages. The pre-training stage of mBART contains two tasks, one is mono-lingual masked language modeling on a 25 languages corpora, and the other is sentence order predication. mBART can provide cross-lingual representations and experiments demonstrate its gains in performance across a wide variety of machine translation tasks.

mBART50 [60] is pre-trained on the same task as mBART but with 50 languages, then further fine-tuned on machine translation task with multiple language pairs jointly. For the machine translation fine-tuning stage, the mBART50 model is not trained from one language to another but from many languages to many other languages, which can enable languages from the similar family to benefit each other [3].

To explore the potential function of large pre-trained language models and their performance under a weak supervised signal, we also evaluate two abstractive models in the zero- and few-shot scenarios. Zero-shot learning and few-shot learning are two problems set up in the field of machine learning. The goal of few-shot learning is that models can predict the correct label of input with only being exposed to a small number (usually hundreds) of examples when training. For zero-shot learning, the model needs to do the same thing when no example belonging to that label is available in the training stage. Both are designed for evaluating model performance with limited data. In the zero-shot scenario, a mono-lingual English model is used for the Chinese-to-English summarisation task. In [52], it was find that the parameters of the mono-lingual

summariser can transfer to other languages with promising performance. In the few-shot scenario, we also rely on the mono-lingual English summariser, which is further fine-tuned on a small number of Chinese-to-English examples (Doc_{zh}, Sum_{en}). To this end, [52] apply two techniques suitable for the few-shot task, Model Agnostic Meta-Learning (MAML; [19]) algorithm and Cross-View Training (CVT; [8]).

MAML consists of nested optimisation iterations. For the inner loop, the model will sample datapoints for each sub-task and then calculate the corresponding new parameters θ'. For the outer loop, the model parameters θ are updated according to each task loss on parameters θ' using another sampled data points. Such settings aims to force the model to learn to learn [54]. Thus, when testing, the model can be quickly adapted to a new task with the small amount of data points.

CVT considers how to leverage the unlabeled data during the training stage. The model includes the primary module and auxiliary modules with different views of the input. The different views are constructed by restricting the feature embedding from the encoder in a different way. For example, views can be built by masking words in different locations. The difference between the output of the primary module and the outputs of auxiliary modules will perform as the supervised signal to train the model. In other words, auxiliary modules should be trained to agree with the primary module. Since all prediction modules share the same encoder, this training technique can drive the encoder to extract better embedding to decrease loss from auxiliary modules. In our project, we assume we will find documents in Chinese without aligned documents in English. They will be regarded as unlabeled documents Doc_{zh} where the corresponding summary Sum_{en} is generated by the current model. The different views of Doc_{zh} will be created by taking the output of different layers from the encoder. To be consistent with [52], we will take the hidden states of the encoder at layers 6 and 11, to form two views. We hope they can provide two levels of abstraction of Doc_{zh}.

4.2 Experimental Setup

4.2.1 Datasets and Splits

We work with the $\mathcal{D}_{zh \to en}$ directions of our XWikis corpus and evaluate model performance on the XWikis parallel test set.

We use the training (95%) and validation (5%) splits from $\mathcal{D}_{zh \to en}$ (c.f. Section 3.1 step 6-9). To train an English mono-lingual summariser (used in the zero- and few-shot

	All	800 ParaLexRank.600	
$D_{en \to en}$	55.16	51.83	51.88
$\mathcal{D}_{de \to en}$	52.05	48.64	48.60
$\mathcal{D}_{fr \to en}$	56.05	51.78	51.86
$\mathcal{D}_{cs \to en}$	53.37	50.20	50.47
$\mathcal{D}_{zh \to en}$	41.38	39.73	39.44

Table 4.1: ROUGE-L recall for source document against reference monolingual summary, where the first 800 tokens, the first 600 tokens, and all tokens of documents are extracted by using LEXRANK model.

scenarios), we need a monol-ingual dataset $\mathcal{D}_{en \to en}$. We re-use that created by [52]. The dataset encompasess a set of Wikipedia titles disjoint from those appearing in the datasets in the XWikis corpus. It has 300,000 instances with 90/5/5 percent of instances in training/validation/test subsets. It follows similar characteristics to the data in the XWikis corpus with an average document and summary length of 884 and 70 tokens, respectively.

4.2.2 Data Pre-process

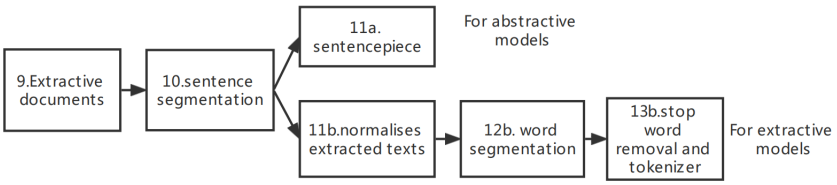

Figure 4.1: Data pre-processing

Due to the hardware limitation that a very long document cannot fit into our GPU memory, we carry out an initial extractive step [40] [52]. We apply the extractive method LEXRANK to represent each paragraph as vectors of TF-IDF values and select the top-ranked ones up to a budget of 600 tokens. We set up an experiment to check the sensitivity of the number of tokens, as shown in Table 4.1. We report ROUGE-L Recall score of mono-lingual document-summary pairs. The results show the extractive step does not ignore too much salient content, though the overall results are lower for the zh-en direction compared with others, which is a quite similar phenomenon mentioned in section 3.2.1.

For the other data processing, as shown in figure 4.1, we apply Stanford API to segment documents and summaries into sentences. For extractive methods, we first normalize all data. For English, we use space as a separator and for Chinese, we apply *Hanlp* to segment Chinese sentences into words to remove the stop words and tokenized. For abstractive methods, we apply sentencepiece [26] to split sentences into sub-tokens. Sentencepiece is a language-independent tokenizer, supporting subword segmentation [59]. Subword segmentation aims to segment tokens into sub-word units so that model can learn morphological knowledge and be applied on unseen tokens. For example, **scientist** may be split into two sub-tokens, **scient** and **ist**. The language model can learn the semantics of each sub-token and can guess the meaning of an unseen token by using the knowledge of sub-tokens the unseen token contains. In this project, we re-use the sentenpiece models provided by mBART and mBART50.

Other experiment settings including the number of training updates, learning rate, optimizer, and so on, keep the same as [52].

4.2.3 Model Settings in Different Scenarios

Extractive. We apply the extractive models described in Section 4.1.1. Since the extractive model cannot support cross-lingual data, we apply the summarize-then-translate [63] paradigm. After extractive models generate the summaries, we employ Google cloud translation API to translate the texts to the target language.

Supervised. We will fine-tune mBART and mBART50 model on the $\mathcal{D}_{zh \to en}$ dataset in a supervise fashion. To validate the model performance on zero- and few-shot scenarios, we also train an English mono-lingual summariser on the separate English dataset $\mathcal{D}_{en \to en}$.

Translate-then-summarize. We will apply the translate-then-summarize pipeline approach as another baseline for abstractive methods. We will first translate the Chinese input documents into English and then use a mono-lingual English summariser.

Zero-shot. The mono-lingual English summariser is directly applied to summarise Chinese documents into English.

Few-shot. These models are based on the mono-lingual English summariser subsequently adapted to the cross-lingual task with a small set of Chinese-to-English

examples $\mathcal{S}_{zh \to en}$. We present experiments with mBART and mBART50 pre-trained models. We evaluate the three few-shot variants of [52] (described in Section 4.1.2). LF-MAML is the light-weight First Order MAML version, FT is a fine-tuned version where only cross-attention and layer normalisation layers are fine-tuned, and CVT incorporates additional unlabelled instances into the adaptation step. We also follow two settings as [52] with $|\mathcal{S}_{zh \to en}|$ being 300 and 1,000 few instances. Note that in each case we take 1/3 for validation, and the rest for training. For CVT, we generate two views, Doc^{m}_{zh} and Doc^{u}_{zh}, for each input document Doc_{zh} in $\mathcal{S}_{zh \to en}$ by taking a middle encoder representation (Doc^{m}_{zh} the hidden states at layer 6) and another by taking an upper encoder representation (Doc^{u}_{zh} the hidden states at layer 11). Intuitively, these provide different levels of abstraction from the input document.

4.3 Results and Analysis

In this section, we compare the cross-lingual summarisation results of zh-en with the other three language pairs. The results are shown in Table 4.2.

Does Zero/Few-Shot Work? Zero-shot performs worse for zh-en. Both mBART and mBART50 obtain relative low scores. Especially, mBART, which has not seen any cross-language alignment during the pre-training stage; it achieves significantly lower results compared with the other three languages. One possible reason is that English shares sub-token units with the other three languages (German, French and Czech). After fine-tuning the mono-lingual English summariser, the knowledge about English tokens can be adapted to these three languages but not Chinese. Our results also match the observation from Chen et al. [6]. Their proposed model is state-of-the-art for zero-shot machine translation, where the model is tested on the languages unseen during supervised training. Their experiments find that the performance of the model is much worse in East Asian languages (Chinese, Korean, and Japanese) than in European languages (German and Romance) under the zero-shot setting.

