Automatic Targeted Evaluation for Document Level NMT

Bachelor's Thesis of

Sicheng Dong

Artificial Intelligence for Language Technologies (AI4LT) Lab
Institut for Anthropomatics and Robotics (IAR)
KIT Department of Informatics

Reviewer:	Prof. Dr. Jan Niehues
Second reviewer:	Prof. Dr. Alexander Waibel
Advisor:	M.Sc. Sai Koneru

29. May 2023 – 29. September 2023

Karlsruher Institut für Technologie
Fakultät für Informatik
Postfach 6980
76128 Karlsruhe

I declare that I have developed and written the enclosed thesis completely by myself, and have not used sources or means without declaration in the text.

PLACE, DATE

.....Sicheng Dong.....

(Sicheng Dong)

Abstract

Machine translation, the technique to automatically translate text from one natural language to another, has experienced a revolution with the introduction of Neural machine translation (NMT), which employs neural networks to generate translations[33]. Despite its great performance[38], most NMT translation models operate on the sentence-level due to the lack of document-level metadata and proper evaluation metrics, leading to potential ambiguity problems[23][29].

Therefore, this study aims to automatically construct a testset containing sentences that need context to disambiguate from document-level parallel corpus to evaluate context-aware models' capability. Our primary focus was on the translation direction from Chinese to English. In the research, we used random evaluation to assess the quality of the testset we got. We found two main reasons leading to contextual ambiguity, namely tense difference and subject difference. We extracted at least 165 sentences having tense difference and 146 sentences demonstrating subject difference. 70% of the extracted tense difference sentences are contextual ambiguous, while 40% of the extracted sentences for subject difference are contextual ambiguous. Compared with the initial rate for contextual ambiguity, it is a high increase.

Apart from this, we also conducted an initial start in the direction from English to Chinese. We proposed reasons leading to contextual ambiguity, namely subject difference as well as the Formality problem.

Further study is required in two aspects. In the direction of EN->ZH, despite the increase of the rate of contextually ambiguous sentence, the size of the testset is too small. In the direction of ZH->EN, it is necessary to continue to extract contextually ambiguous sentences to contruct the final testset.

Zusammenfassung

Die maschinelle Übersetzung beschreibt die Techinik, Text automatisch von einer natürlichen Sprache in eine andere zu übersetzen. Sie hat eine Revolution wegen der Einführung der neuronalen maschinellen Übersetzung erlebt. Die neuronale maschinelle Übersetzung verwendet neuronale Netzwerke, um Translationen zu generieren. Trotz ihrer hervorragenden Leistung, arbeiten die meisten neuronale maschinelle Übersetzungsmodelle auf Satzebene, da Metadaten auf Dokumentebene und geeignete Bwertungsmetriken fehlen, was zu potenziellen Mehrdeutigkeitsproblemen führen.

Ziel dieser Studie ist, aus einem parallelen Korpus auf Dokumentenebene automatisch einen Testsatz zu erstellen, der Sätz enthält, die zu ihrer Disambiguierung Kontext benötigen. Mit diesen Testsatz können wir die Fähigkeit Kontextbezogner Modelle bewerten. Unser Hauptsaugenmerk lag auf der Übersetzungsrichtung von Chinesisch ins Englisch. In der Forschung haben wir eine Wir fanden zwei Hauptgründe für kontextuelle Mehrdeutigkeit, nämlich den Unterschied in der Zeitform und den Unterschied im Subjekt, und extrahierten schließlich 165 Sätze mit Zeitformunterschied und 146 Sätze mit Unterschied im Subjekt. 70% der extrahierten Sätze mit Zeitformunterschied sind kontextuell mehrdeutig, während 40% der extrahierten Sätze mit Unterschied im Subjekt kontextuell mehrdeutig sind. Im Vergleich zur ursprünglichen Rate für kontextuelle Mehrdeutigkeit is es eine deutliche Steigerung. Aber die Größe des Testdatensatzes bleibt jedoch relativ klein.
Außerdem haben wir auch einen ersten Ansatz in der Übersetzungsrichtung von Englisch nach Chinesisch dirchgeführt. Wir haben Gründe für kontextuelle Mehrdeutigkeit dargestellt, nämlich auch den Unterschied im Subjekt sowie das Formalitätsproblem.

In zwei Bereichen besteht weiterer Forschungsbedarf. In der Richtung EN->ZH ist die Größe des Testsatzes relativ zu klein, obwohl Rate der kontextuellen Mehrdeutigkeit sich steigert. In der Richtung ZH->EN is es notwendig, weiterhin kontextuell mehrdeutige Sätze zu extrahieren und das Ergebnis statistisch zu analysieren.

Contents

List of Figures

List of Tables

1 Introduction

1.1 Motivation of Importance

Machine translation (MT) is a sub-field of computational linguistics that investigates the use of software to translate text or speech from one natural language to another[20]. MT plays a crucial role in today's multicultural and globalized society, enabling communication and information dissemination across language barriers. Neural machine translation (NMT), as one of the most promising machine translation approaches, illustrates a high performance score on public benchmarks (Bojar et al., 2016) and rapid adoption in deployments by, e.g., Google (Wu et al., 2016), Systran (Crego et al., 2016), and WIPO (Junczys-Dowmunt et al., 2016)[18].

Current NMT models work on the sentence level, which means each sentence is translated alone. When translating documents, all dependencies between sentences in the document are ignored. This leads to potential ambiguity problems[34]. The most common examples of context-dependent phenomena problematic for MT are coreference (Guillou, 2016), lexical cohesion (Carpuat, 2009), and lexical disambiguation (Rios Gonzales et al., 2017)[3]. As shown in Table 1.1, we take an example of a sentence pair between English and Chinese. For the same Chinese sentence, there are two different English translations. In both English translations, there is a verb in the second sentence which has the same meaning in Chinese and the only difference is the time tense. When the second sentence is translated from Chinese to English, the translation of this verb is ambiguous. It is obvious that there are no words in the current sentence that indicate tense, and the correct tense can only be determined by the antecedent. In these two English sentences, the antecedent "lives" and "lived" show the right time tense.

Table 1.1: Example sentence pair to illustrate how the word is ambiguous based on the context given. Ambiguous words are in bold and the words in context that show the time tense are in italics .

Chinese Sentence	English Translation
他住在这里。他在这里**工作**。	He *lives* here. He **works** here. He *lived* here. He **worked** here.

Compared to sentence-level NMT, document-level NMT performs at the document level and greatly improves the translation quality by conserving the connectivity between sentences in the whole doc-

ument. For example, the lexical cohesion mentioned above is captured by document-level NMT in the translation context and can help solve pronoun translation requiring context outside the current sentence[23].

In practice, however, document-level NMT faces many challenges. The biggest challenge is that existing training data does not have the document metadata that is needed for document training, placing an impediment at the very start of any effort. Moreover, contextual models and baseline models show a minimal difference in document-level metrics like BLEU[27] or COMET[30]. Therefore, it is challenging to find a method to verify the accuracy of a model [23][29].

To solve the problem, in this thesis, we aim to build up a test set that allows us to specifically measure a model's capability to correctly translate ambiguous words based on the context. The test suite consists of those contextually ambiguous sentences.

1.2 Research Question

RQ **How can we extract targeted test sets from document-level parallel training data to reliably and accurately evaluate context-aware NMT systems with minimal human annotation?**

- This question comes from the recognition of the limit of sentence-level translation as well as the challenges faced by document-level translation. We first extract all source sentences that have multiple different translations as the starting point.

  Then, to address the main research question above, we separate it into 3 sub-questions:

  * **How many ambiguous sentences are there in the EN-ZH document-level parallel datasets that need context for disambiguating?**

    · This question is very critical and should be done at the beginning of the study. This gives us a rough idea of the extent to which the entire database needs document-level translation. We did this by calculating the distribution of translations for different quantities and conducting a random evaluation to estimate the rate of contextual ambiguity. The related work can be seen in the section 3.1.

  * **What heuristics can be used to extract sentences that need context to disambiguate?**

    · There are a large number of ambiguous sentences, and it is clearly not feasible to iterate them to find contextually relevant ambiguity. Proper heuristics are needed. We do the extraction by exploring the reasons leading to contextual

ambiguity and extra based on the sentence similarity score. The related work can be seen in section **??** and section 3.3.

