Robust Guidance for Unsupervised Data Selection: Capturing Perplexing Named Entities for Domain-Specific Machine Translation

Seunghyun Ji[1,2], **Hagai Raja Sinulingga**[2], **Darongsae Kwon**[2]
[1]Ahancompany corporation, [2]TelePIX
seunghyun.ji@a-ha.io, {hagairaja, darong.kwon}@telepix.net

Abstract

Low-resourced data presents a significant challenge for neural machine translation. In most cases, the low-resourced environment is caused by high costs due to the need for domain experts or the lack of language experts. Therefore, identifying the most training-efficient data within an unsupervised setting emerges as a practical strategy. Recent research suggests that such effective data can be identified by selecting 'appropriately complex data' based on its volume, providing strong intuition for unsupervised data selection. However, we have discovered that establishing criteria for unsupervised data selection remains a challenge, as the 'appropriate level of difficulty' may vary depending on the data domain. We introduce a novel unsupervised data selection method named 'Capturing Perplexing Named Entities,' which leverages the maximum inference entropy in translated named entities as a metric for selection. When tested with the 'Korean-English Parallel Corpus of Specialized Domains,' our method served as robust guidance for identifying training-efficient data across different domains, in contrast to existing methods.

Keywords: Machine Translation, Data Selection, Unsupervised Method

1. Introduction

With the advent of large-scale models capable of translating numerous languages in various directions(Aharoni et al., 2019), the field of machine translation is entering a new era. For instance, 'No Language Left Behind(NLLB Team et al., 2022)', which demonstrated outstanding performance across a range of languages, was trained on over 40,000 combinations of 200 languages. These models can be regarded as pre-trained or foundational, as they have acquired general knowledge for translation. Nevertheless, they might sometimes face challenges when translating domain-specific data, despite their extensive training on diverse datasets. To address this, fine-tuning the pre-trained models with target domain data can enhance their specialization(Fadaee and Monz, 2018; Zan et al., 2022).

However, when addressing narrow or specialized domains, the model must recognize words that are relatively rare in general corpora. This presents a challenge, as rare words often consist of sparse tokens, such as those composed of single character tokens. Named entities, such as names of persons, organizations, etc., frequently lack synonyms, making it even more perplexing to build contextualized representations, especially in narrow domains. This also underscores the point that acquiring domain-specific translation data is costly, as translators are required who possess not only domain expertise but also familiarity with domain-specific terminology.

To reduce data acquisition costs, one might consider strategically identifying data for labeling rather than making random selections. Several researchers(Paul et al., 2021; Feldman and Zhang, 2020; Sorscher et al., 2022) have suggested various measurement methods aimed at selecting 'effective' data for training. Some of those focus on 'Data difficulty,'(Paul et al., 2021; Meding et al., 2022) identifying data that poses a challenge to a given model. 'Data forgettability(Toneva et al., 2019)' or 'Memorization(Feldman and Zhang, 2020)' could serve as alternative criterion. However, these methods require a supervised setting for selection, which may be inefficient for machine translation. For instance, pruning a dataset is unlikely to yield a better model if the dataset was curated by domain experts (Maillard et al., 2023).

In an unsupervised setting, where training-efficiency should be guessed without a label, Sorscher et al. (2022) demonstrated that the Euclidean distance between a data point's representation and its cluster centroid can serve as an effective criterion for data selection. This approach is supported by several concrete theoretical analyses and provides straightforward guidance for data selection. However, it remains uncertain whether this criterion can be universally applied to parameter-efficient fine-tuning methods(Houlsby et al., 2019; Hu et al., 2022; Liu et al., 2022), which are commonly used. We observed that this measurement method might not always align with training-

This work was initially started in TelePIX, the previous affiliation of the first author.

The code is available in the following hyperlink : https://github.com/comchobo/Capturing-Perplexing-Named-Entities