However, both models (mBART and mBART50) benefit from slightly fine-tuning on a few samples. The overall performance in terms of the four few-shot variants on zh-en is improved by ˜10 ROUGE-L points mBART and ˜5 for mBART50. Still, few-shot remains lower for Chinese than for the other three languages. [52] points out that the summariser can improve on the new cross-lingual task by seen only a few examples. For a distant language such as Chinese, the model may require more examples to train.

		en	de-en	fr-en	cs-en	zh-en
EXT-ORACLE		31.33	23.75	25.01	25.09	25.20
LEAD		25.45	24.95	24.74	24.35	20.05
LEXRANK		25.23	24.22	24.33	23.68	24.53
mBART	Supervised	31.62	32.37	32.18	32.84	31.71
	Translated	—	30.69	30.63	30.39	29.37
	Zero	—	30.10	29.78	28.64	17.83
	Few 300 LF-MAML	—	30.84	30.44	30.15	29.10
	Few 300 FT	—	31.06	30.39	30.36	28.84
	Few 300 CVT	—	30.40	30.12	29.39	28.62
	Few 1K LF-MAML	—	31.19	30.77	31.02	29.35
mBART50	Supervised	32.53	32.95	31.84	33.72	31.85
	Translated	—	31.53	31.35	31.25	29.39
	Zero	—	31.70	30.97	31.14	25.34
	Few 300 LF-MAML	—	31.96	31.17	31.73	29.13
	Few 300 FT	—	31.77	31.39	31.67	29.08
	Few 300 CVT	—	31.77	31.08	31.91	29.20
	Few 1K LF-MAML	—	32.01	31.46	32.00	29.45

Table 4.2: ROUGE-L F1 $X \rightarrow en$ XWikis test sets. Results for de/fr/cs-en are taken from [52] and zh-en are computed in this project.

Can we Beat Machine Translation? In line with previous work[52, 29], the experiments show that Supervised variants are better than Translated ones for zh-en direction on both mBART and mBART50. Zero-shot variants are lower than Translated variants but few-shot variants achieve comparable performance. After training on 1K samples with LF-MAML, the performance of mBART50 even is better than Translated one.

Difference of Chinese Although we specify that the models should only generate English, when checking the generated summaries, we find they mix up with Chinese for all zero/few-shot scenarios. The reason is that the large pre-trained model can learn to copy words from the source document directly [64], to deal with the proper nouns or rare tokens. However, when the models encode input documents (Chinese), they

	Zero	300 FT	300 CVT	300 LF-MAML	1K LF-MAML
mBART	71.2	4.5	4.5	3.6	3.4
mBART50	45.3	2.0	3.3	2.2	1.8

Table 4.3: Number of Chinese characters appearing in the generated summaries over all tokens (%).

Generated Summary Sentences	Correct Token
A. A.汤因比 (1889–1936) was a British historian and diplomat.	Toynbee
The 假山毛榉 is a genus of plants in the family Myrtaceae.	Fagaceae
Benjamin Joseph ... was a French 钢琴ist, composer, and teacher.	Pianist

Table 4.4: Case study for English sentence mix up with Chinese characters

wrongly treat the unseen/unfamiliar Chinese characters as proper nouns and thus directly copy and output them. To assess this mix, we count the number of Chinese characters in models' output summaries for all variants with mBART and mBART50. Table 4.3 shows in zero-shot scenarios, both models output high ratio of Chinese characters, especially mBART. However, the ratios drop dramatically after seeing a few examples, which state both models have learned the cross-lingual alignment.

We further check all Chinese characters that appear in the model generated texts and find that nearly all of them are proper nouns, including person and organization names. We think this phenomenon is due to the under-train of the models. They cannot deal with rare tokens so they output the original Chinese characters. We show and analyze three of these cases in Table 4.4.

In the first case, the family name of A. A. Toynbee is not translated. We find the majority of names that the model cannot translate are family names. Generally, family is more diverse than the given name among western European culture [4], so models have less chances to see them at least once during pre-training or fine-tuning. The second case is similar, the name of special is rarely used in daily life, and might have never been seen by the model. The third case is quite interesting since it is an inner-token mixture of Chinese and English. The corresponding token in the document is 钢琴家 (English: pianist), where 钢琴 means piano and 家 refers to a specialist in a particular skill. The model correctly understands this point but fails to generate piano, finally with a mixture 钢琴*ist* Chinese and English (*ist* suffix indicates a person who is concerned with something, e.g., *scientist* or *dentist*).

Chapter 5

Content Evaluation Via Cross-Lingual Entailment

In Section 2.4.2, we introduce existing work on how to leverage NLI to evaluate content adequacy of outputs by abstractive summarisation models. However, they only focus on mono-lingual English and mainly on the news domain. In this chapter, we explore how to leverage NLI models to evaluate the content overlap of cross-lingual document-summary pairs. We first check the performance of existing multi-lingual NLI (XNLI) models in the cross-lingual setting, then we design several strategies to apply XNLI models as evaluators of content adequacy in cross-lingual abstractive summarisation. Finally, we assess the performance of XNLI models using the human evaluation judgements that we collected in Section 3.2.2.

5.1 X-NLI Model Selection

In this section, we need to find NLI models which can be used as an evaluation tool in the context of CLS, and check their effectiveness before applying them on XWikis. We choose the models which obtain good performance on a multi-lingual NLI dataset, X-NLI [10]. X-NLI includes 15 languages and its test set contains 5000 annotated sentence pairs. We display one example pair for both English and Chinese in table 5.1. Each sentence pair contains two sentences, premise and hypothesis. X-NLI model is aimed to classify the hypothesis as either entailed by, neutral, or contradicted by the premise sentence. In another word, the model will generate a probability distribution for the three labels: entailment, contradiction, and neutral for each sentence pair. X-NLI dataset collects English premises from existing corpus, employs workers to

Language	Premise / Hypothesis	Label
English	There's so much you could talk about on that I'll just skip that.	Entailment
	I won't talk about that, even though there's a lot to cover.	
Chinese	你可以讲的太多了，我就不提了。	Entailment
	即使需要说的很多，但我也不会谈论这个。	

Table 5.1: Example for X-NLI data (test set). The semantics of premises/hypotheses in two languages are the same.

produce corresponding hypotheses, and hire translators to translate sentence pairs into 15 languages.

Then, we choose two state-of-art multi-lingual models on X-NLI dataset, mDeBERTa-v3-base-mnli-xnl (mDeBERTa-v3) [30] and mt5-large-finetuned-mnli-xtreme-xnli (mT5) [65], to verify their performance in our cross-lingual scenario (premise and hypothesis are in different languages).

mDeBERTa-v3 [1] is a multilingual version of DeBERTa-v3 [23]. It leverages transfer learning to store information on language patterns (language knowledge) and specific tasks (task knowledge). We choose the base variant of mDeBERTa-v3 since it reaches the state-of-art performance on X-NLI dataset and is small enough to fit into our available GPUs.

mT5 [2] is a multilingual version of T5 [53] which is pre-trained on a very large scale of crawl-based corpora including 101 languages and achieved state-of-the-art results on many benchmarks when released [65]. We use the large variant of mT5. Including the consideration of hardware limitation, another reason to choose mT5 is that it applies relative position embedding (RPE). RPE incorporates a designed temporal bias term to encode the relative distance between any two tokens. Hence, mT5 can accept any length of the input sequence and the only constraint is memory.

The above models have been fine-tuned on X-NLI training set, i.e., pairs of English premise and English hypothesis. It is expected to transfer to other languages, e.g., French or Chinese, where the input sentence pairs (premise and hypothesis) are also in the same language. The top part of Table 5.2 shows their performance on this setting. However, we need to verify their performance in our cross-lingual scenario where the premise and hypothesis are in different languages. Since the benchmark of CLS from [52] only contains ALL-to-en direction, we follow the same directions. We create

[1] https://huggingface.co/MoritzLaurer/mDeBERTa-v3-base-mnli-xnli

[2] https://huggingface.co/alan-turing-institute/mt5-large-finetuned-mnli-xtreme-xnli

	mDeBERTa-v3	mT5
$\mathcal{D}_{en \to en}$	88.3	88.8
$\mathcal{D}_{fr \to fr}$	83.4	83.9
$\mathcal{D}_{de \to de}$	82.5	84.3
$\mathcal{D}_{zh \to zh}$	81.1	82.3
$\mathcal{D}_{fr \to en}$	84.9	84.9
$\mathcal{D}_{de \to en}$	85.1	85.8
$\mathcal{D}_{zh \to en}$	83.1	84.4

Table 5.2: F1 score for mDeBERTa-v3 and mT5 on X-NLI dataset. $\mathcal{D}_{X \to Y}$ represents that the premise is from lanaguge X and hypothesis is from language Y

cross-lingual sentence pair from X-NLI test set, where the premises are from German, French and Chinese, and the hypotheses are from English. Czech is not included in X-NLI dataset so we cannot test it.

