Understanding Why Polysemous Words Translate Poorly from a Calibration Analysis Perspective

Yucong Wu[1] Yusuke Miyao[1]
[1]The University of Tokyo
wu-yucong725@g.ecc.u-tokyo.ac.jp yusuke@is.s.u-tokyo.ac.jp

Abstract

Neural machine translation models have difficulty in translating polysemous words accurately. We hypothesize that this is because the translation models have calibration errors, i.e., the models give too low probabilities to rare senses and too high probabilities to frequent senses. To test this hypothesis, we propose a calibration analysis framework that observes how calibration errors change under different setups. Calibration errors of models' predictions are measured in terms of sense frequency and part of speech. The results indicate that machine translation models are underconfident in less frequent translations and overconfident in more frequent ones.

1 Introduction

Language learners often feel confused about certain usages of a word or phrase. Especially for those words with multiple meanings, it's hard to grasp the exact meaning or explanation for the whole sentence. Even though they refer to a dictionary, simple and abstract definitions sometimes make them feel more confused about the meaning of polysemous words. Resorting to machine translation unconditionally would not be a good choice [1].

In this paper, we perform calibration analysis on translation probability of polysemous words. Calibration analysis aims to figure out how and when we can trust and accept the translations produced by neural machine translation models. Calibration errors measure the difference between confidence score and prediction accuracy, and is a trustworthiness evaluation in high-stakes tasks like self-driving and medical diagnosis [2]. A model is called calibrated only when its prediction confidence matches its accuracy. Figure 1 shows two translations produced by a calibrated model given the English sentence. We could accept the

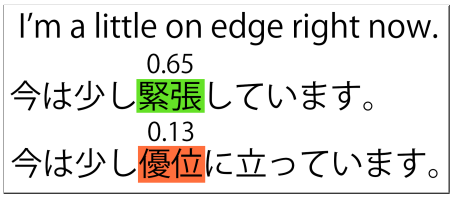

Figure 1: Two translations produced by a calibrated model. The numbers indicate the confidence score for the highlighted token

first sentence as the translation of 'edge' because its confidence is high. For the same reason, we should abandon the second because of the low confidence given by the calibrated model. Language learners can accept translations of the calibrated translation model based on its confidence.

We create a dataset containing polysemous words and their acceptable translation set to support our calibration analysis. We use accuracy-confidence plots to show the calibration errors of models on varied confidence bins. We also compare the tendencies of confidence distributions under different setups including sense frequencies and parts of speech. Our results show that machine translation models are underconfident in less frequent translations but overconfident in more frequent translations, which responses the question in the title.

2 Background

Calibration analysis has been applied to neural networks and machine translation models to help us know when the prediction could be trusted better. Guo et al. [3] observed that modern neural networks suffer from overconfidence about their predictions, and the lack of regularization might be a possible reason. Wang et al. [4] found that neural machine translation is not only overconfident in high-confidence predictions but also underconfident in low-confidence predictions. Kumar et al. [5] claimed that calibration helped improve interpretability, but the end-of-

sentence token is severely overconfident, causing translation to end quickly. Post-hoc calibration strategies including temperature scaling [6] and histogram binning [7] have been proposed to mitigate calibration errors produced by machine translation models.

DiBiMT [8] is an existing polysemous translation dataset. It evaluates whether different senses of polysemous words can be translated correctly. However, the previous polysemous translation dataset suffers from insufficient number of sentences and simple structures as we describe in Section 3.