SIGUL2024 Workshop, pages 307–317

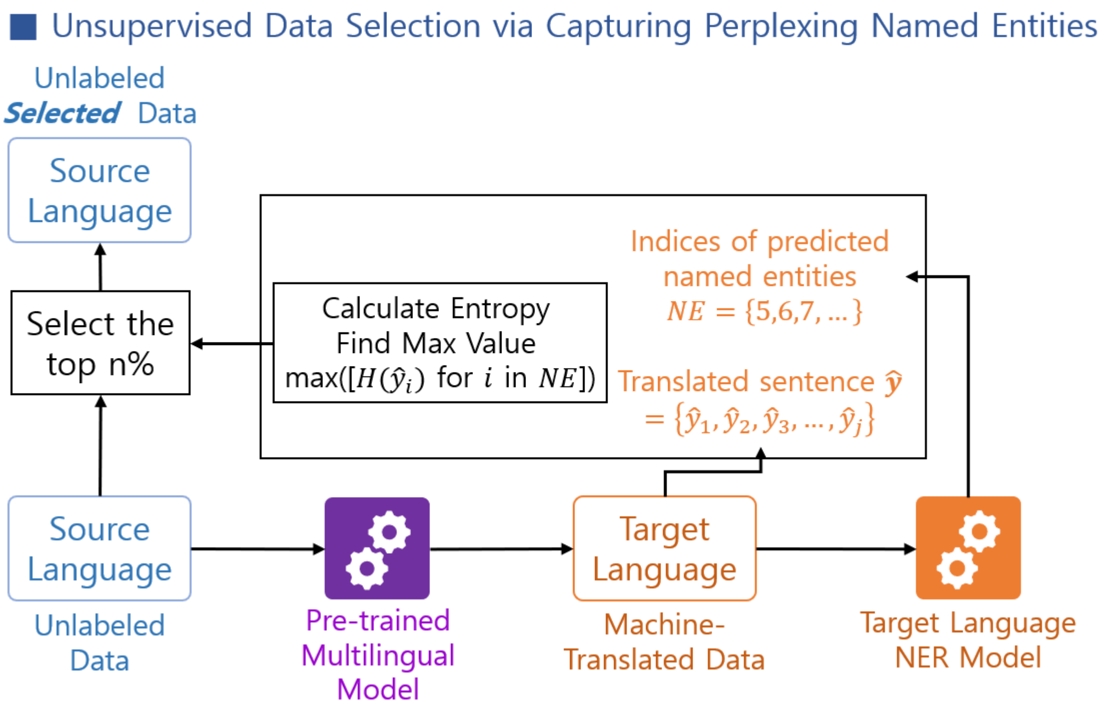

Figure 1: A diagram illustrates our method, which utilizes a pre-trained multilingual model for machine translation and a named entity recognition model that has been fine-tuned on the target language. Our method comprises three steps: 1) capturing named entity tokens in the machine-translated sentences, 2) calculating the inference entropy of those tokens, and 3) using the maximum entropy value as a measure for selection.

efficiency, indicating that it may not consistently correlate with performance improvement, despite using the same pre-trained weights and dataset size. These findings are detailed in Section 5.2.

We propose a novel method for unsupervised data selection, which we refer to as 'Capturing Perplexing Named Entities'. Our method identifies data that should be selected, by assessing the perplexity of named entity tokens translated by a given pre-trained model, as described in Figure 1. The motivations behind this approach are as follow:

- Since named entities in domain-specific data are challenging to translate without recognizing the complex patterns within the domain, they represent one of the most difficult portions to translate. Therefore, these entities should be given priority for efficient domain adaptation.
- The entropy score of a vocabulary distribution can indicate the model's level of perplexity. Given that synonyms for named entities are unlikely to exist, the model should not exhibit a high entropy score for named entities.

In several experiments targeting domain-specific 'Korean to English' translation, our method consistently identified the most training-efficient data. This indicates that our measurement method has a stronger correlation with performance improvement compared to existing methods, which can vary significantly across different data domains. For clarity in our discussion, 'MDS' will serve as the abbreviation for Measurement method for Data Selection, and 'Value by MDS' will denote the specific value it calculates.

2. Related Works

2.1. Named entities in Machine Translation

Translating named entities presents a significant challenge in machine translation(Ugawa et al., 2018), although it is crucial for delivering accurate information(Tjong Kim Sang and De Meulder, 2003). Incorrect translations of named entities, even with few errors, can lead to information distortion. For instance, in Table 1, the human-translated and machine-translated Korean to English-sentences may seem similar. However, a closer examination reveals differences in the individual's name (Steven Strasburg), the league (Major League Baseball), and an adjective (original). Despite these mistakes causing critical distortions, re-

Languages	Data Examples	Scores
Korean	메이저리그 자유계약선수(FA) 최대어 투수 중 한 명인 스티븐 스트라스버그가 원 소속팀 워싱턴과 7년 2억4,500만달러에 도장을 찍었다.	COMET 90.92
English	Steven Strasburg, one of the biggest free agent (FA) pitchers in Major League Baseball, has signed a 7-year, $ 245 million contracts with his original team Washington.	ChrF++ 67.94
Translated	Steven Strasberg, one of the biggest pitchers in the Major League Free Agent (FA) league, signed a seven-year, $ 245 million contract with former team Washington.	BLEU 27.38
Korean	고메스 부상 이후 에버턴 지휘봉을 잡게된 카를로 안첼로티 감독은 지난주 "고메스의 회복이 순조롭게 이뤄지고 있다"고 밝혔다.	COMET 90.99
English	Manager Carlo Ancelotti, who took the helm of Everton after Gomez's injury, revealed last week that "Gomez's recovery is going smoothly."	ChrF++ 64.47
Translated	Coach Carlo Ancelotti, who took over Everton after Gomes' injury, said last week, "Gomes' recovery is progressing smoothly."	BLEU 24.86

Table 1: Example pairs with high COMET and ChrF++ scores but low BLEU scores were selected from sports domain data. The first column represents the source (Korean), the target (English), and the machine-translated (Korean to English) result. Words that may cause critical semantic distortions are highlighted in red. The last column lists the evaluation scores of the machine-translated sentences, calculated using three different metrics.

cent metrics such as COMET(Rei et al., 2020)[1] and ChrF++(Popović, 2015) show scores high enough to be interpreted as satisfactory results. Given that some rare named entities are more common in domain-specific data, building precise contextualized representations of data, which contains named entities, is even difficult to capture by recent deep-model based metrics.