Table 5.2 bottom, shows the results of cross-lingual scenarios. We find the performances on $\mathcal{D}_{X \to en}$ are even better than the mono-lingual $D_{X \to X}$ direction in language other than English. We guess one reason can be that both models are fine-tuned in English-English pairs and our cross-lingual setting benefits from having hypothesis sentences in English.

In conclusion, we think both models, mDeBERTa-v3 and mT5, show reasonable performance to be assessed and applied as content evaluators for document-summary pairs in the CLS task.

5.2 Evaluating Entailment in Document-Summary Pairs

In this section, we discuss how to leverage two X-NLI models to serve as content adequacy evaluators on CLS task. Following previous work [43] [27], we assume that a summary should be entailed by the document. In other words, the content conveyed by the sentences in a summary should be supported by the corresponding document.

We can assess entailment between each sentence in a summary and each sentence in the document using the X-NLI models. However, due to the nature of abstractive summarisation, a summary sentence can aggregate information from multiple sentences in a document. In this case, by evaluating one document sentence at a time, entailment will fail for each of the summary sentences. Laban et al.[27] have also taken all document sentences as the premise and evaluated entailment for each sentence in

the summary. Thus, we can divide the discussion into different ways we can take sentences from the document (premises) and sentences from the summary (hypothesis) for entailment evaluation. At a high level, we split the discussion into two parts, the one-to-one relationship and the many-to-one relationship.

5.2.1 One-to-One Relationship

We generate a XNLI sentence pair (premise, hypothesis) by a (document, summary) pair at the sentence level. Specifically, the document is split into sentences, as premises labeled from $\text{Doc}_X = \{P_1, ..., P_M\}$. The summary is also split into sentences, as hypotheses labeled from $\text{Sum}_Y = \{H_1, ..., H_N\}$.

The X-NLI models run over each sentence pair (P_m, H_n) and generate a probability distribution over three labels: entailment, contradiction, and neutral. Laban et al. [27] revealed that using different combinations of three labels only slightly influences the result. For simplicity, we only pick the probability of entailment for the following experiments. After that, we form an $M \times N$ matrix E consisting of the entailment scores E_{mn}, where m and n are the indexes of sentences in the document and summary, respectively.

In the one-to-one relationship, we assume each summary sentence H_n is supported by only one document sentence P_m. Intuitively, for the summary sentence, only the document sentence P_m with the top entailment score (the strongest support) should be selected. We apply the max operator on each matrix column and obtain a $1 \times N$ vector. We take the mean operator on the vector to produce the final score for a document-summary pair.

5.2.2 Many-to-One Relationship

In the many-to-one relationship, we assume each summary sentence H_n can be supported by more than one document sentence $\{P_j, \cdots, P_k\}$. We refer to them as P_{many}. The core problem is, **how to find P_{many} in the document**, which can correctly entail H_n. We heuristically propose three strategies to search for such sentences.

Strategy 1: In one-to-one scenarios, we apply the max operator on each column to choose the P_m with the highest entailment score. Similarly, Strategy 1 takes $\{P_j, \cdots, P_k\}$ with top C entailment probabilities. In other words, we use the results E_{mn} from the one-to-one relationship to select C candidate document sentences. The sentences

$\{P_j, \cdots, P_k\}$ with the top C highest supporting sentences will be kept in their original position and concatenated to form the P_{many}. We apply X-NLI models on the created pairs of premise and hypothesis (P_{many}, H_n) to obtain new entailment scores.

Strategy 2: The idea of Strategy 2 is the same as Strategy 1, though the direction used to create the one-to-one matrix E is reversed. This time we run the X-NLI models on one-to-one pairs (H_n, P_1), ..., (H_n,P_M), where the summary sentence H_n acts as the premise and each document sentence $\{P_1, \cdots, P_M\}$ as the hypothesis. Then, we select the top C document sentences which are most likely to be entailed by H_n. As in Strategy 1, these sentences will be concatenated to form the P_{many} and the X-NLI models are applied on the created pairs (P_{many}, H_n) to obtain new entailment scores.

Merge: Merge leverages two matrices from strategy 1 and 2. The strategy element-wise adds them together and use the new matrix E to select top C document sentences to form P_{many} and create the pair (P_{many}, H_n). The later procedure is the same as Strategy 1 and 2.

Sentences in XWikis documents and summaries aggregate different number of content units which hinders one-to-one entailment assessment. We analyse three special cases illustrated in Figures 5.1, 5.2, and 5.3. Suppose there is a document-summary pair, where the document contains three sentences $\{P_1, P_2, P_3\}$ and the summary only contains one sentence H_1. The task is to correctly select C=2 sentences from the document that can entail the H_1. We set the example in such a way that P_1 and P_3 will make P_{many} and P_3 is a weaker support sentence. P_2 is constructed to mislead the strategy, except in Case 1. A, B, C, D, E, and F are units of semantic information. We assume each semantic unit is independent of the others and the premise can entail the hypothesis if and only if the premise contains more or the same semantic units than the hypothesis contains.

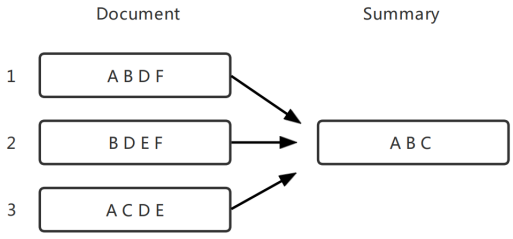

Figure 5.1: Case 1: document sentences are longer than the summary sentence.

As the Figure 5.1 shows, if the document sentences are longer than H_1 (contain

more semantics units), Strategy 1 works well. In this case, concatenated P_1 and P_3 as P_{many} can entail the summary sentence and obtain a high score.

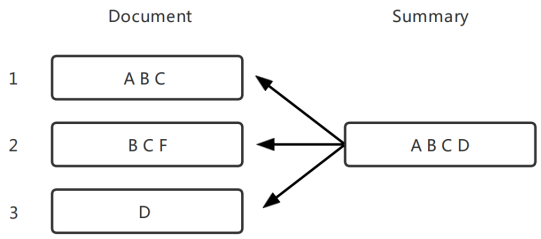

Figure 5.2: Case 2: document sentences are shorter than the summary sentence.

In Case 2 (Figure 5.2), P_3 obtains a relatively low likelihood to entail H_1 but (in some cases, e.g., paraphrases) could be entailed by H_1. Thus, Strategy 1 will select P_1 and P_2 as P_{many} while Strategy 2 correctly selects P_1 and P_3. From this case it follows that in some cases, if a sentence in the document contains fewer semantic units supporting the summary sentence, Strategy 2 can capture the entailment relation.

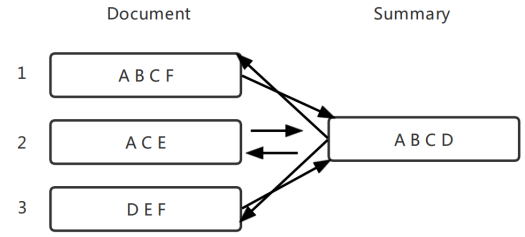

Figure 5.3: Case 3: document sentences have the similar length as the summary sentence.

Case 3 is more complex since each document sentence is mixed with supportive and not supportive information for the summary sentence. In this case, we can see that neither Strategy 1 nor Strategy 2 will select P_3 since the overlap is low for either P_3 *entail* H_1 or H_1 *entail* P_3.

A natural question here is, why we do not use all document sentences concatenated as the premise, which can indeed include all information from the document? [28] points out that this idea can suffer from a mismatch in input granularity between NLI datasets (sentence-level), and factual detection (document-level). The noise from the whole document can also hurt the classification. Still, for the sake of evaluation in our setting we include this as an alternative strategy, namely **Document-to-One**.

Here we review the current issues. Usually, the sentence from the summary aggregates information from multiply document sentences. However, if we merge too many

	Content
Input	placeholder: The black-browed albatross (Thalassarche melanophris), also known as the black-browed mollymawk, is a large seabird of the albatross family Diomedeidae. {mask}
Target sentences	The black-browed albatross (Thalassarche melanophris) is also known as the black-browed mollymawk. The black-browed albatross is a large seabird of the albatross family Diomedeidae.