Expected calibration error (ECE) is a popular metric to evaluate the extent of calibration errors of a model's prediction [9]. The confidence axis $[0, 1]$ is divided into bins with equal sizes, i.e., $B = \{\mathfrak{X}_1, \cdots, \mathfrak{X}_{|B|}\}$. X is the set of all predictions. ECE is calculated among $|B|$ bins by using the weighted average of absolute differences between accuracy $Acc(\mathfrak{X})$ and confidence $Conf(\mathfrak{X})$.

$$ECE = \sum_{i=1}^{|B|} \frac{|\mathfrak{X}_i|}{|X|} |Conf(\mathfrak{X}_i) - Acc(\mathfrak{X}_i)| \tag{1}$$

$Conf(\mathfrak{X})$ is defined as the average of all prediction probabilities (Eq. 2), where $\Phi(x)$ is the prediction probability on the sample x from the bin $\mathfrak{X}$. $Acc(\mathfrak{X})$ is defined as the ratio of correct predictions (Eq. 3), where $\hat{y}(x)$ indicates the prediction given by the model and $\mathbb{I}()$ is the indicator function:

$$Conf(\mathfrak{X}) = \frac{1}{|\mathfrak{X}|} \sum_{x \in \mathfrak{X}} \Phi(x) \tag{2}$$

$$Acc(\mathfrak{X}) = \frac{1}{|\mathfrak{X}|} \sum_{x \in \mathfrak{X}} \mathbb{I}(\hat{y}(x) = y) \tag{3}$$

3 Method

In this research, we investigate the translation quality of polysemous words using calibration error analysis. We collect a translation corpus containing polysemous words to support our calibration analysis. All of our corpora are obtained or transformed from *Projekt Deutscher Wortschatz* [10] Corpora English and WordNet [11] example sentences. We create a dataset with two thousand sentences containing polysemous words from the collected corpora. Our dataset consists of triples $(s, l, \mathcal{G})$, where s is a source sentence, l is a dictionary form (lemma) of a target polysemous word, and $\mathcal{G}$ is a set of acceptable translations for l. An example is:

("I'm a little on edge right now.", edge, { 緊張 })

Statistics	Ours	DiBiMT
Lemmas	129	200
Sentences	1901	599
Average Length	21.55	8.81
Sentences per Lemma	15	3

Table 1: Statistics of translation datasets containing polysemous words

Table 1 compares the major statistics of our created dataset with DiBiMT. It shows that our dataset provides a larger number of sentences for each polysemous word, and the average length indicates our dataset contains sentences with more complicated structures.

Calibration analysis is to observe how ECE and translation accuracy would change in different groups of polysemous words. The groups are divided based on sense frequencies or POS tags. ECE is used to measure the extent of calibration errors for translation probability of polysemous words. S represents the source sentence, X for the polysemous word, $Y = Y_1 \cdots Y_N$ for the translation sequence. $M(S, X, Y)$ indicates the index set in Y corresponding to X. The translation probability $\Phi(X, Y)$ of polysemous word is defined below:

$$\Phi(X, Y) = \prod_{t \in M(S,X,Y)} P(Y_t | X, Y_{<t}) \tag{4}$$

We run M2M100 [12] translation models on our dataset using its initial setup. $Conf(\mathfrak{X})$ can be calculated by substituting $\Phi(x)$ in Eq. 2 with Eq. 4. $|B|$ is set to 10. A translation is correct if and only if the corresponding translation of lemma is in $\mathcal{G}$.

4 Experiment

4.1 Analysis on Sense Frequency

We divide all senses into Most Frequent Sense (MFS), Frequent Senses (FS+), and Less Frequent Sense (LFS) based on their prior distribution in the corpus. MFS indicates the sense with the largest frequency, FS+ represents those senses in a descending order whose frequencies are cumulatively greater than 70%, and other senses are called LFS.