One current approach to translate named entities precisely, integrates a knowledge base(Zhao et al., 2020) or employs a transliteration model once tokens are identified as named entities(Sharma et al., 2023). However, these strategies often rely on specialized algorithms that act as a workaround, rather than directly boosting the translation model's performance or robustness. Multi-task learning has demonstrated improvements in translation performance when additional annotations for named entities are provided(Xie et al., 2022). However, this method may incur significantly higher labeling costs.

2.2. Data Selection for Training

Throughout several training cycles, metrics such as forgetting scores(Toneva et al., 2019), memorization(Feldman and Zhang, 2020), diverse ensembles(Meding et al., 2022), and normed gradients(Paul et al., 2021) could be used as one of the measurement methods for data selection (MDS). EL2N, which quantifies the error magnitude, acts as a training-free MDS. However, these methods require annotations, limiting their application to supervised settings only. As high-quality data has been shown to significantly outperform large volumes of low-quality or synthetic data(Maillard et al., 2023), it is generally recommended that the data with elaborate annotations should not be pruned.

In an unsupervised setting, one might explore data uniqueness—for example, by measuring the Euclidean distance between a data representation and its centroid(Sorscher et al., 2022) (referred to as Selfsup)—as a form of unsupervised MDS. Measuring uncertainty, which could be estimated by the entropy of the probability distribution, also might be one of MDS(Brown et al., 1990; Wu et al., 2021). However, empirical evidence suggests that when training with small datasets, excessively unique data (indicated by high values in MDS Selfsup) may impede training(Sorscher et al., 2022). Therefore, selecting data using the appropriate type of MDS and determining the optimal value for MDS are crucial. Nonetheless, establishing a standard for this is challenging, to the best of our knowledge.

In machine translation, reference-free Quality Estimation (QE) methods, which operate as an un-

[1]We used `https://huggingface.co/Unbabel/wmt22-comet-da` to evaluate using COMET score.

supervised MDS, are gaining focus. One strategy involves the intuition of 'seeking perplexing data' by identifying attention distractions or uncertainties(Peris and Casacuberta, 2018). More sophisticated reference-free QE algorithms, which can be implemented using deep models(Rei et al., 2021), have demonstrated competitive results when compared to their reference-requiring counterparts(Rei et al., 2020). However, these methods, relying on sentence embedding models, are often confounded by even slight literal differences. We have observed and discussed this phenomenon in Section 2.1.

3. Existing Methods

We consider the multilingual translation model as a 'pre-trained model', with subsequent training on specific data referred to as 'fine-tuning'.

3.1. EL2N

Paul et al. (2021) previously used the average error from several minimally trained models to identify data that could not be easily trained in a few epochs. This method requires paired data for its computations, hence categorized as a supervised approach. Intuitively, the EL2N value from a pre-trained model signifies an average error or incorrect confidence, enabling the identification of the most problematic data for a given model. If $\boldsymbol{Y}$ and $\hat{\boldsymbol{Y}}$ represent the original and translated sentences in the target language, respectively, EL2N can be described as follows:

$$\text{EL2N}(\boldsymbol{Y}, \hat{\boldsymbol{Y}}) = \frac{1}{L}\sum_{l=1}^{L} \|\boldsymbol{y}_l - \hat{\boldsymbol{y}}_l\|$$
$$L = min(|\boldsymbol{Y}|, |\hat{\boldsymbol{Y}}|)$$

where $\hat{\boldsymbol{y}}$ represents the predicted token distribution, and $\boldsymbol{y}$ is the actual label. Given that the translated sentence may contain a different number of tokens from original sentence, we chose the shorter token length, represented by the cardinality of $\boldsymbol{Y}$ and $\hat{\boldsymbol{Y}}$.