* **What heuristics can be applied to filter out non-contextual ambiguous sentences ?**
    - After extraction, some non-contextual ambiguous sentences can still exist. Finding ways to filter them can enhance the quality of the testset. We do the filtration by analyzing the non-contextual reasons for ambiguity and employing a language model. The related work can be seen in the section 3.4 and section 3.5.

1.3 Thesis Structure

In Chapter 1 we emphasized the importance of neural networks and the ambiguous problems that could be encountered when only translating at the sentence level and the challenge of document-level NMT when evaluating context-based parallel datasets. Chapter 2 introduced core principles in Machine learning, Machine translation, Neural machine translation, and Evaluation metrics, including Sequence-to-sequence models, Recurrent neural networks, BLEU scores, etc. which are fundamental for this thesis.

In Chapter 3 we displayed the construction of the Chinese-to-English test step by step as well as pointed out the problems encountered. The entire dataset construction process consists of 5 steps: Dataset evaluation, Reasons for ambiguity, Extraction of contextual ambiguity, Further filtration, and Language model filtration. Then we conducted an initial start for the construction of the testset in an opposite direction (English -> Chinese), including the fundamental characteristics of datasets and the potential reasons leading to contextual ambiguity.

Finally, in Chapter 4 we summarised the findings and outlined future research directions and potential speculations.

2 Background & Related Work

Three lines of work are related to our paper: the basic knowledge and network types of Artificial Neural Networks (described in Section 2.1), the fundamental knowledge about Machine Translation, especially Neural Machine Translation as well as several neural machine models (described in Section 2.2), and research focused on how to evaluate the translation. (described in Section 2.3)

2.1 Artificial Neural Network

Artificial Neural Networks, abbreviated as ANNs or neural networks, are mathematical or computational models that mimic the structure and functionality of human biological neural networks in the fields of machine learning and cognitive science. They are used to estimate or approximate functions[1].

The basic component of one Artificial Neural Network is called Artificial Neuron. As shown in Figure 2.1, a simple Artificial Neuron consists of input, weights, bias, activation function, and output. In the field of natural language processing, the input can be a single word or sentence. For each input, a weight is assigned to measure the degree of influence of that input on the neuron. Input values are multiplied by their corresponding weights and then summed to form a weighted sum. Bias is another constant value added to the weighted sum to adjust the threshold. The activation function is a non-linear part of the artificial neuron and it takes the sum of bias and weighted sum to generate the output.

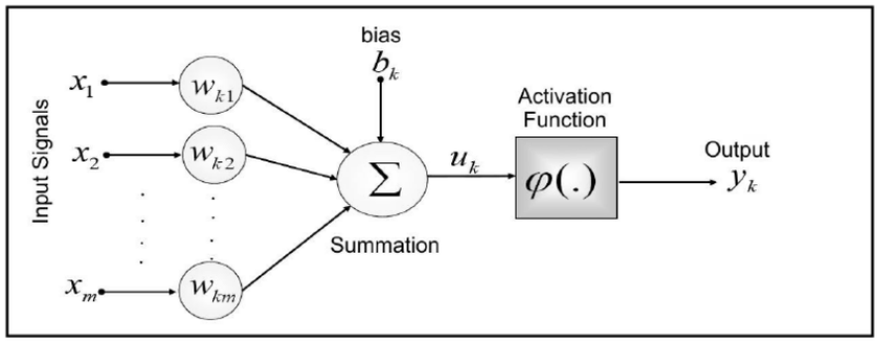

Figure 2.1: A simple artificial neuron structure
[39]

An Artificial Neural Network comprises multiple Artificial Neurons and employs different topological structures to process various data types. Typically, a Neural Network can be divided into three layers: input layer, hidden layer, and output layer. The two most common topological structures of Neural Networks are depicted in Figure 2.2 Left Feedforward Neural Network (FNN)[4] and Figure 2.2 Right Recurrent Neural Network (RNN)[31].

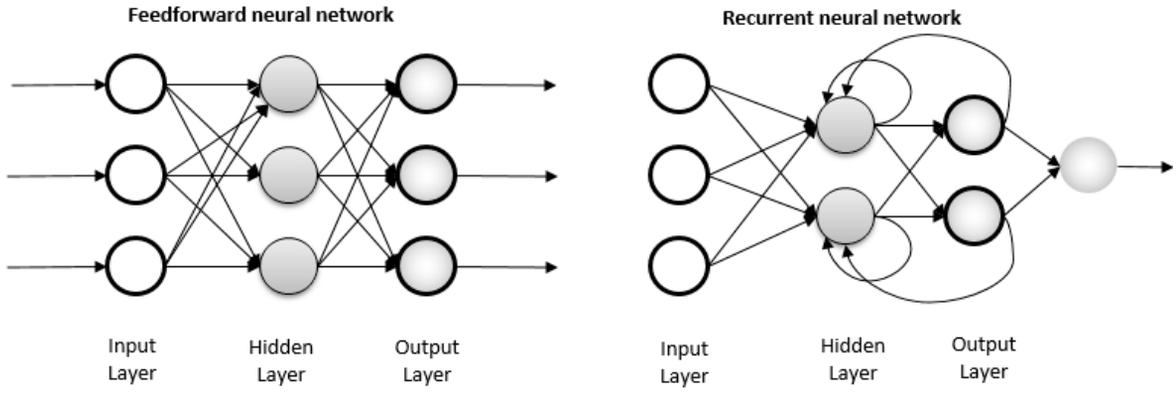

Figure 2.2: Feedforword neural network and Recurrent neural network [28]

FNN is one of the simplest network architectures. It contains one input layer, one or multiple hidden layers, and one output layer. The information flows unidirectionally from input to output without any recurrent connections[32]. However, when the task is to predict the next word in one sentence, simple FNN faces challenges since it lacks key information about the previous words. Instead, RNN has recurrent connections with hidden states before. At any given time step t, its output not only depends on the current input at time t but also relies on the hidden state generated at the previous time step [17]. And the hidden state before storing the sequence information.

However, RNN has also its limits. When the input sequence is too long, the gradient used to update the weights gets smaller and smaller since it has been biased many times. This leads to the vanishing gradient[37]. To solve this problem, a special kind of RNN called Long Short Term Memory (LSTM) was applied. It comprises three gate structures: an input gate, an output gate, and a forget gate. These gates control how much information should enter, leave, or be forgotten[42].

2.2 Machine Translation

Machine Translation (MT) is the use of automated software that translates text without human involvement[21]. However, this task is inherently challenging due to the flexibility and ambiguity of language, making it difficult to determine a single best translation for a given sentence.

Basically, the machine translation metrics can be divided into two categories: Rules-based Machine Translation and Data-driven Machine Translation. Furthermore, Data-driven machine Translation contains two main strategies, namely Statistical Machine Translation (SMT) and Neural Machine Translation. (NMT)

The main idea of Rule-based Machine Translation is to get the translation by introducing the semantic, syntax, and linguistic knowledge of both source and target language[24]. Statistical Machine Translation aims to build up a statistical model by analyzing a large amount of parallel language corpus to do the translation. It turns the translation problem into a possibility problem: Given the source language S, what's the conditional possibility of target language T[6]. And it uses the learned model to maximize the conditional possibility of getting the optimal translation result. Practical implementations of SMT are generally phrase-based systems (PNMT) which translate sequences of words or phrases where the lengths may differ[44].

In this thesis, we mainly focus on neural machine translation.

2.2.1 Neural Machine Translation

Neural Machine Translation is the development of Statistical Machine Translation. It uses Neural Networks to learn the translation model instead of analyzing the large parallel language corpus. Unlike the discrete representation of statistical machine translation, Neural Machine Translation uses continuous space representation to indicate the words, phrases, or sentences[2]. When constructing a translation model, it relies totally on Neural Networks to map from source sentences to target sentences[44].

Here we introduce multiple basic NMT models below:

2.2.2 Sequence to Sequence Model

The Sequence-to-sequence model (Seq2seq) aims at mapping an input sequence to an output sequence[8]. The most common structure is the encoder-decoder architecture, as shown in Figure 2.3.