Table 5.3: One example of template

document sentences as the premise, will lead to a mismatch for the XNLI model since it was fine-turned on at sentence level and so lacks the ability to capture information from long text [27]. To deal with such a dilemma, we split the summary sentence into several sub-sentences, which can decrease the information that each summary sentence contains. In other words, we reshape the hard Case 3 to the easier Case 1.

To split summary sentences, we introduce a pre-trained language model T5-large [53]. We collect 100 instances from [1], and apply prompt learning to train a prefix template. Prompt learning is a paradigm to quickly adapt pre-trained language models to downstream NLP tasks. It augments the input text with a textual template according to the desired task to carry out. Thus, it transfers the downstream tasks to the pre-trained language model.

We use the open-prompt package [15] with prefix turning [35]. The model was trained for 15 epochs using the Adam optimizer (learning rate=3e-4) with linear learning rate decay scheduling. We provide one example of how to truncate the whole sentence into sub-sentences in table 5.3. After inserting a placeholder as the prompt before the input sentence, the model needs to learn what to generate for the content of the mask and calculates the loss with respect to the target sentences. More examples can be found in Appendix A.4. Basically, we only split the document sentence whose length is higher than a threshold. We take care of some potential issues when creating training samples. We confirm the subjective of each new sub-sentence is correct and the relationship inner original sentence (e.g. causal relationship) is not broken after sub-sentence segmentation. The total number of summary sentences is 53 and grows to 87 after segmentation.

5.3 Experimental Setup

In this section, we design experiments to assess how well XNLI models (using the strategies introduced in the previous section) align or can predict human judgements at the sentence and sub-sentence levels, on the zh-en direction of XWikis dataset.

As human judgements, we re-use a fine-grained version of the annotations collected in the human validation experiment in Section 3.2.2. Fine-grained human judgements range from **entail**, **partly_entail**, to **not_entail**. These judgements can be seen as ordinal variables and we map them to 1, 0.5, and 0 for convenience. We vote the judgements from three annotators on each summary sentence. At the sub-sentence level, **entail** and **not _entail** judgements will be directly assigned to each sub-sentence. We re-annotate sub-sentences derived from sentences labeled as **partly_entail**.

5.4 Entailment for Approximation of Human Judgements on Content Adequacy

In this section, we evaluate X-NLI models with human evaluation. We compute correction based on Spearman's correlation coefficients. Spearman correlation is to measure how well the relationship between two variables can be described using a monotonic function [31]. Such a score will be high when two variables share a similar ranking. We analyse the 5 strategies previously introduced, One-to-One, Strategy 1, Strategy 2, Merge, and Document-to-One. We only include the analyses with mT5 as it shows better performance; results for mDeBERTa-v3 can be found in Appendix A.3. Table 5.4 shows the results of the correlation analysis between the XNLI model (with different strategies) and human judgements at sentence and sub-sentence levels, on zh-en direction of XWikis.

Does the XNLI model correlate with human judgements on summary sentences? From the results in Table 5.4, moderate correlation (0.493 and 0.521) is found between XNLI model (mT5) and human judgements. Considering that mT5 has not been fine-tuned on the XWikis corpus neither on the cross-lingual setting (premise and hypothesis are in different languages), we consider the XNLI model can provide acceptable and promising performance to judge the content adequacy of sentence summaries.

Which strategy works best? How does the value of C influence strategies? For both sentence and sub-sentence scenarios, the Document-to-One strategy performs

Sum.Sub-Sentence	C=2	4	6	8	10	12	14	16
One-to-One	0.399	-	-	-	-	-	-	-
Strategy 1	0.411	0.336	0.372	0.346	0.343	0.353	0.370	0.386
Strategy 2	0.312	0.319	0.294	0.396	0.412	0.390	0.381	0.418
Merge	0.479	**0.493**	0.453	0.355	0.370	0.370	0.378	0.384
Document-to-One	0.372	-	-	-	-	-	-	-
Sum.Sub-Sentence	C=2	4	6	8	10	12	14	16
One-to-One	0.401	-	-	-	-	-	-	-
Strategy 1	**0.521**	0.449	0.417	0.371	0.359	0.374	0.365	0.365
Strategy 2	0.427	0.415	0.422	0.432	0.423	0.416	0.440	0.450
Merge	0.485	0.450	0.409	0.371	0.417	0.402	0.395	0.415
Document-to-One	0.395	-	-	-	-	-	-	-

Table 5.4: Correlations between human annotators and mT5. Each cell shows the score for label entailment.

the worst. This aligns with the observations in recent work by Laban et al. [27], who find that when taking too many sentences as the premise, the train/test mismatch issue appears. One-To-One is a little higher than Document-to-One while all other strategies are higher than One-to-One for most values of C, which evidences that the summary sentence indeed aggregates information from multiple document sentences. Merge obtains the best results at sentence level when $C = 4$. However, we cannot confirm that each summary sentence needs four document sentences to be supported (or not) on average. When C increases the train/test mismatch issue becomes severe getting closer to the Document-to-One strategy.

Generally, with the increment of C, the score of strategy 1 is decreased and then oscillated, while the score of strategy 2 is first increased and then oscillated. It is interesting that strategy 2 can obtain a high correlation with humans though the value of C is high, which points to the direction $S_n \rightarrow D_m$ (case 2 in figure 5.2) providing a better signal even at a high C.

Merge is relatively robust to the value of C but its performance still drops with a high C. When the value of C is in [2, 8], its value is higher than Strategy 1 and 2. This means that the two strategies make different predictions on the same summary sentence, and simply summing their results can better capture the entailment relations. When C increases, the sentences selected by the three strategies are similar so the scores are also close.

text	The	Bay	of	Pigs	invasion	is	conducted	by
1-gram	0.11	0.16	0.12	0.20	0.23	0.11	0.21	0.16
2-grams	-	0.15	0.15	0.19	0.27	0.21	0.26	0.20
text	Cuban	exiles	who	opposed	Fidel	Castro's	Cuban	Revolution.
1-gram	0.11	0.17	0.12	0.13	0.25	**0.25**	0.14	0.07
2-grams	0.10	0.23	0.09	0.10	0.11	**0.31**	0.24	0.08

Table 5.5: The entailment probability changed by masking 1-gram and 2-grams in the sentences. The score given to the full sentence is 0.12 and the human judgement is **partly_entail** due to **Cuban exiles** is not mentioned in the document.

Where the XNLI model is more likely to make a wrong prediction? We define a wrong prediction as two situations (1) the human judgement is **entail** but the score of model output is low and (2) the human judgement **not_entail** but the score of the model is high. As Figure 5.4 shows, situation 2 is rare but situation 1 is very frequent. It shows that mT5 may assign a rather low entailment score to a summary sentence which in fact can be supported by the document (according to human judgement). Thus, we manually check several cases and conclude the main reasons are the following.

Lack of translation. To know which tokens cause the low entailment score in a sentence, we iteratively mask 1-gram and 2-grams in the sentence and compute the sentence entailment score at each step. We assume that if the score of entailment is increased after removing certain token(s), such tokens are not understood by the XNLI model. The example in Table 5.5 shows at each token entry the sentence entailment score obtained when we mask that token (or two consecutive tokens). We observe that the problematic tokens are proper names, such as *Fidel Castro's* and *Bay of Pigs Invasion*, and we confirm the corresponding tokens exist in the associated Chinese document. Although the sanity check in Table 5.2 suggests both mT5 and mDERBert-v3 can be applied in our cross-lingual setting, XWikis contains more diverse and difficult proper nouns than the X-NLI dataset. The XNLI models cannot align some proper names between Chinese and English, leading to the translation issue.

Ontological and world knowledge. In the first example in Table 5.6, the summary says *an airplane from Barcelona–El Prat Airport in Spain to Düsseldorf Airport in Germany* while the document only mentions the two cities but not which country the city belongs to. When we remove the names of two countries, the entailment score is

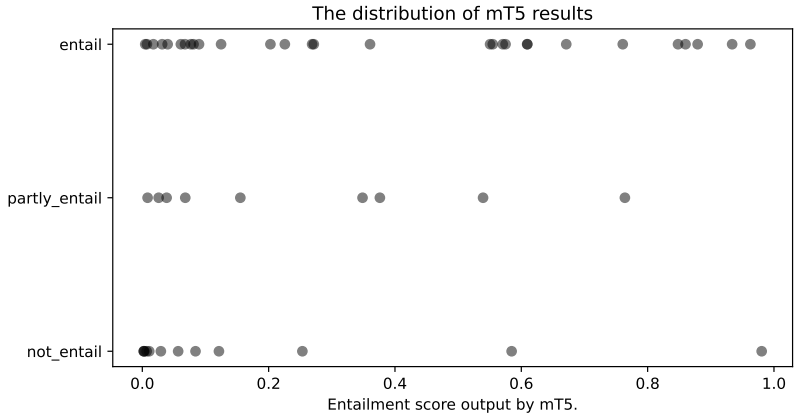

Figure 5.4: The distribution between human evaluation and prediction from mT5. Strategy 1, C=2.

increased. This problem is due to the weak ability of XNLI models with ontological and world knowledge. Ontological knowledge refers to the relation across tokens, such as hypernymy/hyponymy or meronymy/holonymy for nouns (cat $\rightarrow$ animal) [51]. World knowledge refers to all other than ontological knowledge, such as causation relations (hungry $\rightarrow$ eat) or knowledge about named entities [57]. Though some evidence [11, 67] shows large pre-trained models have learned common sense knowledge from massive data, it is still hard for them to handle complex world or ontological knowledge, especially in the encyclopedias domain. In this example, to successfully handle this case, when the XNLI model grasps the content Düsseldorf from the document, it needs to know that Düsseldorf is a city in Germany (world knowledge) and the relationship between country and city (hypernymy/hyponymy).