Figure 2 demonstrates the calibration errors within MFS, FS+, and LFS groups, respectively. We discover that the

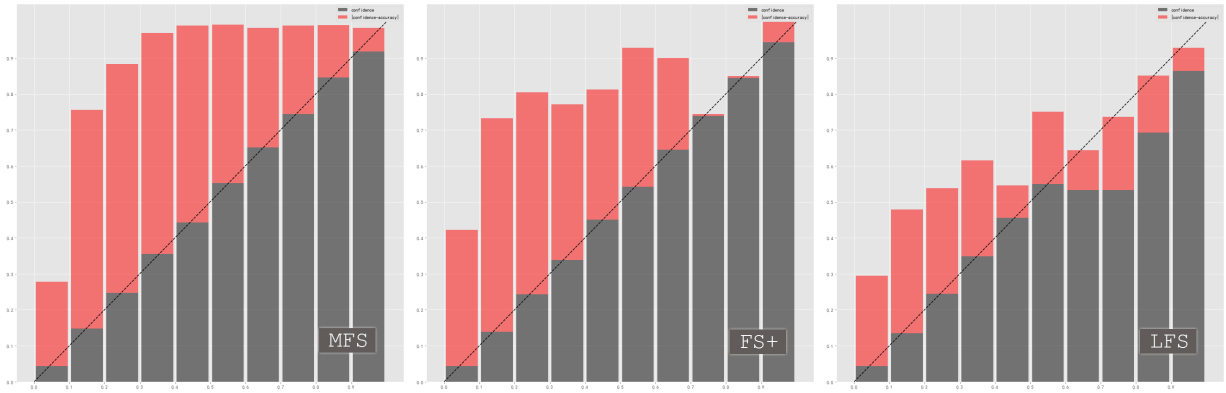

Figure 2: Accuracy-confidence plot for MFS, FS+, and LFS. The X axis shows confidence bins and the Y axis shows accuracy. Black bars indicate the minimum of accuracy and confidence, and red bars represent the difference between confidence and accuracy. The 45° lines represent perfect calibration, i.e., accuracy matches confidence exactly. Under/over confidence depends on whether the red bar is above/below the line.

	MFS	**FS+**	**LFS**	**MFS with FS+**
ECE	38.98	37.23	**23.03**	38.58
Accuracy	0.865	0.727	**0.506**	0.834

Table 2: ECE and accuracy for different sense frequencies

model is underconfident in low-confidence bins and overconfident in high-confidence bins when considering MFS samples only. The calibration of models changes slightly in FS+. In the highest confidence bin, it becomes underconfident. It shows well-calibrated in near high-confidence bins but still under-confident in low-confidence bins.

Comparing LFS with MFS and FS+, we find that the borderline between underconfident and overconfident bins moves to the left side of the axis from the 9th bin to the 7th. The difference between accuracy and confidence within each bin, i.e., the area of red bars, also declines. Despite its low accuracy, it demonstrates that the model is much more calibrated in less frequent sense sample sets.

Table 2 exhibits ECE and translation accuracy when only MFS, FS+, and LFS are used, as well as the mix of MFS and FS+. We can see the marginal decrease in ECE and accuracy in the LFS set. Although the accuracy decreases in LFS, the lower ECE shows that the model is much more calibrated in less frequent sense samples. It would be misleading if the machine translation model shows high confidence scores in low-accurate predictions. Comparing FS+ to MFS, we can also observe a slight decrease in ECE and accuracy. Less frequent sense would reduce translation accuracy and lower expected calibration errors.

Figure 3 shows the confidence distribution of correct

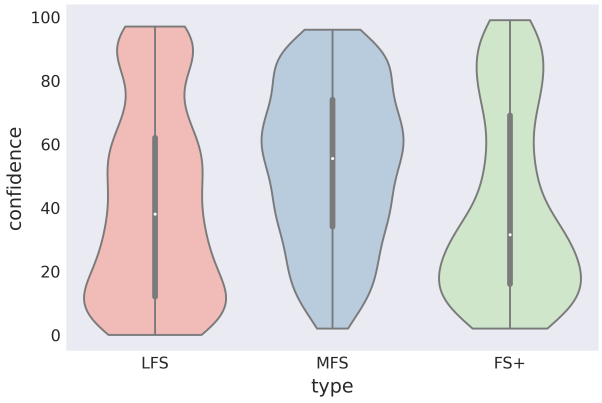

Figure 3: Kernel density estimation (KDE) curves of confidence distributions in LFS, MFS, and FS+. The wider the horizontal line is, the larger the distribution density is. (Gaussian Kernel, bandwidth = 12)