3.2. Entropy

Brown et al. (1990) demonstrated that uncertainty in prediction is quantifiable by entropy. Various studies have reported performance improvements by employing entropy to select data for training(Jiao et al., 2021; Wu et al., 2021). Building on this concept, we considered entropy as an indicator of the pre-trained model's perplexity regarding specific sentences, selecting them as candidates for fine-tuning. The entropy of the vocabulary distribution is defined as:

$$H(\hat{\boldsymbol{y}}) = \frac{1}{V}\sum_{i \in V} -P(\hat{y}_i)logP(\hat{y}_i)$$

where V is a vocabulary. We adopted averaged entropy as MDS which is as follows:

$$AvgEntropy(\hat{\boldsymbol{Y}}) = \frac{1}{L}\sum_{l=1}^{L} H(\hat{\boldsymbol{y}}_l)$$

where L is a length of the sentence $\hat{\boldsymbol{Y}}$.

However, given that the optimal entropy level may differ by token types, such as adjectives or synonyms, we hypothesized that employing $AvgEntropy$ as an MDS might lead the model to become either overconfident or overly cautious.

3.3. Selfsup

Sorscher et al. (2022) observed that within clustered image representations, data points distant from their centroids often exhibit unique patterns, which have high Euclidean distance to the centroid. However, its effectiveness as an MDS for fine-tuning translation models remains unverified. To adapt this approach to the language domain, we utilized sentence embeddings for the source data and applied k-means clustering. If x_A represents a sentence embedding of source language data x, clustered around centroid A, then the MDS Selfsup can be described as:

$$Selfsup(x_A) = ||x_A - A||$$

If the sentence embeddings are well-aligned, MDS Selfsup is expected to capture training-efficient data for fine-tuning. Although recent sentence embedding models demonstrate decent performance, their accuracy in domain-specific data remains questionable. Our findings provide support for this doubt, as illustrated in Table 1, where the COMET score failed to detect semantic distortion.

3.4. Reference-free COMET

Rei et al. (2021) proposed a Reference-free COMET, which was trained to estimate quality without reference, only with source and translated sentences. Reference-free COMET was designed to predict quality annotations using a sentence embedding model. Its output range is 0 to 1, where 1 denotes the best quality. We expected that Reference-free COMET as an MDS would be inversely proportional to the training-efficiency since it would detect examples that the model could not translate well.

4. Proposed Method

Our hypothesis posits that complex patterns possessed by named entities are essential for fine-tuning. This is particularly true in domain-specific machine translation, where rare words and expressions occur frequently but are not present in the general domain. By incorporating these characteristics into data selection, we measured the maximum entropy while translating named entities, which are unlikely to have alternative answers. In summary, our method specifically targets perplexing named entities.

```
// Dataset X in source language consists
   of sentences x
// f_pre is pre-trained multilingual model
// d is an index of segments.
// len is an amount of data to sample.
def PruneByMDS(X, d, len = 2000):
    X' ← empty dictionary
    for x in X:
        ŷ ← f_pre(x)
        X'['Value by MDS'].insert(f_MDS(ŷ))
        X'['Sentence'].insert(x)
    X'.sortby(['Value by MDS'])
    X' ← X'.split_into(4)
    X' ← X'.select(d)
    X' ← X'.sample(len)
    return X'
```

Figure 2: Pseudo code for the experiment data preparation. We sorted and split the data into 4 segments based on each value by MDS. Then, we sampled 2,000 sentences from each segment for fine-tuning.

$$PerEnts(\hat{\boldsymbol{Y}}) = max(\{H(\hat{\boldsymbol{y}}_x) | x \in NE(\hat{\boldsymbol{y}})\})$$

where $NE(\hat{\boldsymbol{y}})$ represents a set of named entity token indices in the machine-translated sentence $\hat{\boldsymbol{y}}$, predicted by a named entity recognition model. We will use the abbreviation 'PerEnts,' to refer to our method.

5. Experiments

5.1. Settings for experiments

We attempted to evaluate our method, which is one of the unsupervised MDSs, with various datasets. We sorted the data based on the values of each MDS and divided it into four segments to verify that each MDS is proportional to training-efficiency. If it is proportional and invariant across data domains, it can be regarded as 'robust guidance' for unsupervised data selection. We also conducted multiple data samplings for fine-tuning to precisely assess the capabilities of MDSs. This process follows the same cycle as described in the pseudo-code, shown in Figure 2. Note that the highest segment index (3 in our case) represents data subsets with the highest values according to each MDS.

Models and Datasets As a pre-trained translation model, we used 'NLLB-1.3B(NLLB Team et al., 2022)[2]' multilingual model. We then employed the 'Korean-English Parallel Corpus of Specialized Domains(Flitto, 2021) [3]', published by the National Information Society Agency of South Korea, as the domain-specific dataset. Given the scarcity of open datasets in the Korean language available for public download, we adopted this approach despite its limited access being restricted to nationals. There are 'Law, Medical, Travel, Sports' domains, showing each distribution in Table 2. The 'Law' domain consists of precedents from the Supreme Court of South Korea. The 'Sports' domain includes various articles about international sports events. The other domains were compiled from domain-specific articles, thus containing names of locations (in the Travel domain) or names of medicines (in the Medical domain).