Co-reference problem. Co-reference is when two different linguistic expressions refer to the same entity in a text (e.g., the first time the entity is introduced is referred to as *Barack Obama* and in sub-sequence sentences is referred as *he*). The co-reference problem occurs as we analyse entailment at the summary sentence level. As the second example in Table 5.6 shows, the subject of the summary sentence is **it** but the model does not what **it** refer to as the sentence is evaluated in isolation. Thus, the entailment score is low. After correctly substituting **it** by the name of the movie it refers to, the entailment score is doubled.

Does sub-sentence work? Comparing top and bottom blocks Table 5.4, all strategies work better at sub-sentence level. Strategy 1 obtains the best performance, which is higher than the best result in at the sentence level. This shows that in a finer granularity

	Content	Score
Original	Germanwings Flight 9525 was a scheduled international passenger flight... from Barcelona–El Prat Airport in Spain to Düsseldorf Airport in Germany.	0.57
Modified	Germanwings Flight 9525 was a scheduled international passenger flight...from Barcelona–El Prat Airport to Düsseldorf Airport.	0.68
Original	It is based on emoji faces, smileys and graphics used in electronic messages.	0.08
Modified	The Emoji Movie is based on emoji faces, smileys and graphics used in electronic messages.	0.15

Table 5.6: The entailment probability changed by modifying the summary sentence. The human judgements for two examples are **entail** and **partly_entail**. In the second example, **smileys and graphics** is not mentioned in the corresponding document.

(sub-sentence level), each summary sentence contains less information and is easier to be predicted by document sentences. The sub-sentence level creates more cases of the Case 1 discussed in Figure 5.1. As mentioned before, Strategy 1 is specially suitable for dealing with Case 1, thus its best performance at the sub-sentence level.

Comparing the distributions at sentence and sub-sentence level in Figures 5.4 and 5.5, we can see that the situation 1 (models assign a small score to a summary sentence that can be entailed by the document) is less frequent. A desired effect of the sub-sentence level analysis is that summary sentences judged as **partly_entail** will be better predicted at sub-sentence level, potentially providing separate **entail** and **not_entail** judgements. We show an example of this in Table 5.7. The original summary sentence introduces a person, including time, held positions, and interpersonal relations, which conveys several semantic units. After sub-sentence splitting, such semantic units are split into two sentences. Then, the XNLI model can better assess the relation between the summary and document sentences.

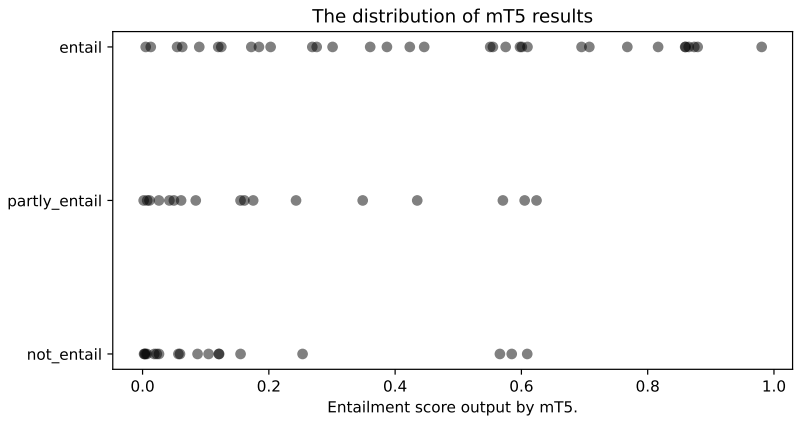

Figure 5.5: The distribution between human evaluation and prediction from mT5 at the sub-sentence level. Strategy 1, C=2.

	Content	Score
Summary Sentence	Adam Silver joined the NBA in 1992 and has held various positions within the league, becoming chief operating officer and deputy commissioner under his predecessor and mentor David Stern in 2006.	0.27
Sub-sentence1	Adam Silver joined the NBA in 1992 and has held various positions within the league.	0.91
Sub-sentence2	Adam Silver became chief operating officer and deputy commissioner under his predecessor and mentor David Stern in 2006.	0.04

Table 5.7: The entailment probability at the sentence and sub-sentence levels. The human judgements are **partly_entail**, **entail**, and **not_entail** for the summary sentence and its sub-sentence 1 and 2 respectively.

Chapter 6

Conclusions

We draw conclusions for the three main parts of this project.

- We extend an existing cross-lingual summarisation dataset XWikis with a more distant language, Chinese, which we hope will be a valuable resource for research.
- We evaluate a wide range of models on the cross-lingual summarisation task, including translated, supervised, and zero/few-shot variants. Compared with the other directions (fr-en, de-en, cs-en), the overall performance is lower in zh-en direction. Specifically, the result of zh-en is much worse in the zero-shot scenario but can be improved after slight fine-tuning on a few samples. The conclusion from the previous works [29, 52] that Supervised models are better than Translated ones, still holds for zh-en.
- We explore how to leverage XNLI models to evaluate cross-lingual document summary pairs. We propose several strategies and the experiments show the predictions from them correlate with human evaluation. We also find that applying sub-sentence segmentation to summary sentences can help the model to capture alignments in a document-summary pair and thus obtain better correlation scores.

In our future work, we can apply XNLI models on XWikis to automatically check the content adequacy of the extracted cross-lingual document-summary pairs. To do so, we need to translate the score for entailment to human-readable labels, in other words, to construct a map from [0,1] to {entail, not_entail}. For labels, we can aggregate the human evaluation labels not_entail and partly_entail together, and keep entail still, to form a binary classification (entail vs not_entail). Figure 6.1 shows an attempt using the ROC curve. The threshold for the best curve (Strategy 1) is 0.269. We can treat the

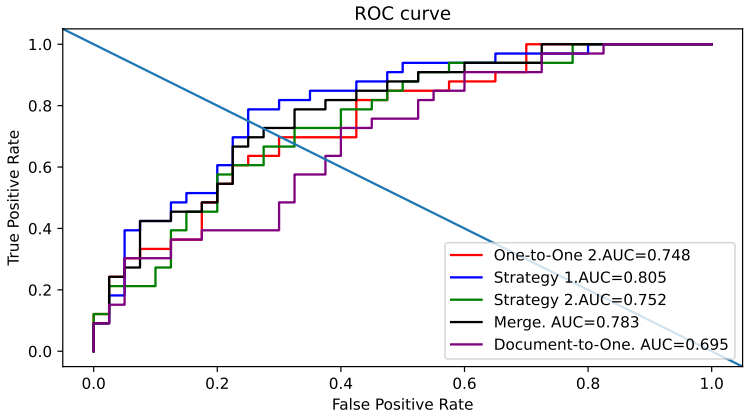

Figure 6.1: ROC curve at the sub-sentence level. C=2.

entailment probability of a document summary sentence pair above this value as **entail** and the remaining ones as **not_entail**.

Another direction is to consider the loss from the XNLI model as an extra supervised signal, to better fine-tune a large pre-trained multi-lingual language model on the CLS task. The addition of the XNLI model may help the model generate more faithful summaries.

Our proposed strategies are interpretable, which means humans can know which portion of documents supports the summary by viewing the entailment scores. It is also possible to incorporate the entailment relation into visual analysis tools for summarisation such as SummViz [61]. Exploring the interpretability of the NLI model on the CLS task can be another future work.

Bibliography

[1] Roee Aharoni and Yoav Goldberg. Split and rephrase: Better evaluation and stronger baselines. In Iryna Gurevych and Yusuke Miyao, editors, *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 2: Short Papers*, pages 719–724. Association for Computational Linguistics, 2018.