The Encoder accepts input sequences (source sentences) and translates them into a fixed-length state vector. The State Vector contains all the information about the input sequence. The Decoder then reads the state vector and generates the output sequence step by step. At each time step, the output of the Decoder is used as input to the next time step[11].

The common implementation of the Encoder and Decoder is to use LSTM, as shown in Figure 2.4.
So for the input x_1, x_2, x_3 at the time step t, let the f represent the LSTM transition and the hidden

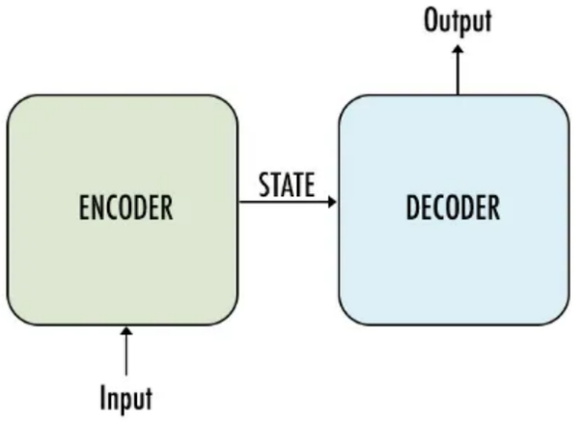

Figure 2.3: Encoder-Decoder Model
[11]

state ht of them is[14]:

$$h_t = f(x_t, h_{t-1})$$

Let the q represents the Encoder vector calculation function and the Encoder vector c is[14]:

$$c = q(h_1, h_2, h_3)$$

The Decoder then outputs the result at time $t'$with $y_{t'-1}$representing the output at last time step and $s_{t'-1}$representing the hidden state at last time step, g representing the LSTM transition[14]:

$$s'_{t'} = g(y_{t'-1}, c, s'_{t'-1})$$

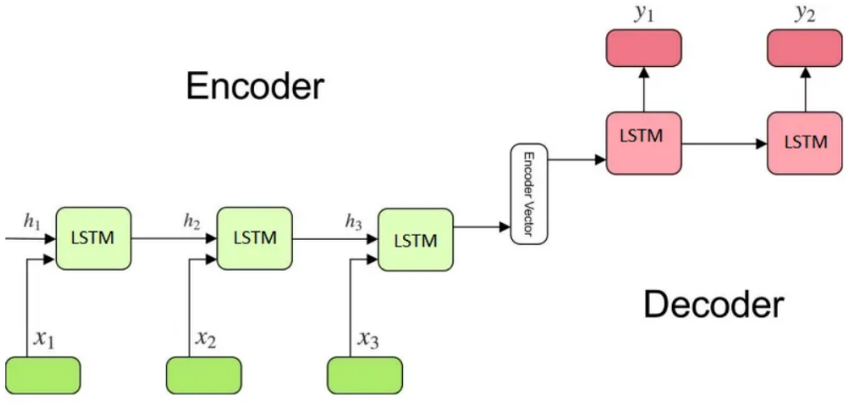

Figure 2.4: LSTM Implementation for Encoder-Decoder model
[12]

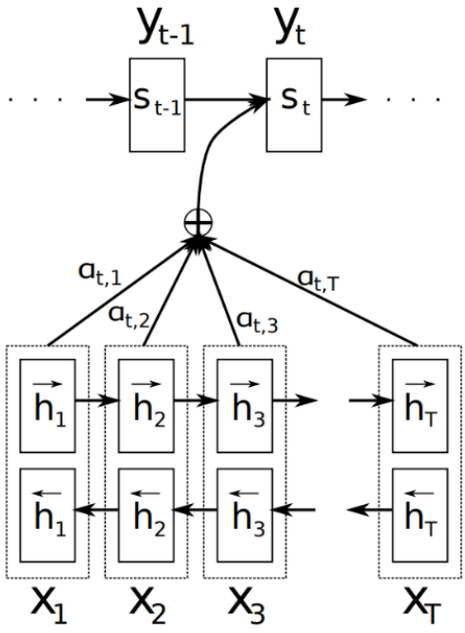

Figure 2.5: Illustrate of trying to generate the t-th target word y_t given a source sentence $(x_1, x_2 \dots x_T)$ [2]

2.2.3 Attention

However, as you may have noticed, the most critical Encoder vectors are only determined by the last hidden state and the last input, no matter how long the sequence is. When the sequence is too long, the information may easily be distorted, which affects the performance[40].

To solve this problem, we introduce a new mechanism called Attention to Seq2Seq model[38]. It enables the model to focus on different parts of the input sequence when generating each output. In the Encoding phase, the Encoder not only transmits the last hidden state but also transmits all the previous hidden states to the decoder. When decoding, Attention gives each input hidden state a weight to calculate the weighted. As shown in Figure 2.5, $h1, h2, \dots, h_T$ are all encoder states and S_t is the decoder state at time step t and $\alpha_{t,1}, \dots, \alpha_{t,T}$ are the weights assigned to them. y_t it the weighted sum from them. The weights can be calculated by a special function called *score*. We apply it for each encoder hidden states and the result of this function is called Attention scores. We apply $softmax$ function on Attention scores to get the weights mentioned above[40].

2.2.4 Transformer

Transformer is one of the most popular language models that use Attention in recent years[38].It relies only on the Attention to process sequence data and its translation quality and speed greatly exceeds that of other models in the same time.

Two special attention mechanisms are used here, namely Self-Attention and Multi-Head Attention. Self-Attention allows transformer to assign various weights to each input position, so that it can focus on different parts of the given sequence. 'Self' means that it focus on relationships in the same sequence. Multi-Head Attention introduces multiple Self-Attention heads. Each attention head has a different set of weight assignments, which makes Transformer capable of observing various aspects of the same input sequence.

As illustrated in Figure 2.6, Transformer is also a kind of Seq2seq model with the Encoder and Decoder architecture. Unlike RNN, it doesn't process input sequentially, so an additional position information for each word is calculated by Positional Encoding before sending into the Encoder[9]. It uses Multi-Head Attention in both Encoder and Decoder and other layers like Add & Norm layers, Feed Forward layers are also included in the model to improve the performance of the structure.

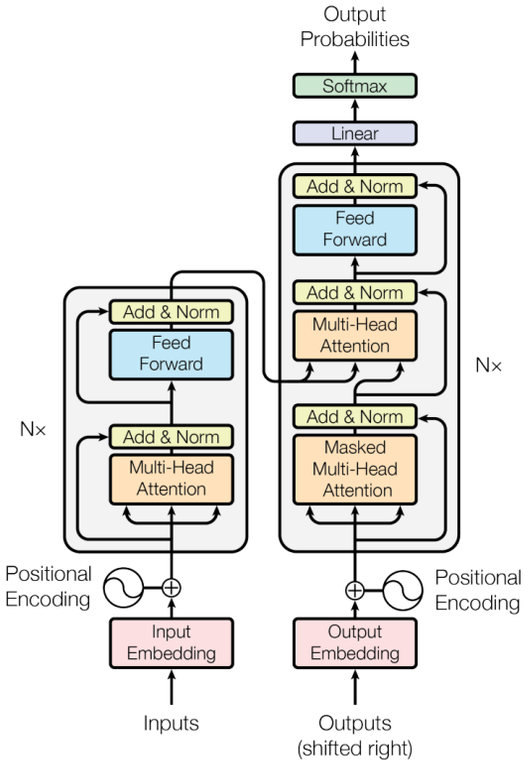

Figure 2.6: Transformer model
[38]

2.3 Language Model

Language Modeling is widely used in various areas of NLP, including text conversion, speech recognition, etc. Language modeling uses techniques such as statistical analysis or neural network to predict the probability of occurrence of a given text in a sentence[19].

Different algorithms and network structures are applied, such as RNN, LSTM or Encoder-Decoder mentioned before. Among all neural Language modelling, two language models are most popular, one is the Casual Language Model and the other is the Masked Language Model[22]. We focus mainly on the first one.