Data Domain	Train / Test
Medical	200k / 25k
Travel	160k / 20k
Law	120k / 15k
Sports	160k / 20k

Table 2: The number of sentences of 'Korean-English Parallel Corpus of Specialized Domains' dataset, released with train/test splits.

Training and Hyperparameters Given the potential variability in domain-specific translation, such as extremely unique domains or low-resource environments, we randomly sampled 2,000 sentences from each segment, regarding the pre-defined seeds. We employed IA3 training(Liu et al., 2022) to simulate practical fine-tuning environments. For hyperparameters, we set the epoch to 10, and the batch size to 32, and searched for the best learning rate from three options [1e-2, 2e-2, 3e-2] during each fine-tuning trial. Given that fine-tuning with a low-resource dataset might result in high variance between models, we took the average scores of three fine-tuned models, using sampled data with 3 different seeds.

Implementations of MDSs Since our method requires named entity recognition model in the target language, which is English in our case, we employed the 'd4data/biomedical-ner-all[4]' fine-tuned model to capture entities in the 'medical' domain dataset, such as names of medicines. For datasets in other domains, we used 'RashidNLP/NER-Deberta[5]' model, trained with Few-NERD dataset(Ding et al., 2021), which we conjectured far more comprehensive than CoNLL-2003 dataset(Tjong Kim Sang and De Meulder,

[2] https://huggingface.co/facebook/nllb-200-distilled-1.3B

[3] This research (paper) used datasets from 'The Open AI Dataset Project (AI-Hub, S. Korea)'. All data information can be accessed through 'AI-Hub (www.aihub.or.kr)'.

[4] https://huggingface.co/d4data/biomedical-ner-all

[5] https://huggingface.co/RashidNLP/NER-Deberta

MDSs	Average Performance		
	BLEU	ChrF++	COMET
Not fine-tuned	21.42	45.57	76.39
Random	33.71	56.90	80.71
Supervised method			
EL2N (Paul et al., 2021)	34.01	57.25	80.84
Unsupervised methods			
Entropy (Jiao et al., 2021)	33.64	57.05	80.86
Selfsup (Sorscher et al., 2022)*	33.85	57.11	80.81
Reference-Free COMET (Rei et al., 2021)*	33.88	**57.22**	**80.92**
PerEnts (ours)	**34.09**	57.19	80.82

Table 3: Average test-set performance across 4 domains. We divided the dataset for each domain into four segments after sorting by each MDS and sampled 2,000 sentences three times from each segment. Given our conjecture that invariance across data domains is an important characteristic of an unsupervised MDS, we reported scores fine-tuned with subsets from either the highest (3) or lowest (0), denoted with an asterisk) segment. The highest scores among the unsupervised MDSs are highlighted in bold.

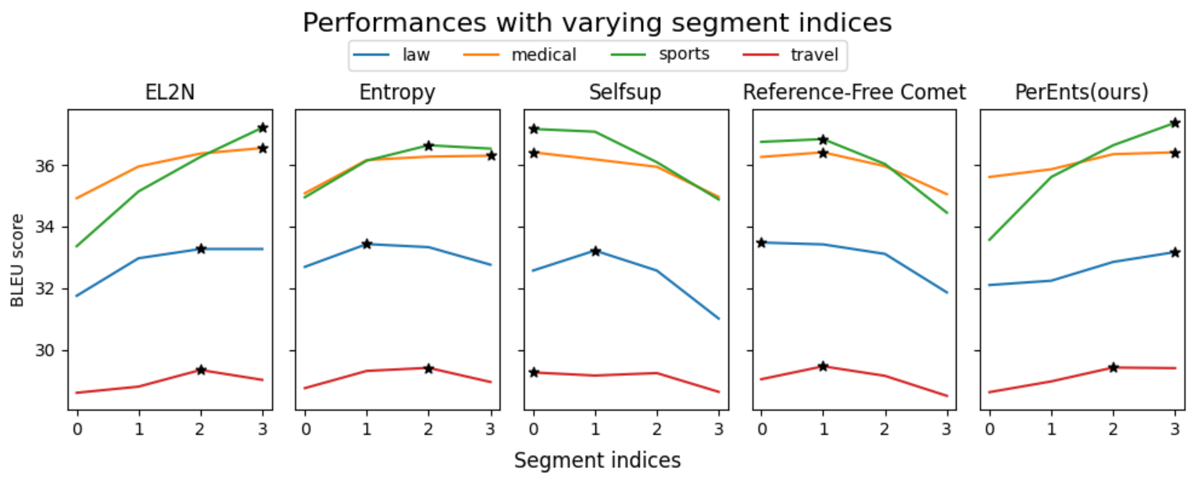

Figure 3: The scores for each segment index across the four domains. The best BLEU scores among the segment indices were marked with a black star. Experimental results demonstrated that our method consistently identified the most training-efficient data by selecting the highest segment (3), whereas other methods varied by data domain.