[2] Reinald Kim Amplayo, Peter J. Liu, Yao Zhao, and Shashi Narayan. Smart: Sentences as basic units for text evaluation, 2022.

[3] Naveen Arivazhagan, Ankur Bapna, Orhan Firat, Dmitry Lepikhin, Melvin Johnson, Maxim Krikun, Mia Xu Chen, Yuan Cao, George F. Foster, Colin Cherry, Wolfgang Macherey, Zhifeng Chen, and Yonghui Wu. Massively multilingual neural machine translation in the wild: Findings and challenges. *CoRR*, abs/1907.05019, 2019.

[4] RL Arthaud, AN Hohneck, CH Ramsey, and KC Pratt. The relation of family name preferences to their frequency in the culture. *The Journal of Social Psychology*, 28(1):19–37, 1948.

[5] Phyllis B. Baxendale. Machine-made index for technical literature - an experiment. *IBM J. Res. Dev.*, 2(4):354–361, 1958.

[6] Guanhua Chen, Shuming Ma, Yun Chen, Dongdong Zhang, Jia Pan, Wenping Wang, and Furu Wei. Towards making the most of cross-lingual transfer for zero-shot neural machine translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 142–157, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[7] Elizabeth Clark, Asli Celikyilmaz, and Noah A. Smith. Sentence mover's similarity: Automatic evaluation for multi-sentence texts. In Anna Korhonen, David R.

Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 2748–2760. Association for Computational Linguistics, 2019.

[8] Kevin Clark, Minh-Thang Luong, Christopher D. Manning, and Quoc V. Le. Semi-supervised sequence modeling with cross-view training. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii, editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 1914–1925. Association for Computational Linguistics, 2018.

[9] Trevor Anthony Cohn and Mirella Lapata. Sentence compression as tree transduction. *Journal of Artificial Intelligence Research*, 34:637–674, 2009.

[10] Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel R. Bowman, Holger Schwenk, and Veselin Stoyanov. Xnli: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2018.

[11] Joe Davison, Joshua Feldman, and Alexander Rush. Commonsense knowledge mining from pretrained models. 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 1173–1178, Hong Kong, China, November 2019. Association for Computational Linguistics.

[12] Mingkai Deng, Bowen Tan, Zhengzhong Liu, Eric Xing, and Zhiting Hu. Compression, transduction, and creation: A unified framework for evaluating natural language generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7580–7605, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[13] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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, June 2019. Association for Computational Linguistics.

[14] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics, 2019.

[15] Ning Ding, Shengding Hu, Weilin Zhao, Yulin Chen, Zhiyuan Liu, Haitao Zheng, and Maosong Sun. Openprompt: An open-source framework for prompt-learning. In Valerio Basile, Zornitsa Kozareva, and Sanja Stajner, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics, ACL 2022 - System Demonstrations, Dublin, Ireland, May 22-27, 2022*, pages 105–113. Association for Computational Linguistics, 2022.

[16] Esin Durmus, He He, and Mona Diab. FEQA: A question answering evaluation framework for faithfulness assessment in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5055–5070, Online, July 2020. Association for Computational Linguistics.

[17] Günes Erkan and Dragomir R. Radev. Lexrank: Graph-based lexical centrality as salience in text summarization. *J. Artif. Intell. Res.*, 22:457–479, 2004.

[18] Tobias Falke, Leonardo F. R. Ribeiro, Prasetya Ajie Utama, Ido Dagan, and Iryna Gurevych. Ranking generated summaries by correctness: An interesting but challenging application for natural language inference. In Anna Korhonen, David R. Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 2214–2220. Association for Computational Linguistics, 2019.

[19] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning, ICML*

2017, Sydney, NSW, Australia, 6-11 August 2017, volume 70 of *Proceedings of Machine Learning Research*, pages 1126–1135. PMLR, 2017.

[20] Yvette Graham. Re-evaluating automatic summarization with BLEU and 192 shades of ROUGE. In Lluís Màrquez, Chris Callison-Burch, Jian Su, Daniele Pighin, and Yuval Marton, editors, *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*, pages 128–137. The Association for Computational Linguistics, 2015.

[21] Max Grusky, Mor Naaman, and Yoav Artzi. Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies. In Marilyn A. Walker, Heng Ji, and Amanda Stent, editors, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, pages 708–719. Association for Computational Linguistics, 2018.

[22] Han He and Jinho D. Choi. The stem cell hypothesis: Dilemma behind multi-task learning with transformer encoders. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5555–5577, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[23] Pengcheng He, Jianfeng Gao, and Weizhu Chen. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. *CoRR*, abs/2111.09543, 2021.

[24] Wojciech Kryscinski, Nitish Shirish Keskar, Bryan McCann, Caiming Xiong, and Richard Socher. Neural text summarization: A critical evaluation. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 540–551. Association for Computational Linguistics, 2019.

[25] Wojciech Kryscinski, Bryan McCann, Caiming Xiong, and Richard Socher. Evaluating the factual consistency of abstractive text summarization. In *Proceedings*

of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 9332–9346, Online, November 2020. Association for Computational Linguistics.

[26] Taku Kudo and John Richardson. 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, November 2018. Association for Computational Linguistics.

[27] Philippe Laban, Tobias Schnabel, Paul N. Bennett, and Marti A. Hearst. Summac: Re-visiting nli-based models for inconsistency detection in summarization. *Trans. Assoc. Comput. Linguistics*, 10:163–177, 2022.

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

[29] Faisal Ladhak, Esin Durmus, Claire Cardie, and Kathleen R. McKeown. Wikilingua: A new benchmark dataset for cross-lingual abstractive summarization. *CoRR*, abs/2010.03093, 2020.

[30] Wouter van Atteveldt Andreu Salleras Casas Laurer, Moritz and Kasper Welbers. Less annotating, more classifying – addressing the data scarcity issue of supervised machine learning with deep transfer learning and bert - nli. *Open Science Framework*.

[31] Ann Lehman. *JMP for basic univariate and multivariate statistics: a step-by-step guide*. SAS Institute, 2005.

[32] Anton Leuski, Chin-Yew Lin, Liang Zhou, Ulrich Germann, Franz Josef Och, and Eduard Hovy. Cross-lingual c*st*rd: English access to hindi information. *ACM Transactions on Asian Language Information Processing*, 2(3):245–269, sep 2003.

[33] Anton Leuski, Chin-Yew Lin, Liang Zhou, Ulrich Germann, Franz Josef Och, and Eduard H. Hovy. Cross-lingual c*st*rd: English access to hindi information. *ACM Trans. Asian Lang. Inf. Process.*, 2(3):245–269, 2003.

[34] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel R. Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 7871–7880. Association for Computational Linguistics, 2020.

[35] Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 4582–4597. Association for Computational Linguistics, 2021.

[36] Han-Teng Liao, King-wa Fu, and Scott A. Hale. How much is said in a microblog?: A multilingual inquiry based on weibo and twitter. In David De Roure, Pete Burnap, and Susan Halford, editors, *Proceedings of the ACM Web Science Conference, WebSci 2015, Oxford, United Kingdom, June 28 - July 1, 2015*, pages 25:1–25:9. ACM, 2015.

[37] Jung-Min Lim, In-Su Kang, and Jong-Hyeok Lee. Multi-document summarization using cross-language texts. In Noriko Kando and Haruko Ishikawa, editors, *Proceedings of the Fourth NTCIR Workshop on Research in Information Access Technologies Information Retrieval, Question Answering and Summarization, NTCIR-4, National Center of Sciences, Tokyo, Japan, June 2-4, 2004*. National Institute of Informatics (NII), 2004.

[38] Chin-Yew Lin and Franz Josef Och. Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. In Donia Scott, Walter Daelemans, and Marilyn A. Walker, editors, *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics, 21-26 July, 2004, Barcelona, Spain*, pages 605–612. ACL, 2004.

[39] Feifan Liu and Yang Liu. Exploring correlation between ROUGE and human evaluation on meeting summaries. *IEEE Trans. Speech Audio Process.*, 18(1):187–196, 2010.

[40] Yang Liu and Mirella Lapata. Hierarchical transformers for multi-document summarization. In Anna Korhonen, David R. Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 5070–5081. Association for Computational Linguistics, 2019.

[41] Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. Multilingual denoising pre-training for neural machine translation. *Trans. Assoc. Comput. Linguistics*, 8:726–742, 2020.

[42] Timothy May. David w. anthony, the horse, the wheel, and language: How bronze-age riders from the eurasian steppes shaped the modern world. princeton, nj: Princeton university press, 2007. xii+ 553 pp. isbn: 978-0-691-5887-0 (hbk.). 35.00. *Itinerario*, 32(3):102–103, 2008.

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

[44] Ramesh Nallapati, Feifei Zhai, and Bowen Zhou. Summarunner: A recurrent neural network based sequence model for extractive summarization of documents. In Satinder P. Singh and Shaul Markovitch, editors, *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA*, pages 3075–3081. AAAI Press, 2017.