The Causal Language Model is used to predict the possibility of next word based on the given context[41]. Its mechanism is very similar to the decoder mechanism in a Transformer, and it excludes the influence of the later tokens on the former ones. When each token calculates the probability, it will only take into account the content on the left side of the tokens, and will not take into account the content of the position after it[7].

Famous Causal Language Models include model like GPT-2 and etc[15]. In this thesis, we use language model called DistilGPT2, which is an pre-trained language model under the smallest version of Generative Pre-trained GPT-2[13]. It is expected to be smaller and easier to run compared with the baseline model.

2.4 Evaluation

2.4.1 Sentence Similarity

Sentence similarity is a measure of the extent to which two sentences express the same meaning. It is widely used in fields like duplicate recognition, paraphrase and so on[26].
One of the most used metrics is called cosine similarity. Cosine similarity measures the similarity between two vectors of an inner product space[16]. Then it calculates the cosine of the angle between two vectors and assesses if two vectors are in the same direction. The detailed formula is below with the vectors A and B:

$$\text{Cosine Similarity}(\mathrm{A}, \mathrm{B}) = \frac{\mathrm{A} \cdot \mathrm{B}}{\|\mathrm{A}\|\|\mathrm{B}\|}$$

The result of the cosine similarity ranges from -1 to 1. When the score equals 1, it means the direction of two vectors is entirely the same and thus, the words or sentences in comparison are highly similar. When the score approaches 0, the words or sentences given have no similarity. When a score is smaller than 0, it indicates that they are dissimilar[10]. As shown in Figure 2.7, in the left part, France is considered to be similar to Italy, so the score is 1. while on the right side, the ball and crocodile are not relevant, so the score is 0.

2.4.2 BLEU Scores

BLEU represents bilingual evaluation understudy, which is an algorithm for evaluating the translation quality of a given text made by machines. The main idea of this mechanism focuses on the degree of

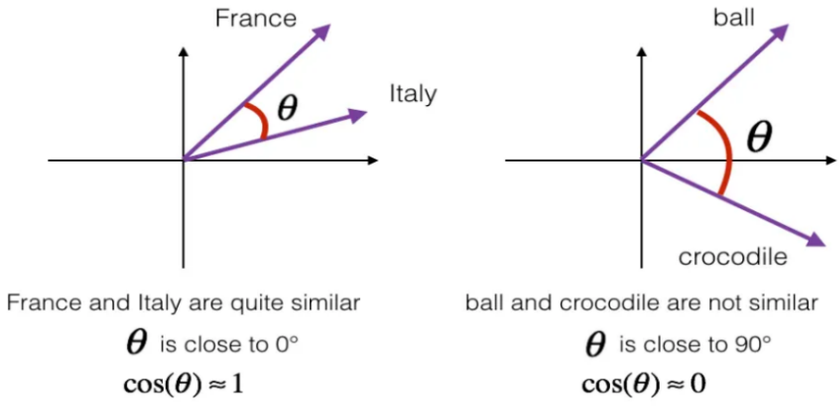

Figure 2.7: Consine scores for two situations
[25]

closeness between machine translation and human translation[5].
The BLEU calculation formula is below[35]:

$$\text{BLEU} = \text{BP} \times \exp\left(\sum_{i=1}^{n} w_i \log(p_i)\right) \tag{2.1}$$

- BP(Berevity Penalty): Factor used to reduce the impact of sentence length on BLEU scores
- w_i: Weight factor of each n-gram
- p_i: Precision of i-gram, represents the number of i-grams in machine translation divided by the maximal number of i-grams in the reference translation.

2.4.3 Contrastive Evaluation

Contrastive evaluation is another approach to assessing a model's performance. Unlike the BLEU metric above, it doesn't focus on the translation itself. It tests a model's ability to distinguish between good translations and bad translations[23]. As Input, we give two sentence pairs: (source sentence, target sentence) and (source sentence, contrastive sentence) and use a language model to calculate the score for both. Only when the target sentence has a higher score than any other contrastive sentences, the target sentence will be considered as a good translation[23].

3 Method

Check https://github.com/OscarDDD/Testset-for-contextual-ambiguity for the entire testset.

This section illustrated the approach to extracting sentences that need context to disambiguate from ZH-EN document-level parallel datasets. We mainly focused on the ZH->EN direction. The whole process is below:

1. First, the used datasets were evaluated. Details can be seen in section 3.1.
2. After that, we explored the reasons leading to ambiguity and extracted those matching sentences. Details can be seen in section 3.2.
3. Next, we explained the approaches to extracting those contextually ambiguous sentences from sentences we got in step 2 and analyzed the non-contextual reasons leading to ambiguity. Details can be seen in section 3.3.
4. Then, we filtered out sentences related to the reasons in the previous step. Details can be seen in section 3.4.
5. Finally, we employed a causal language model to further filter out those non-contextual ambiguities. Details can be seen in section 3.5.

At the end of this section, we also provided an initial start for the EN->ZH direction. All data is from the **open parallel corpus (OPUS)**[36] and **WMT23 Discourse-Level Literary Translation (Literary)**[43].

3.1 Datasets assessment

In this section, we performed some basic assessments of the datasets to establish their fundamental characteristics and checked if they could be used for the research below.

As shown in Table 3.1, we first checked for each file in datasets if the total number of lines equals the sum of the number of duplicated lines and number of unique lines to ensure the completeness and correctness of the datasets

File name	Total number of lines	Number of duplicated lines	Number of unique lines
OPUS.eng	17, 451, 546	5, 236, 893	12, 214, 653
OPUS.zho	17, 451, 546	4, 702, 524	12, 749, 022
Literary.eng	1, 939, 187	219, 646	1, 719, 541
Literary.zho	1, 939, 187	218, 377	1, 720, 810

Table 3.1: The table indicates the line number, duplicated line number and unique line number of datasets. *eng* means files contains English sentences and *zho* means files contain Chinese sentences.

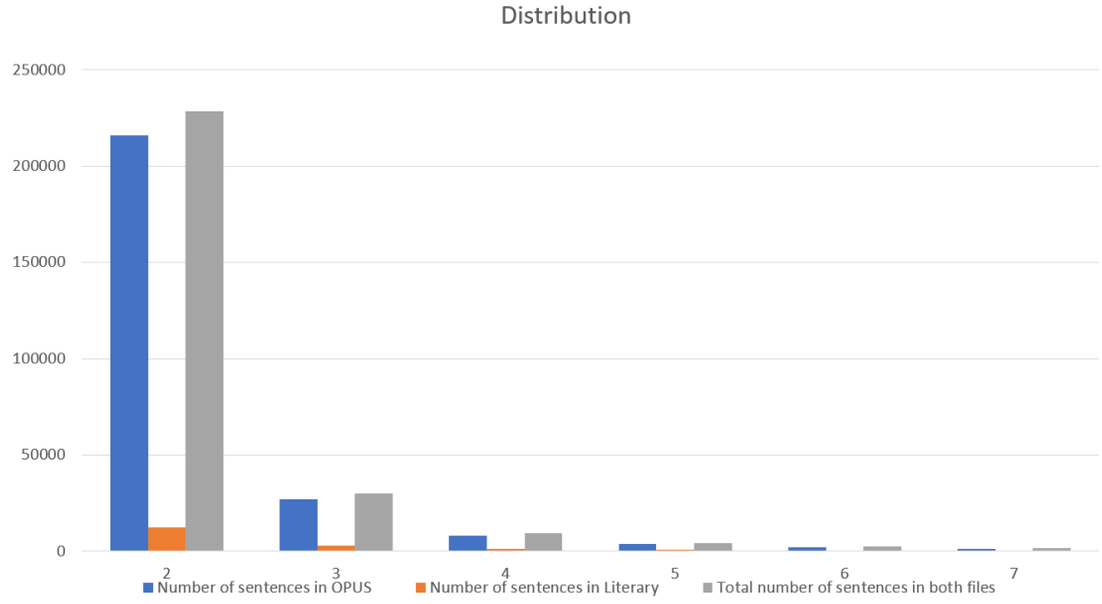

Figure 3.1: The figure indicates the distribution of different varieties of translations. The X-axis represents the number of different English translations for Chinese sentences, and the Y-axis represents the number of such Chinese sentences.