2003). To implement MDS Selfsup, we used the monolingual sentence embedding model 'BM-K/KoSimCSE-roberta-multitask[6]', which is specialized for the Korean (source) language. Lastly, 'Unbabel/wmt23-cometkiwi-da-xl' was employed for Reference-Free COMET(Rei et al., 2021)[7].

5.2. Main Results

We employed BLEU(Post, 2018), ChrF++(Popović, 2015), and COMET scores(Rei et al., 2020)[8] for evaluation, as presented in Table 3. The fine-tuned models were evaluated using pre-split test sets. It is important to note that, identifying the optimal value for each MDS requires access to every segment index, necessitating a complete parallel corpus for comparison. To simulate a practical strategy where access is limited, we reported averaged scores by selecting either the highest (3) or lowest (0) segment index. For instance, the domain-average

[6] https://huggingface.co/BM-K/KoSimCSE-roberta-multitask

[7] https://huggingface.co/Unbabel/wmt23-cometkiwi-da-xl

[8] We used https://huggingface.co/Unbabel/wmt22-comet-da to evaluate using COMET score.

	The numbers of *Correctly Guessed* / *Newly Guessed* named entities				
	EL2N (Supervised)	Entropy	Selfsup	Reference-Free COMET	PerEnts (Ours)
Law	652/3842	721/3788	589/3837	690/3732	702/3968
Travel	2242/17389	2079/17693	1610/18159	2035/17686	1944/18554
Sports	1822/8087	1841/8785	1875/8648	1900/8736	1922/8442

Table 4: We observed the number of named entities that models could guess for each domain test dataset. Among the words translated by the NLLB model for each test set, named entities (NEs) were stored and classified as a 'Pre-trained Named Entities'. Additionally, NEs observed in the learning datasets created by each method were stored and classified as an 'Observed Named Entities'. If an NE inferred from a model's test data is not present in either the Pre-trained or Observed, it is categorized as 'Newly Guessed'. Furthermore, if such a guess is accurate, it is classified as 'Correctly Guessed'.

score for EL2N was determined by selecting segment index 3, while for MDS Selfsup, segment index 0 was chosen.

Our method, referred to by the abbreviation 'PerEnts,' achieved the highest BLEU score among the MDSs, even surpassing the supervised method (EL2N). Although other existing methods outperformed ours for COMET and ChrF++ scores, we propose that the BLEU score might be the most critical metric for domain-specific translation due to its ability to capture semantic distortion, as demonstrated in Table 1.

Additionally, to assess the robustness of the MDSs, we calculated the average scores across four different domains, as presented in Figure 3. The best performing segment index, selected by other MDSs, was neither 0 nor 3, suggesting that these MDSs are sensitive to the data domain. We conjectured that this observation could complement the assertion by Sorscher et al. (2022) that 'The best selection strategy depends on the amount of initial data.' Even though the same pre-trained weights and the same volume of data were used for each fine-tuning procedure, the data domain could play an important role as a factor. Furthermore, our selection of a well-regarded monolingual sentence embedding model[9] for implementing MDS Selfsup did not result in decent performance, supporting the idea that the sentence embedding model could be confounded by slight literal differences.

5.3. Experiments for Generalizability

Fine-tuning on overly complex or specialized domains can lead to overfitting, which undermines generalization. Particularly, our method, which identifies data with complex named entities, may be prone to overfitting. To verify this, we evaluated the generalizability of each model trained with data generated by MDSs. Initially, for each test set, words translated by the NLLB model were stored and classified as a 'Pre-trained Named Entities'. Similarly, named entities identified in the training datasets selected by each MDSs were cataloged as an 'Observed Named Entities'. While translating test data, a new named entity predicted by a model, which is not in Pre-trained or Observed Named Entities, it is considered 'Newly Guessed'. If such a guess is accurate, it is deemed 'Correctly Guessed'. The counts of Newly Guessed and Correctly Guessed named entities are presented in Table 4.

We could observed that our method do not just memorize named entities in a given train dataset. Although obvious correlations between 'Correctly Guessed Named Entities' were not exposed, our method can help a model to guess correct named entities, without an abuse generating named entities.

MDSs	Averaged Performance		
	BLEU	ChrF++	COMET
PerEnts	**34.09**	57.16	80.82
*Mean	33.94	**57.19**	80.82
Selfsup	**33.85**	**57.11**	**80.81**
*Multilingual	33.3	56.78	80.02

Table 5: The results of MDS variants. '*Mean' denotes that it averaged entropy instead of choosing max in our method(PerEnts), and 'Multilingual' adopted a multilingual sentence embedding model for 'Selfsup'. Both variants used the same segment index to achieve the highest average performance.