[45] Shashi Narayan, Shay B. Cohen, and Mirella Lapata. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1797–1807, Brussels, Belgium, October-November 2018. Association for Computational Linguistics.

[46] Ani Nenkova and Kathleen R. McKeown. Automatic summarization. *Found. Trends Inf. Retr.*, 5(2-3):103–233, 2011.

[47] Ani Nenkova and Rebecca J. Passonneau. Evaluating content selection in summarization: The pyramid method. In Julia Hirschberg, Susan T. Dumais, Daniel Marcu, and Salim Roukos, editors, *Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics,*

HLT-NAACL 2004, Boston, Massachusetts, USA, May 2-7, 2004, pages 145–152. The Association for Computational Linguistics, 2004.

[48] Graham Neubig and Kevin Duh. How much is said in a tweet? A multilingual, information-theoretic perspective. In *Analyzing Microtext, Papers from the 2013 AAAI Spring Symposium, Palo Alto, California, USA, March 25-27, 2013*, volume SS-13-01 of *AAAI Technical Report*. AAAI, 2013.

[49] Khanh Nguyen and Hal Daumé III. Global voices: Crossing borders in automatic news summarization. *CoRR*, abs/1910.00421, 2019.

[50] Peter Norvig. English letter frequency counts:mayzner revisited or etaoin srhldcu, 2012.

[51] Sebastian Padó and Ido Dagan. Textual Entailment. In *The Oxford Handbook of Computational Linguistics, Second Edition*. Oxford University Press.

[52] Laura Perez-Beltrachini and Mirella Lapata. Models and datasets for cross-lingual summarisation. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 9408–9423. Association for Computational Linguistics, 2021.

[53] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(1), jan 2020.

[54] Jürgen Schmidhuber. *Evolutionary principles in self-referential learning, or on learning how to learn: the meta-meta-... hook*. PhD thesis, Technische Universität München, 1987.

[55] Thomas Scialom, Paul-Alexis Dray, Sylvain Lamprier, Benjamin Piwowarski, Jacopo Staiano, Alex Wang, and Patrick Gallinari. QuestEval: Summarization asks for fact-based evaluation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6594–6604, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[56] Thomas Scialom, Sylvain Lamprier, Benjamin Piwowarski, and Jacopo Staiano. Answers unite! unsupervised metrics for reinforced summarization models. 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 3246–3256, Hong Kong, China, November 2019. Association for Computational Linguistics.

[57] Satoshi Sekine, Kentaro Inui, Ido Dagan, William B Dolan, Danilo Giampiccolo, and Bernardo Magnini. Proceedings of the acl-pascal workshop on textual entailment and paraphrasing. In *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*, 2007.

[58] Thibault Sellam, Dipanjan Das, and Ankur P. Parikh. BLEURT: learning robust metrics for text generation. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel R. Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 7881–7892. Association for Computational Linguistics, 2020.

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

[60] Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. Multilingual translation with extensible multilingual pretraining and finetuning. *CoRR*, abs/2008.00401, 2020.

[61] Jesse Vig, Wojciech Kryscinski, Karan Goel, and Nazneen Rajani. Summvis: Interactive visual analysis of models, data, and evaluation for text summarization. In Heng Ji, Jong C. Park, and Rui Xia, editors, *Proceedings of the Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL 2021 - System Demonstrations, Online, August 1-6, 2021*, pages 150–158. Association for Computational Linguistics, 2021.

[62] Xiaojun Wan, Huiying Li, and Jianguo Xiao. Cross-language document summarization based on machine translation quality prediction. In Jan Hajic, Sandra

Carberry, and Stephen Clark, editors, *ACL 2010, Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, July 11-16, 2010, Uppsala, Sweden*, pages 917–926. The Association for Computer Linguistics, 2010.

[63] Jiaan Wang, Fandong Meng, Duo Zheng, Yunlong Liang, Zhixu Li, Jianfeng Qu, and Jie Zhou. A survey on cross-lingual summarization. *arXiv preprint arXiv:2203.12515*, 2022.

[64] Song Xu, Haoran Li, Peng Yuan, Youzheng Wu, Xiaodong He, and Bowen Zhou. Self-attention guided copy mechanism for abstractive summarization. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel R. Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 1355–1362. Association for Computational Linguistics, 2020.

[65] Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 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, June 2021. Association for Computational Linguistics.

[66] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020.

[67] Xuhui Zhou, Yue Zhang, Leyang Cui, and Dandan Huang. Evaluating commonsense in pre-trained language models. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 9733–9740. AAAI Press, 2020.

[68] Junnan Zhu, Qian Wang, Yining Wang, Yu Zhou, Jiajun Zhang, Shaonan Wang, and Chengqing Zong. NCLS: neural cross-lingual summarization. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of the 2019*

Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019, pages 3052–3062. Association for Computational Linguistics, 2019.

Appendix A

Discussion for data creation

A.1 Factor selection

An open research problem in the linguistic field is to quantify the expressiveness across language, in another word, to count the number of words/characters used to express the same information.

Neubig and Duh [48] collect data from the Twitter platform and find that Chinese and Japanese are more expressive than the other 24 languages. Liao et al. [36] further establish a comparison based on the parallel corpus, openly available Universal Declaration of Human Rights, and the translated subtitles from TED talks. Their results are shown below figure A.1.

We compare the text of UDHR and TED Talks and believe the former is more closer to our Wikipedia domain and thus pick 4 as the ratio of characters between Chinese and English required to express the same information. In addition, the average number of characters of an English word is 4.79 [50]. By these two numbers, we calculate the ratio between the English words and Chinese characters is 1.2.

A.2 Natural language processing for Chinese text

During the whole project, there are several problems specific to Chinese texts. They can be divided into two classes, the variants of Chinese and Chinese word segmentation.

The Chinese wiki dump used for this project contains a mix of multiple variants of Chinese (simplified Chinese, traditional Chinese, and so on). Even in an article, it is possible that some paragraph is traditional Chinese while others are simplified Chinese. We finally use simplified Chinese. The reasons are:

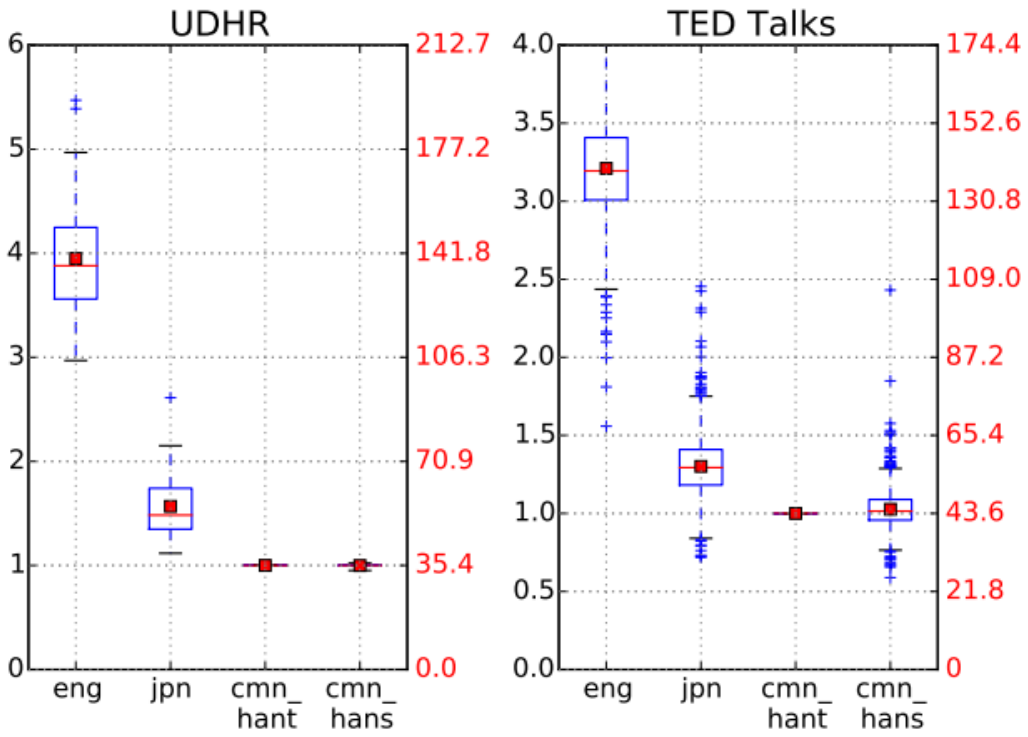

Figure A.1: Relative ratio of characters in English (eng), Japanese (jpn), and Chinese using simplified characters (cmn hans) required to express the same content compared to Chinese using traditional characters (cmn hant) as the baseline [36].