Then, to extract context-based ambiguous sentences, we need to first ensure that the given datasets maintained the document structure, rather than being randomly shuffled during translation. Otherwise, the context was meaningless. We checked it manually and found out that both OPUS and Literary were document-level, so they were used in the research below.

As illustrated in Table 3.2, we subsequently analyzed the entire corpus and extracted a set of instances where a single Chinese sentence corresponded to multiple English translations. In total, we got 279,991 Chinese sentences 260,981 for OPUS and 19,010 for Literary, as illustrated in This was also the **start point** of our research. We aimed to extract contextual ambiguous sentences and filter out non-contextual ones from this set.

FileName	Number of Chinese sentences
OPUS	260,981
Literary	19,010

Table 3.2: Number of Chinese sentences that have multiple various translations

As shown in Figure 3.1, we then split the set of ambiguous sentences into groups based on the number of translations of one sentence to obtain the distribution of translations for different quantities. There were 215,920 sentences in OPUS and 12,325 sentences in Literary that had 2 different varieties of English translations. It was obvious that most ambiguous Chinese sentences had two different English translations, so the research below was focused on the ambiguous sentences with 2 varieties of translations.

Before applying any methods, we conducted an initial assessment of the rate of those contextual ambiguous sentences on the entire dataset. We did it by randomly choosing ten sentences and assessing them manually, as shown in Table 3.3. Both initial rates were 0%. This was caused by various reasons, such as synonyms or translation errors, etc, which did not need context to classify. Although both rates here are 0, the datasets still can contain a certain amount of contextual ambiguity due to the large number of ambiguous sentences.

FileName	Contextual	Non-contextual	Initial Rate
OPUS	0	10	0%
Literary	0	10	0%

Table 3.3: Initial Rate of contextual ambiguous sentences

3.2 Reasons for the ambiguity

We discovered the possibilities leading to different translations through two methods. The first one involved speculation on ambiguous words based on the syntax, semantics, and linguistic rules of both languages. The second method involved calculating the similarity score of the ambiguous sentences and detecting inconsistencies within the pairs that had a higher score. Sentences with a lower score could be just

A higher score indicated that the two sentences had less difference, making it easier for us to observe commonalities. For those

After assessing multiple instances, we discovered two possible reasons for sentences that caused contextual ambiguity. The first one was the difference in tenses. In Chinese, verbs did not have any tense changes, while in English, verbs should change depending on time tense, as illustrated in Table 1.1 before. The sentence itself should not contain time information explicitly, it must rely on context for classification.

The second reason involved the differences of subjects between the two translations. When Chinese sentences did not explicitly mention a specific subject or mention a subject that had different meanings, English translations introduced one new subject or a word showing one of its meanings based on the context information. As illustrated in Table 3.4, Chinese sentences only conveyed information in reference. However, translations inserted subjects ”he” and ”she” based on contextual antecedents, namely ”father” and ”girlfriend”. Obviously, in this situation, the sentence itself also comprised nothing about the subject, so the context was required.

source:	”在哪里?”	有请代表
variety1:	” Where is **he** ? ”	Welcome the **representatives**.
variety2:	” Where is **she** ?”	Welcome the **representative**.
context1:	**He** is looking for his father urgently.	5 **representatives**come here for the meeting.
context2:	**He** is looking for his girlfriend urgently.	Prof. Li is here as the **representative** today.
reference:	” Where ? ”	Welcome **representative** || **representatives**.

Table 3.4: Subject Difference Example

In order to validate that the two scenarios mentioned above were not exceptions, we extracted all sentences matching the phenomena and statistically counted the number of ambiguous sentence pairs with different tenses and subjects. In total, we got 20,786 sentences for tense difference and 9,409 sentences for subject difference. It was important to mention that all the sentences found here did not necessarily have real contextual tense and subject differences since they might contain detection problems or other bad translations. Details were in the next section 3.3.

Filename	Tense Difference	Subject Difference
OPUS	18,979	8,489
Literary	1,807	920

Table 3.5: Number of Chinese sentences that have tense or subject differences

3.3 Extraction of contextually ambiguity

In this section, we started from the above extracted ambiguous sentences containing issues with tenses or subjects and calculated the cosine similarity between pairs of them. We focused on finding the similarity score where most of the contextually ambiguous sentences are found.

Before calculating, we first normalized the sentences to exclude the influence of ambiguous words on our overall characteristics of sentences. This involved converting all verbs into their original form and replacing all subjects with the word ”Subject”, as shown in Table 3.6 . Here ”acknowledged” was transformed to ”acknowledge” and the ”expression” was transformed to ”Subject”.

Original sentence:	We **acknowledged** this.
Sentence after Time Normalization:	We **acknowledge** this.
Original sentence:	The **expression** on his face is interesting.
Sentence after Subject Normalization:	The **Subject** on his face is interesting.

Table 3.6: Examples of normalization tense difference and subject difference

After that, we selected multiple similarity thresholds and divided the normalized datasets into different threshold groups. We conducted a random evaluation like before for each of them. We had set four thresholds here, which were 0.9, 0.95, 0.9, and 1.0 respectively, and got the number of sentences for each threshold set as well as the contextual ambiguous rate.

	OPUS	Literary	Contextual sentence rate for OPUS	Contextual sentence rate for Literary
Threshold set (0.9)	13,885	360	10%	0%
Threshold set(0.95)	11,665	155	20%	10%
Threshold set(0.98)	8,214	77	20%	30%
Threshold set(1.00)	1,534	32	50%	20%

Table 3.7: Number of sentences and contextual sentence rate in different threshold sets for tense difference

As shown in Table 3.7, the group with similarity 1 contained the highest proportion of contextual ambiguous sentences in OPUS, at 50%. In Literary, based on the graphs alone, the group with a similarity of 0.98 got the best rate, however, we still considered groups with similarity 1 to perform the best, because most of the contextual ambiguous sentences found in 0.98 had actually a score of similarity 1. Obviously, OPUS demonstrated a higher rate compared to Literary, which was due to what the dataset contained. OPUS comprised multiple written documents such as legislative acts or meeting records, while Literary consisted of novels. Written documents often offered a clearer context to help decide the tense. In contrast, Novels contained large amounts of dialogue or thoughts from people, making it hard to tense decisions.

There were two main non-contextual reasons that led to ambiguity. The first reason was due to wrong verb tense detection. As shown in Table 3.8, when verbs were used as participial adjectives or clausal complement, morphological changes were not caused by tenses. In the first example, ”relating” and ”related” here were both used as adjectives to describe the antecedent ”homework”, while in the second example, the morphological change between ”playing” and ”play” was due to prepositional combination. The second reason was that the context did not provide enough tense information to distinguish between the two situations. In addition to the problems above, there were some other translation flaws like case differences, redundant translations, etc.

participial adjectives:	The **homework** relating to this is here.
	The **homework** related to this is here.
clausal complement:	**He** likes to play basketball.
	He likes playing basketball.

Table 3.8: Non-contextual reasons for tense difference

	OPUS	Literary	Contextual sentence rate for OPUS	Contextual sentence rate for Literary
Threshold set (0.9)	5,421	261	10%	0%
Threshold set(0.95)	4,057	151	20%	20%
Threshold set (0.98)	2,784	78	10%	10%
Threshold set(1.00)	1,255	52	10%	30%

Table 3.9: Number of sentences in different threshold set for subject difference

As illustrated in Table 3.9, the overall accuracy for the subject difference was consistently low, and the distribution was quite even, ranging from around 10% to 30%. The main reason leading to non-contextual ambiguity here was the unrecognizable distinction between subjects. For example, considering pronouns and synonyms, they were sometimes replaceable by each other, no matter what

context was given. As illustrated in Table 3.10, relative pronouns like ”that” and ”which” in the first example, demonstrative pronouns like ”This” and ”That” in the second example as well as synonyms like ”Reform” and ”Change” in the third example, they were not classified under most circumstances.

relative pronouns:	It is the flower **that** I bought yesterday.
	It is the flower **which** I bought yesterday.
demonstrative pronouns:	**This** is true.
	That is true.
synonym:	**Reform** will be seen.
	Change will be seen.