5.4. Additional Study

Since the intuition for each MDS could be implemented in various forms, we implemented some MDS variants. e.g., adopting average entropy instead of max for our method. We also employed multilingual sentence embedding model 'sentence-

[9] https://huggingface.co/BM-K/KoSimCSE-roberta-multitask

transformers/LaBSE(Feng et al., 2022)[10]' for implementing MDS Selfsup. The results are reported in Table 5. Although there were less significant degradations, it can be argued that our method's focus on finding maximum entropy more effectively captures the 'unlearned parts.' and it reveals a limitation in the representation ability of multilingual sentence embedding models.

6. Limitations

We attempted to verify our method under various situations and data domains. However, it's important to note that our experiments were conducted with a single translation direction and a single data size (2k). We acknowledge that testing on multiple translation directions and diverse amounts of datasets could potentially provide a more comprehensive validation of MDSs, including our method. Additionally, the impact of utilizing named entities may vary by language, e.g., languages that use uppercase letters. Although we recognize the importance of diverse environments and theoretical analysis, limited experiments were done based on a strategic decision to verify generalizability for practical usage. We believe that these limitations could be interesting topics for future research, exploring which measurement method can generally affect the performances of fine-tuned models.

7. Conclusion

To identify the most training-efficient data for annotating in domain-specific machine translation, we explored various measurement methods that could serve as a benchmark for selection, collectively referred to as 'MDS.' We recognized named entities as 'complex patterns' requiring highly confident prediction. As a result, we introduced 'Capturing Perplexing Named Entity' as one of the MDSs. This approach has seen effective as a guidance for selecting training data, even in unsupervised settings. Despite the common challenge of identifying effective data for annotation in deep learning—a challenge that we could not directly address in terms of the relationship between memorizable patterns and generalizability due to a lack of theoretical analysis—we hope our findings will pave the way for more in-depth research in the future.

8. Bibliographical References

[10] https://huggingface.co/sentence-transformers/LaBSE

Sweta Agrawal, Chunting Zhou, Mike Lewis, Luke Zettlemoyer, and Marjan Ghazvininejad. 2022. In-context examples selection for machine translation.

Roee Aharoni, Melvin Johnson, and Orhan Firat. 2019. Massively multilingual neural machine translation. 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 3874–3884, Minneapolis, Minnesota. Association for Computational Linguistics.

Orevaoghene Ahia, Julia Kreutzer, and Sara Hooker. 2021. The low-resource double bind: An empirical study of pruning for low-resource machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3316–3333, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yamini Bansal, Behrooz Ghorbani, Ankush Garg, Biao Zhang, Colin Cherry, Behnam Neyshabur, and Orhan Firat. 2022. Data scaling laws in nmt: The effect of noise and architecture. In *International Conference on Machine Learning*, pages 1466–1482. PMLR.

Peter F Brown, John Cocke, Stephen A Della Pietra, Vincent J Della Pietra, Frederick Jelinek, John Lafferty, Robert L Mercer, and Paul S Roossin. 1990. A statistical approach to machine translation. *Computational linguistics*, 16(2):79–85.

Ning Ding, Guangwei Xu, Yulin Chen, Xiaobin Wang, Xu Han, Pengjun Xie, Haitao Zheng, and Zhiyuan Liu. 2021. Few-NERD: A few-shot named entity recognition dataset. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3198–3213, Online. Association for Computational Linguistics.

Marzieh Fadaee and Christof Monz. 2018. Back-translation sampling by targeting difficult words in neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 436–446, Brussels, Belgium. Association for Computational Linguistics.

Vitaly Feldman and Chiyuan Zhang. 2020. What neural networks memorize and why: Discovering the long tail via influence estimation. *Advances in Neural Information Processing Systems*, 33:2881–2891.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891, Dublin, Ireland. Association for Computational Linguistics.

Adi Haviv, Ido Cohen, Jacob Gidron, Roei Schuster, Yoav Goldberg, and Mor Geva. 2022. Understanding transformer memorization recall through idioms. *arXiv preprint arXiv:2210.03588*.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for nlp. In *International Conference on Machine Learning*, pages 2790–2799. PMLR.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Wenxiang Jiao, Xing Wang, Zhaopeng Tu, Shuming Shi, Michael R Lyu, and Irwin King. 2021. Self-training sampling with monolingual data uncertainty for neural machine translation. *arXiv preprint arXiv:2106.00941*.

Martin Joos. 1936. *Language*, 12(3):196–210.

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

Haokun Liu, Derek Tam, Mohammed Muqeeth, Jay Mohta, Tenghao Huang, Mohit Bansal, and Colin A Raffel. 2022. Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning. *Advances in Neural Information Processing Systems*, 35:1950–1965.

Jean Maillard, Cynthia Gao, Elahe Kalbassi, Kaushik Ram Sadagopan, Vedanuj Goswami, Philipp Koehn, Angela Fan, and Francisco Guzman. 2023. Small data, big impact: Leveraging minimal data for effective machine translation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2740–2756, Toronto, Canada. Association for Computational Linguistics.