translations. We extracted all correct translations and group these translations into LFS, MFS, and FS+ according to their sense frequencies to draw this figure. Observing the distribution of LFS, there is a peak on low confidence bins of the KDE curve. Low confidence for most predictions indicates that models are uncertain about their decision even though it is correct. The current LFS group consists of all the samples we concern about, i.e., rare senses of polysemous words. The peak on low confidence bins indicates that the model is inclined to put too low probabilities on these tokens. This conclusion gives an available response toward the question in the title.

MFS shows different tendencies than LFS and FS+. Its peak is in the upper of the confidence axis and MFS puts less weight on the low confidence bin. The distribution

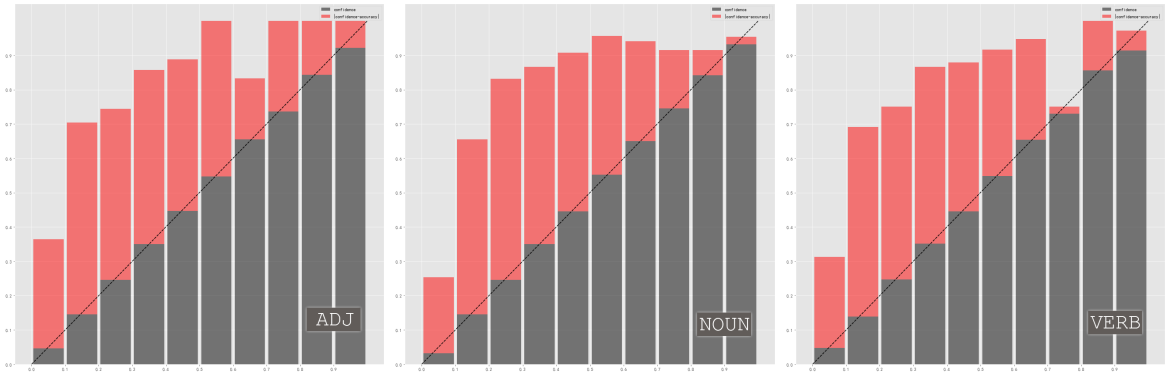

Figure 4: Accuracy-confidence plot for adjectives, nouns, and verbs.

	Noun	Verb	Adjective	All
Percentage	54.59	30.56	14.84	100
ECE	**33.11**	34.78	39.68	34.62
Accuracy	0.846	0.690	**0.658**	0.771

Table 3: ECE and accuracy for different POS

shows that models are confident about their translations of the most frequent sense. The reason is thought to be that models have seen sufficient numbers of MFS samples during training, and learned the pattern for collocation of MFS and other words, and hence uncertainty like LFS and FS+ disappears. The KDE curves of MFS, FS+, and LFS differ in shapes. LFS and FS+ are similarly bowling-shaped, whereas MFS is like a spindle.

4.2 Analysis on Part of Speech

Figure 4 illustrates calibration errors for adjectives, nouns, and verbs. All three major POS are underconfident in low-confidence bins but differ a little in high-confidence bins. Nouns and verbs are overconfident in high-confidence bins, while adjectives are underconfident in high-confidence bins. Verbs have higher accuracy in low-confidence bins than nouns. Adjectives are underconfident among all confidence bins.

Table 3 shows that noun senses have the smallest ECE, and adjectives have the largest. Adjectives have the lowest accuracy, but nouns have the highest accuracy.