Table 3.10: Non-contextual reasons for subject difference

3.4 Further Filtering

As a result of the previous analysis, certain kinds of non-contextual ambiguous sentences occurred frequently. In this section, we filtered out those sentences to enhance the quality of the testset. Here we focused mainly on the threshold set that has a similarity score of 1 since it had a higher rate and sentences had fewer differences, making it easier to do more processing.

For tense differences, we filtered out the circumstances when verbs were used as participial adjectives and clausal complements. As shown in Table 3.15, after filtering, there were a total of 18,040 sentences left in OPUS, with 1,386 sentences having a similarity score of 1. For WMT Literary, we got 328 sentences and 32 of them had a similarity score of 1. The evaluation rate for OPUS rose from 40% to 60%, while the rate for WMT Literary remained the same since no sentence was filtered out.

	OPUS	Literary
Total number of sentences that have two varieties of translations	18,040	1,775
Threshold set (0.9)	13,884	328
Threshold set (0.95)	11,324	155
Threshold set (0.98)	8,022	76
Threshold set (1.00)	1,386	32

Table 3.11: Numbers of ambiguous sentences for tense difference after filteration

Name	Threshold	Non-contextual	Contextual	Good example rate
OPUS	1	4	6	60%
Literary	1	6	4	20%

Table 3.12: Rate of contextual ambiguous sentences for tense difference in threshold 1

	OPUS	Literary
Total number of sentences that have two varieties of translations	8788	857
Number of sentence (0.9)	4,821	226
Number of sentence (0.95)	3,498	127
Number of sentence (0.98)	2,279	65
Number of sentence (1.00)	962	43

Table 3.13: Numbers of ambiguous sentences for subject difference after filteration

Name	Threshold	Non-contextual	Contextual	Good example rate
OPUS	1	8	2	20%
WMT Literary	1	7	3	30%

Table 3.14: Rate of contextual ambiguous sentences for subject difference in threshold 1

	OPUS	Literary
Total number of sentences that have two varieties of translations	2420	566
Number of sentence (0.9)	1,455	168
Number of sentence (0.95)	1014	95
Number of sentence (0.98)	587	46
Number of sentence (1.00)	179	32

Table 3.15: Number of ambiguous sentences for each threshold set due to subject differences after filtering synonyms

For subject, we first ruled out ambiguity due to similar pronouns. As illustrated in Table 3.13, there were a total of 8,788 sentences left in OPUS, with 962 sentences having a similarity score of 1. For WMT Literary, we got 857 sentences and 43 of them had a similarity score of 1. However, the evaluation rate for OPUS and Literary remained the same since there were still large amounts of synonyms and other problems.

We initially attempted to eliminate synonyms by using NLTK method, but this approach was proved to be ineffective since the detection method was problematic. It was so strict that the good ones were also filtered out. As illustrated in Table **??**, "He" and "It" were considered as synonyms. This significantly decreased the whole size of the datasets. Only 25% of the sentences were left after filtration, as shown in Table 3.15.

3.5 Language Model Filtering

In this section, a causal language model was employed to further resolve the non-contextual ambiguity by assessing the contextual relevance of ambiguous words. We hoped language models can find out those small relationships between sentences. For tense difference, ambiguous words indicated the words that showed the tense. For subject difference, ambiguous words meant the subject which produced differences. The main idea involved calculating the difference between the conditional probabilities of the ambiguous words in the presence versus absence of context as well as with the wrong context. We explained this more precisely with the example in Table 3.16, which showed a tense ambiguity due to context and we calculated 3 groups of probability for it:

Context	Ambiguous sentences
I was at home.	I was doing my homework.
I am at home.	I am doing my homework.

Table 3.16: Example with tense difference to explain language model process

Possibility of ambiguous words with context:

$$P1 = P(\text{'was'}|\text{'I was at home. I'})$$
$$P2 = P(\text{'am'}|\text{'I am at home. I'})$$

Possibility of ambiguous words without context:

$$P3 = P(\text{'was'}|\text{'I'})$$
$$P4 = P(\text{'am'}|\text{'I'})$$

Possibility of ambiguous words given the context from others:

$$P5 = P(\text{'was'}|\text{'I am at home. I'})$$
$$P6 = P(\text{'am'}|\text{'I was at home. I'})$$

Test if the following conditions are fulfilled:

Condition1: P1 - P3 and P2 - P4 is positive

Condition2: P1 - P5 and P2 - P6 is positive

Condition3: P1 + P2 - P3 - P4 - P5 - P6 is positive

The magnitude of difference illustrated the correlation degree. The higher the score, the greater the impact.

In practice, we chose DistilGPT2 as the applied language model. Before employing it, multiple preprocessing should be done. Firstly, we checked if there was only one difference between the two sentences in one pair since we only focused on the probability of one ambiguous word here. Secondly, ambiguous words were split into multiple subwords when tokenizing sometimes. In this case, we averaged the probabilities of all the subwords as the final probability of the word. The remaining number of sentences after filtration was in Table 3.17. Since the size of Literary in both situations was small, we made our research below only with OPUS.

	OPUS (tense)	Literary (tense)	OPUS (subject)	Literary (subject)
Number of sentences after filtering sentences with multiple differences	1,078	26	941	43

Table 3.17: Number of sentences after preprocessing

We first got the number of sentences matching each condition above, as shown in Table 3.18. Then we set thresholds for each condition result and conducted a random evaluation.

Condition	Number of sentences with tense difference	Number of sentences with subject difference
1	186	149
2	188	150
3	26	59

Table 3.18: Number of sentences satisfying various conditions

The tense evaluation result for tense condition1, condition2 and condition3 were illustrated in Table 3.23, Table 3.24 and Table 3.25. In every three scenarios, we achieved a relatively high rate compared with the before. We chose a testset satisfying condition 2 as out final result since it not only achieved a good rate but it also had a relatively larger size. Although the testset satisfying condition 3 had the highest rate, its small size excluded itself.

For subject difference, the results satisfying each condition were shown in Table 3.26, Table 3.27, and Table 3.28. Obviously, the testset satisfying condition 1 performed the best with a rate of around 40%, which was twice as high as the others. So we chose it as out final testset. However, despite its increase, the rate still hasn't exceeded 50%. We took three speculations on the reasons for the low rate. First, most non-contextual reasons mentioned above were filtered out. However, there were still lots of synonyms as well as the use of both singular and plural forms. Often there was no clear contextual information for further distinguishing. The second reason was that we only considered one context before calculating the scores since it sometimes needs context farther away to determine the subject. However, it was problematic if we gave more sentences before, because the language model then needed to consider more factors, making the final result contain more uncertainty. The third reason and most likely reason was that the dataset itself contained relatively few instances of contextual subject ambiguity because all of our random evaluations illustrated a low rate in the end.

For both testsets, although we improved their overall contextually ambiguous rates step by step, their sizes were also decreasing. Obviously, some contextually ambiguous sentences were also filtered out by our methods. Therefore, more research should be done here to find a solution to enlarge the sizes.

3.6 EN->ZH

The entire content above illustrated how to extract instances of ambiguity when translating from Chinese into English. In this section, we shifted our focus to the opposite direction, namely from English into Chinese. We provided an initial start and proposed multiple potential reasons leading to ambiguity.

As illustrated in Figure 3.19, there were in total 534,659 English sentences that had ambiguity problems, namely 516,450 for OPUS and 18,209 for Literary. The distribution of this was in Figure 3.2

Filename	Number of Chinese sentences
OPUS	516,450
Literary	18,209

Table 3.19: Number of English sentences that have multiple various translations

below. We got 395,061 and 11586 sentences with two varieties of English translation from OPUS and Literary. The same, most English sentences had two varieties of Chinese translations and.

The initial random evaluation rate could be seen in Table 3.20. For both datasets the result rate was 0%. The reason was not only because there still persisted synonym problems but also due to the reversal of the sentence structure.

To identify the possible causes of ambiguity, we first considered the possibilities mentioned above, namely tense difference and subject difference, and assessed if they are possible here. Since Chinese verbs did not have tense changes, tense differences should be excluded. However, subject differences still existed. As illustrated in Table 3.21, ”you” refers to an individual or a group of people, and the appropriate translation depends on the context.