Kristof Meding, Luca M. Schulze Buschoff, Robert Geirhos, and Felix A. Wichmann. 2022. Trivial or impossible — dichotomous data difficulty masks model differences (on imagenet and beyond). In *International Conference on Learning Representations*.

Pedro Mota, Vera Cabarrao, and Eduardo Farah. 2022. Fast-paced improvements to named entity handling for neural machine translation. In *Proceedings of the 23rd Annual Conference of the European Association for Machine Translation*, pages 141–149, Ghent, Belgium. European Association for Machine Translation.

Xuan-Phi Nguyen, Shafiq Joty, Kui Wu, and Ai Ti Aw. 2020. Data diversification: A simple strategy for neural machine translation. *Advances in Neural Information Processing Systems*, 33:10018–10029.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia-Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation.

Mansheej Paul, Surya Ganguli, and Gintare Karolina Dziugaite. 2021. Deep learning on a data diet: Finding important examples early in training. In *Advances in Neural Information Processing Systems*.

Álvaro Peris and Francisco Casacuberta. 2018. Active learning for interactive neural machine translation of data streams. In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, pages 151–160, Brussels, Belgium. Association for Computational Linguistics.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting bleu scores. *WMT 2018*, page 186.

Nasim Rahaman, Aristide Baratin, Devansh Arpit, Felix Draxler, Min Lin, Fred Hamprecht, Yoshua

Bengio, and Aaron Courville. 2019. On the spectral bias of neural networks. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 5301–5310. PMLR.

Vikas Raunak, Arul Menezes, and Marcin Junczys-Dowmunt. 2021. The curious case of hallucinations in neural machine translation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1172–1183, Online. Association for Computational Linguistics.

Ricardo Rei, Ana C Farinha, Chrysoula Zerva, Daan van Stigt, Craig Stewart, Pedro Ramos, Taisiya Glushkova, André F. T. Martins, and Alon Lavie. 2021. Are references really needed? unbabel-IST 2021 submission for the metrics shared task. In *Proceedings of the Sixth Conference on Machine Translation*, pages 1030–1040, Online. Association for Computational Linguistics.

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

Radhika Sharma, Pragya Katyayan, and Nisheeth Joshi. 2023. Improving the quality of neural machine translation through proper translation of name entities. In *2023 6th International Conference on Information Systems and Computer Networks (ISCON)*, pages 1–4.

Ben Sorscher, Robert Geirhos, Shashank Shekhar, Surya Ganguli, and Ari S. Morcos. 2022. Beyond neural scaling laws: beating power law scaling via data pruning. In *Advances in Neural Information Processing Systems*.

Xabier Soto, Dimitar Shterionov, Alberto Poncelas, and Andy Way. 2020. Selecting backtranslated data from multiple sources for improved neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3898–3908, Online. Association for Computational Linguistics.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 142–147.

Mariya Toneva, Alessandro Sordoni, Remi Tachet des Combes, Adam Trischler, Yoshua Bengio, and Geoffrey J. Gordon. 2019. An empirical study of example forgetting during deep neural network learning. In *International Conference on Learning Representations*.

Arata Ugawa, Akihiro Tamura, Takashi Ninomiya, Hiroya Takamura, and Manabu Okumura. 2018. Neural machine translation incorporating named entity. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3240–3250, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Minghao Wu, Yitong Li, Meng Zhang, Liangyou Li, Gholamreza Haffari, and Qun Liu. 2021. Uncertainty-aware balancing for multilingual and multi-domain neural machine translation training. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7291–7305.

Shufang Xie, Yingce Xia, Lijun Wu, Yiqing Huang, Yang Fan, and Tao Qin. 2022. End-to-end entity-aware neural machine translation. *Machine Learning*, pages 1–23.

Changtong Zan, Keqin Peng, Liang Ding, Baopu Qiu, Boan Liu, Shwai He, Qingyu Lu, Zheng Zhang, Chuang Liu, Weifeng Liu, et al. 2022. Vega-mt: The jd explore academy translation system for wmt22.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. 2017. Understanding deep learning requires rethinking generalization. In *International Conference on Learning Representations*.

Yang Zhao, Lu Xiang, Junnan Zhu, Jiajun Zhang, Yu Zhou, and Chengqing Zong. 2020. Knowledge graph enhanced neural machine translation via multi-task learning on sub-entity granularity. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4495–4505.

Yikai Zhou, Baosong Yang, Derek F. Wong, Yu Wan, and Lidia S. Chao. 2020. Uncertainty-aware curriculum learning for neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6934–6944, Online. Association for Computational Linguistics.

9. Language Resource References

Flitto. 2021. *Korean-English Parallel Corpus of Specialized Domains*. AI Hub.