(1) The evaluation metric ROUGE is based on the token overlap. However, the main difference across variants of Chinese is character/token. This drives we must choose one of the variants.

(2) The pre-trained language models used in the project, mBART and mBART50 have been trained on the corpora, where most of the Chinese texts are from the simplified version. We consulted one person in the NLP group of the University of Edinburgh, who said this leads to models' bad performance in traditional texts.

Thus, we choose the simplified version and convert all other variants to the simplified Chinese, which introduces the next question: how to do the conversion?

Generally, for most texts, there exists one-to-one mapping from other variants to the simplified version at the character level. We tried several online tools, but the result is not ideal. Wikipedia contains many terminologies, such as "hacker", and "electronegativity", which are translated into different characters across the variants. Such characters cannot be converted directly. We finally apply a conversion table from the wiki community to deal with such terminologies and then apply other package to do

one-to-one character mapping.

The other problem is Chinese word segmentation, Word segmentation plays a key role in natural language processing including counting and evaluation metrics. The existing four languages in XWikis use space as the separator between tokens. However, there is no space in a sentence for Chinese texts. In fact, Chinese word segmentation itself is a research question and there are many existing pre-trained models in different granularity. Another method, which is a more popular method for research on Chinese text, is to segment the sentence at the character level. In another word, treating each Chinese character as a token. To prevent noise from the word segmentation model, in this project, we apply the latter method, but it contains many problems, particularly in the encyclopedia domain. For example, one token "Edinburgh" will be translated into three Chinese characters " 爱丁堡 ". Here " 爱丁堡 " is a transliterated word, where three characters are used to simulate the pronunciation of Edinburgh. Each character has its own meaning but only merging them together can express "Edinburgh". If we segment the sentence into characters, such semantic information will be lost. As mentioned before, Wikipedia contains a large number of terminologies, names of books, movies, famous people, and so on. Most of them are made by phonetics translators, which leads to this problem being more severe.

A.3 Results for mDebert-v3

The results are shown in figure A.1.Note that there is no result for Document-to-One and the results for C larger than 4 since mDEBERT-v3 cannot receive the input more than 512 tokens.

A.4 Data for template training

Sum.Sub-Sentence	C=2	4
One-to-One	0.280	-
Strategy 1	0.277	0.375
Strategy 2	0.361	0.323
Merge	0.263	0.392
Sum.Sub-Sentence	C=2	4
One-to-One	0.358	-
Strategy 1	0.366	0.431
Strategy 2	0.437	0.418
Merge	0.332	0.372

Table A.1: Correlations between human annotators and mDEBERT-v3. Each cell shows the score for label entailment.

```
Aidsham Hall , which was finished in 1931 , is located in Sri Lanka and has an architectural style known as Tudor and Jacabian .
The architecture style of Adisham Hall is Tudor and Jacabian . Aidsham Hall is located in Sri Lanka . Adisham Hall was finished in 1931 .

Adisham Hall is located in Haputalethe , Sri Lanka , where Sri Jayawardenepura Kotte is the capital and the Tamil language is spoken .
Adisham Hall is located in Haputale , Sri Lanka . Sri Jayawardenepura Kotte is the capital of Sri Lanka . The language of Sri Lanka is the Tamil language .

The Alan B. Millar Hall was completed on 1st June 2009 and currently hosts the US Mason School of Business .
The Alan B. Millar Hall was completed on 1st June 2009 . The Mason School of Business is the tenant of the Alan B Miller Hall in the United States .

1634 : The Bavarian Crisis is written by Virginia DeMarce and Eric Flint and was preceded by Grantville Gazette III .
1634 : The Bavarian Crisis was preceded by Grantville Gazette III . 1634 : The Bavarian Crisis is written by Virginia DeMarce and Eric Flint .

A Severed Wasp , ISBN number 0-374-26131-8 , is currently in print .
A severed wasp can be found in print . A Severed Wasp has the ISBN number 0-374-26131-8 .

Alan Shepard was born in New Hampshire and he served as the Chief of the Astronaut Office .
Alan Shepard served as the Chief of the Astronaut Office . Alan Shepard was born in New Hampshire .

1634 : The Bavarian Crisis ' which was written by Eric Flint , is available in hardcover .
1634 : The Bavarian Crisis was written by Eric Flint . 1634 : Th Bavarian Crisis is available in hardcover .

Eric Flint is the author of 1634 : The Bavarian Crisis , which was preceded by 1634 : The Ram Rebellion .
Eric Flint is the author of 1634 : The Bavarian Crisis . 1634 : The Bavarian Crisis was preceded by 1634 : The Ram Rebellion .

AZ Alkmaar has has 17023 members and their ground is the AFAS Stadion .
AZ Alkmaar 's ground is the AFAS Stadion . AZ Alkmaar has 17023 members .

Alexander L. Wolf is the leader of the Association for Computing Machine which is the publisher of ACM Transactions on Information Systems .
Association for Computing Machinery is the publisher of ACM Transactions on information Systems . Alexander L. Wolf is the leader of the Association for Computing Machine .

Aenir , which was written by Garth Nix has the OCLC number 45644811 and the ISBN number 0-439-17684-0 .
The OCLC number of Aenir is 45644811 . Aenir was written by Garth Nix . The ISBN number of Aenir is 0-439-17684-0 .

Arena in Dublin part of Leinster is owned by Live National Entertainment .
Arena is owned by Live Nation Entertainment . Arena is in Dublin which is part of Leinster .

Akita Museum of Art is located in Akita , Japan .
Akita Museum of Art is located in Japan . Akita Museum of Art is an art museum in the city of Akita .

The book ' A Long Long Way ' with the OCLC number 57399346 is available in hardcover and has 292 pages .
A Long Long Way is 292 pages long . The OCLC number of A Long Long Way is 57392246 . A Long Long Way is available in hardcover .

Amdavad ni Gufa is located in Ahmedabad , in India , the country where T S Thakur and Sumitra Mahajan are leaders .
The leader of India is called T S Thakur and another leader there is Sumitra Mahajan . Amdavad ni Gufa is located in Ahmedabad in India .

The book , A Loyal Character Dancer , has the ISBN number of 1-56947-301-3 which has 360 pages .
The book , A Loyal Character Dancer , has the ISBN number of 1-56947-301-3 . A Loyal Character Dancer has 360 pages .

The Amsterdamsche Football Club Ajax Amateurs is the complete name for the AFC Ajax -LRB- amateurs -RRB- who played in the Topklasse in the 2014-2015 season .
Simples: AFC Ajax -LRB- amateurs -RRB- played in the Topklasse in the 2014-2015 season . The full name of AFC Ajax -LRB- amateurs -RRB- is Amsterdamsche Football Club Ajax Amateurs .

The novel A Long Long Way was followed by The Secret Scripture published by Faber and Faber .
The novel A Long Long Way was followed by The Secret Scripture . The Secret Scripture was published by Faber and Faber .

Birmingham , which is lead by the Labour politician John Clancy , is the home town of the architect John Madin who designed 103 Colmore Row .
Birmingham is lead by the Labour politician John Clancy. Birmingham is the home town of the architect John Madin.  John Madin designed 103 Colmore Row.

Eric Flint 's book , `` 1634 : The Bavarian Crisis , '' was preceded by `` Grantville Gazette II . ''
Eric Flint is the author of 1634 : The Bavarian Crisis . 1634 : The Bavarian Crisis was preceded by Grantville Gazette II .

Agremiação Sportiva Arapiraquense , managed by Vica , has 17000 members.
Agremiacao Sportiva Arapiraquense are managed by Vicas . Agremiação Sportiva Arapiraquense has 17000 members .

AFC Ajax -LRB- amateurs -RRB- , who have 5000 members played in the 2014 season.
AFC Ajax -LRB- amateurs -RRB- played in the 2014 season . AFC Ajax -LRB- amateurs -RRB- has 5000 members .

200 Public Square is located at Public Square , Cleveland and has a floor count of 45 .
200 Public Square is located at Public Square , Cleveland . 200 Public Square has a floor count of 45 .

Asser Levy Public Baths are in New York City of which New Netherland is a part .
The location of Asser Levy Public Baths are New York City . New York City is a part of New Netherland .

Alfred Giles , architect of The Asher and Mary Isabelle Richardson House died in Kendall County , Texas .
Alfred Giles was the architect of Asher and Mary Isabelle Richardson House . The architect Alfred Giles died in Kendall County , Texas .

Max Huiberts owns AZ Alkmaar which has 17023 members .
Max Huiberts owns AZ Alkmaar. AZ Alkmaar has 17023 members .
```

Figure A.2: Part of training samples for template learning

Appendix B

Screenshot for human evaluation

Figure B.1: Screenshot for human evaluation webpage