There was another special case for Chinese - the Formality. As illustrated in Table 3.22, ”you” here was translated into ” 你” in variety 1 and ” 您” in variety 2. The use of ” 您” in variety 2 indicated respect for the person because the antecedent of it was an old man.

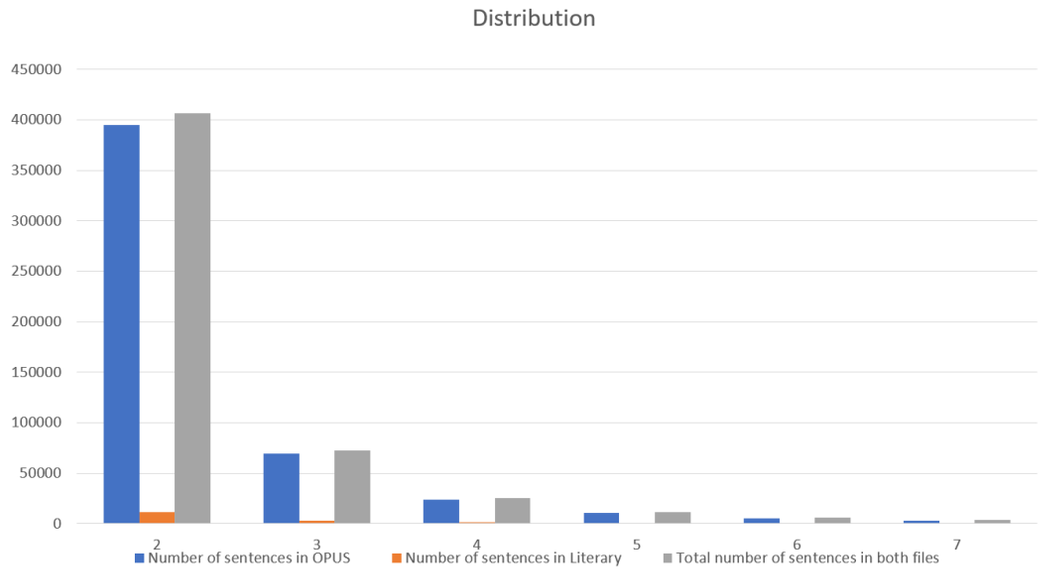

Figure 3.2: Distribution of different varieties of translations

FileName	Need Context	do not need context	lnitial Rate
OPUS	0	10	0%
Literary	0	10	0%

Table 3.20: Initial rate of contextual ambiguous sentences

source:	”What are **you** doing?”
variety1:	“**你们**干什么!”
variety2:	“**你**干什么!”
context1:	马森一挥手，身后**家族佣兵团**的高手顿时冲了进来，就只听见一阵刀剑出鞘的声音，七八把武器就把林云围了起来。
context2:	一边说，**他**一边站起来，关上门，“咔哒”一声，上了锁

Table 3.21: Subject difference example

source:	Who are you ?
variety1:	**你**是谁?
variety2:	**您**是谁?
context1:	我望着这位**老者**说:”
context2:	我打量着面前的**小混混**说:”

Table 3.22: Formality example

Number of sentences	Difference by condition1	Contextual	Non-contextual	Probability
106	e^{-5}	7	3	70%
180	e^{-7}	5	5	50%

Table 3.23: Number of sentences and contextual ambiguous sentences rate for condition1 in tense difference for different thresholds

Number of sentences	Difference by condition2	Contextual	Non-contextual	Probability
116	e^{-6}	7	3	70%
165	e^{-7}	7	5	70%

Table 3.24: Number of sentences and contextual ambiguous sentences rate for condition2 in tense difference for different thresholds

Number of sentences	Difference by condition3	Contextual	Non-contextual	Probability
26	0	8	2	80%

Table 3.25: Number of sentences and contextual ambiguous sentences rate for condition3 in tense difference for different thresholds

Number of sentences	Difference by condition1	Contextual	Non-contextual	Probability
96	e^{-4}	4	6	40%
146	e^{-6}	4	6	40%

Table 3.26: Number of sentences and contextual ambiguous sentences rate for condition1 in subject difference for different thresholds

Number of sentences	Difference by condition2	Contextual	Non-contextual	Probability
96	e^{-4}	2	8	20%
144	e^{-6}	2	8	20%

Table 3.27: Number of sentences and contextual ambiguous sentences rate for condition2 in subject difference for different thresholds

Number of sentences	Difference by condition3	Contextual	Non-contextual	Probability
59	0	2	8	20%

Table 3.28: Number of sentences and contextual ambiguous sentences rate for condition3 in subject difference for different thresholds

4 Conclusion & Future work

In this thesis, we primarily focused on exploring how to extract contextual ambiguous sentences from a large parallel dataset to build a document-level testset automatically to evaluate the performance of context-aware models accurately and easily. We mainly emphasized the direction of **ZH->EN** as well as provided an initial start for the opposite direction.

OPUS	Step Numbering	Number of sentences for tense difference	Number of sentences for subject difference	Rate for tense difference	Rate for subject difference
	3	1534	1255	50%	10%
	4	1386	962	60%	20%
	5	165	146	70%	40%

Table 4.1: This table shows the probability of getting a contextually relevant ambiguous sentence in the testset after each step. Step numbering is the number set at the beginning of chapter 3. Rate is the probability of getting a contextually relevant ambiguous sentence out of ten arbitrarily selected sentences

Literary	Step numbering	Number of sentences for tense difference	Number of sentences for subject difference	Rate for tense difference	Rate for subject difference
	3	32	52	20%	30%
	4	32	43	20%	30%

Table 4.2: This table shows the probability of getting a contextually relevant ambiguous sentence in the testset after each step. Step numbering is at the beginning of Chapter 3. Rate is the probability of getting a contextually relevant ambiguous sentence out of ten arbitrarily selected sentences

In the ZH->EN direction, we found two reasons leading to ambiguity, namely tense difference and subject difference, by analyzing semantic rules of both source and target languages as well as observing

those most similar ambiguous sentences. Then we employed techniques such as Sentence similarity calculation, Language model filtration as well as Ambiguity reasons analysis to extract contextual ambiguous sentences while eliminating non-contextual ones. We conducted a random evaluation after each step to choose the best result for the further research study. The rates after each step were illustrated in Table 4.1 for OPUS and Table 4.2 for Literary. As the final result, we built up a testset for the tense difference comprising 165 sentences with an estimated rate of 70% for OPUS and 32 sentences with an estimate rate of 20% for Literary. Regarding the subject difference, the result was not as high as expected, we had only 43 sentences with a low rate of 30% for Literary and 146 sentences with a rate of 40% for OPUS. However, both are a great increase compared with the initial rate of contextual ambiguity at 0%.

In the EN->ZH direction, we provided some basic evaluations about the dataset and proposed the total number and the distribution of ambiguous sentences as well as multiple reasons leading to ambiguity, namely subject difference and especially the Formality problem.

In the following part, we introduce multiple potential future research directions to further enhance and complete our study:

- In the Language model filtration section mentioned above, we only evaluated the instances by considering one sentence before. However, in practice, some antecedents appear in more distant contexts. How to evaluate with a variable size of context for evaluating is a good research direction.
- We only presented an initial start for the testset in the direction of EN-> ZH. Further research is required to assess the multiple detailed characteristics, including the quantity of those mentioned ambiguity phenomena, as well as to extract and filter a final testset.
- We have already got the testset, but have not yet evaluated the performance of the recently mentioned context-aware models on it. Future research can involve using this testset to compare the performance differences among various models and validate the importance of a certain architecture, like a multi-encoder.
- We have focused mainly on Chinese-English parallel corpora. In the future, we can expand this to multilingual parallel corpora, such as Chinese-Germany or Chinese-France and analyze commonalities and differences leading to ambiguity, and compare the model's performance.

Bibliography

[1] *Artificial neural network*. https://en.wikipedia.org/wiki/Artificial_neural_network.