Figure 5 analyzes the confidence distribution of correct translations for nouns, verbs, and adjectives. If a lemma has two senses with different POS, it will be much easier for models to discriminate its sense and translate it into correct expressions. Because nouns are the largest senses, there are plenty of lemmas having both noun and other POS senses. Therefore, the density would focus on and above

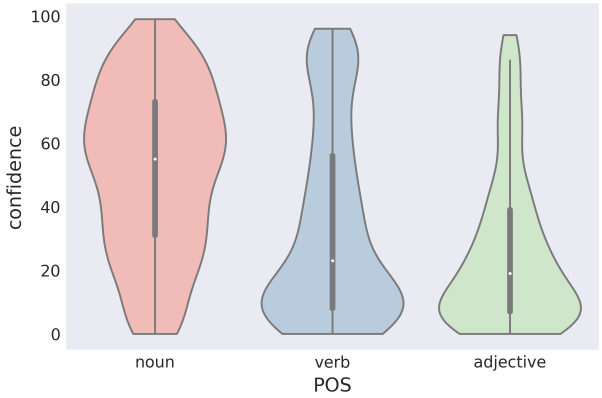

Figure 5: KDE curves for nouns, verbs, and adjectives. (Gaussian kernel, bandwidth = 12)

the middle bins and the KDE curve of the noun distribution looks like bell-shaped distribution. Verbs and adjectives have similar spindle-shapes in confidence distribution, i.e., the lower confidence bins have a higher frequency density whereas higher confidence bins have a lower density.

5 Conclusion

We created a machine translation test set containing polysemous words based on *Projekt Deutscher Wortschatz* and WordNet. We applied calibration analysis to the quality of polysemous word translation to investigate why polysemous words translate poorly. The proposed framework contains a comprehensive analysis on sense frequency and POS. Our analysis on sense frequency provides an effective perspective to show how the calibration errors of models change as sense frequency decreases. We also discover that nouns, verbs, and adjectives differ in confidence distribution shapes. Correct translations in LFS tend to report relatively low confidence, and this could response to the question in the title.

References

[1]Frederick Liu, Han Lu, and Graham Neubig. Handling homographs 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)**, pages 1336–1345, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

[2]Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving, 2019.

[3]Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In **Proceedings of the 34th International Conference on Machine Learning - Volume 70**, ICML'17, page 1321–1330. JMLR.org, 2017.

[4]Shuo Wang, Zhaopeng Tu, Shuming Shi, and Yang Liu. On the inference calibration of neural machine translation. In **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**, pages 3070–3079, Online, July 2020. Association for Computational Linguistics.

[5]Aviral Kumar and Sunita Sarawagi. Calibration of encoder decoder models for neural machine translation. **CoRR**, abs/1903.00802, 2019.

[6]PLATT J. C. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. **Advances in Large Margin Classifiers**, 1999.

[7]Bianca Zadrozny and Charles Elkan. Obtaining calibrated probability estimates from decision trees and naive bayesian classifiers. In **Proceedings of the Eighteenth International Conference on Machine Learning**, ICML '01, page 609–616, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers Inc.

[8]Niccolò Campolungo, Federico Martelli, Francesco Saina, and Roberto Navigli. DiBiMT: A novel benchmark for measuring Word Sense Disambiguation biases in Machine Translation. In **Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pages 4331–4352, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[9]Mahdi Pakdaman Naeini, Gregory F. Cooper, and Milos Hauskrecht. Obtaining well calibrated probabilities using bayesian binning. In **Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence**, AAAI'15, page 2901–2907. AAAI Press, 2015.

[10]Uwe Quasthoff. Projekt deutscher wortschatz. In Gerhard Heyer and Christian Wolff, editors, **Linguistik und neue Medien. Proc. 10. Jahrestagung der Gesellschaft für Linguistische Datenverarbeitung.**, Wiesbaden, 1998. Deutscher Universitätsverlag.

[11]George A. Miller. WordNet: A lexical database for English. In **Human Language Technology: Proceedings of a Workshop held at Plainsboro, New Jersey, March 8-11, 1994**, 1994.

[12]Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Edouard Grave, Michael Auli, and Armand Joulin. Beyond english-centric multilingual machine translation. **J. Mach. Learn. Res.**, 22(1), jul 2022.