[2] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. *Neural Machine Translation by Jointly Learning to Align and Translate*. 2016. arXiv: 1409.0473 [cs.CL].

[3] Rachel Bawden et al. "Evaluating Discourse Phenomena in Neural Machine Translation". In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. New Orleans, Louisiana: Association for Computational Linguistics, June 2018, pp. 1304–1313. DOI: 10.18653/v1/N18-1118. URL: https://aclanthology.org/N18-1118.

[4] G. Bebis and M. Georgiopoulos. "Feed-forward neural networks". In: *IEEE Potentials* 13.4 (1994), pp. 27–31. DOI: 10.1109/45.329294.

[5] *BLEU*. https://en.wikipedia.org/wiki/BLEU.

[6] Peter F. Brown et al. "A Statistical Approach to Machine Translation". In: *Comput. Linguist.* 16.2 (June 1990), pp. 79–85. ISSN: 0891-2017.

[7] *Causal language modeling*. https://huggingface.co/docs/transformers/tasks/language_modeling.

[8] Kyunghyun Cho et al. *Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation*. 2014. arXiv: 1406.1078 [cs.CL].

[9] Stefania Cristina. *The Transformer Model*. https://machinelearningmastery.com/the-transformer-model/.

[10] PMP Deepthi Viswanathan Nair. *Unlocking the potential of Cosine similarity in AI*. https://www.linkedin.com/pulse/unlocking-potential-cosine-similarity-ai-viswanathan-nair-pmp/.

[11] Pradeep Dhote. *Seq2Seq-Encoder-Decoder-LSTM-Model*. https://pradeep-dhote9.medium.com/seq2seq-encoder-decoder-lstm-model-1a1c9a43bbac.

[12] Pradeep Dhote. *Seq2Seq-Encoder-Decoder-LSTM-Model*. https://pradeep-dhote9.medium.com/seq2seq-encoder-decoder-lstm-model-1a1c9a43bbac.

[13] *DistilGPT2*. https://huggingface.co/distilgpt2.

[14] *Encoder-Decoder (seq2seq)*. http://en.diveintodeeplearning.org.s3-website-us-west-2.amazonaws.com/chapter_natural-language-processing/seq2seq.html.

[15] *GPT-2*. https://huggingface.co/gpt2.

[16] Jiawei Han, Micheline Kamber, and Jian Pei. "2 - Getting to Know Your Data". In: *Data Mining (Third Edition)*. Ed. by Jiawei Han, Micheline Kamber, and Jian Pei. Third Edition. The Morgan Kaufmann Series in Data Management Systems. Boston: Morgan Kaufmann, 2012, pp. 39–82. ISBN: 978-0-12-381479-1. DOI: https://doi.org/10.1016/B978-0-12-381479-1.00002-2. URL: https://www.sciencedirect.com/science/article/pii/B9780123814791000022.

[17] *Introduction to Recurrent Neural Network*. https://www.geeksforgeeks.org/introduction-to-recurrent-neural-network/.

[18] Philipp Koehn and Rebecca Knowles. "Six Challenges for Neural Machine Translation". In: *CoRR* abs/1706.03872 (2017). arXiv: 1706.03872. URL: http://arxiv.org/abs/1706.03872.

[19] en Lutkevich. *language modeling*. https://www.techtarget.com/searchenterpriseai/definition/language-modeling.

[20] *Machine Translation*. https://en.wikipedia.org/wiki/History_of_machine_translation.

[21] *Machine Translation*. https://www.atanet.org/client-assistance/machine-translation/.

[22] Sourabh Mehta. *Decoding the Mechanics of Masked and Casual Language Models*. https://machinehack.com/story/decoding-the-mechanics-of-masked-and-casual-language-models.

[23] Mathias Müller et al. "A Large-Scale Test Set for the Evaluation of Context-Aware Pronoun Translation in Neural Machine Translation". In: *Proceedings of the Third Conference on Machine Translation: Research Papers*. Brussels, Belgium: Association for Computational Linguistics, Oct. 2018, pp. 61–72. DOI: 10.18653/v1/W18-6307. URL: https://aclanthology.org/W18-6307.

[24] Gergely D. Németh. *Machine Translation: A Short Overview*. https://towardsdatascience.com/machine-translation-a-short-overview-91343ff39c9f.

[25] Jair Neto. *Best NLP Algorithms to get Document Similarity*. https://medium.com/analytics-vidhya/best-nlp-algorithms-to-get-document-similarity-a5559244b23b.

[26] *nlp-recipes*. https://microsoft.github.io/nlp-recipes/examples/sentence_similarity/.

[27] Kishore Papineni et al. "Bleu: a Method for Automatic Evaluation of Machine Translation". In: *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. Philadelphia, Pennsylvania, USA: Association for Computational Linguistics, July 2002, pp. 311–318. DOI: 10.3115/1073083.1073135. URL: https://aclanthology.org/P02-1040.

[28] Engin Pekel and Selin Kara. "A COMPREHENSIVE REVIEW FOR ARTIFICAL NEURAL NETWORK APPLICATION TO PUBLIC TRANSPORTATION". In: *Sigma Journal of Engineering and Natural Sciences* 35 (Mar. 2017), pp. 157–179.

[29] Matt Post and Marcin Junczys-Dowmunt. *Escaping the sentence-level paradigm in machine translation*. 2023. arXiv: 2304.12959 [cs.CL].

[30] Ricardo Rei et al. *COMET: A Neural Framework for MT Evaluation*. 2020. arXiv: 2009.09025 [cs.CL].

[31] Robin M. Schmidt. *Recurrent Neural Networks (RNNs): A gentle Introduction and Overview*. 2019. arXiv: 1912.05911 [cs.LG].

[32] Rifayat Showrav. *Unleashing the Power of Neural Networks: Exploring the Depths of Artificial Intelligence?* https://medium.com/mlearning-ai/unleashing-the-power-of-neural-networks-exploring-the-depths-of-artificial-intelligence-998d4ce29d55.

[33] Felix Stahlberg. *Neural Machine Translation: A Review*. https://jair.org/index.php/jair/article/view/12007.

[34] Dario Stojanovski. "Modeling contextual information in neural machine translation". June 2021. URL: http://nbn-resolving.de/urn:nbn:de:bvb:19-284113.

[35] Elior Sulem, Omri Abend, and Ari Rappoport. "BLEU is Not Suitable for the Evaluation of Text Simplification". In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, Oct. 2018, pp. 738–744. DOI: 10.18653/v1/D18-1081. URL: https://aclanthology.org/D18-1081.

[36] Jörg Tiedemann. "Parallel Data, Tools and Interfaces in OPUS". In: *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*. Ed. by Nicoletta Calzolari (Conference Chair) et al. Istanbul, Turkey: European Language Resources Association (ELRA), May 2012. ISBN: 978-2-9517408-7-7.

[37] *Vanishing gradient problem*. https://en.wikipedia.org/wiki/Vanishing_gradient_problem.

[38] Ashish Vaswani et al. *Attention Is All You Need*. 2023. arXiv: 1706.03762 [cs.CL].

[39] Mauricio Veronez et al. "Regional Mapping of the Geoid Using GNSS (GPS) Measurements and an Artificial Neural Network". In: *Remote Sensing* 3 (Dec. 2011). DOI: 10.3390/rs3040668.

[40] Lena Voita. *Sequence to Sequence (seq2seq) and Attention*. https://lena-voita.github.io/nlp_course/seq2seq_and_attention.html.

[41] Tomas Vykruta. *Understanding Causal LLM's, Masked LLM's, and Seq2Seq: A Guide to Language Model Training Approaches*. https://medium.com/@tom_21755/understanding-causal-llms-masked-llm-s-and-seq2seq-a-guide-to-language-model-training-d4457bbd07fa.

[42] *What are the advantages and disadvantages of using long short-term memory (LSTM) cells over simple RNN cells?* https://www.linkedin.com/advice/1/what-advantages-disadvantages-using-long-short-term.

[43] *WMT23 Discourse-Level Literary Transla*. http://www2.statmt.org/wmt23/literary-translation-task.html.

[44] Yonghui Wu et al. *Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation*. 2016. arXiv: 1609.08144 [cs.CL].