Evaluation of MT Models on Polish to Korean Translations (Formal & Informal)

L. Kim, A. Lee, S. Prochowski

Special Topics: Natural Language Processing Final Project

Evaluation of MT Models on Polish to Korean Translations (Formal & Informal)

Austin Lee, Lara Kim, & Sabina Prochowski

Abstract

This paper compares the performance of three different existing pre-trained models (Facebook Hugging Face Model, Yandex.Translate, & OpenNMT) on Polish-Korean translations using both, direct translation and pivot translation via English. To evaluate the translation quality, we use the BLEU (bilingual evaluation understudy) score, a commonly used metric for MT evaluation. We also investigate the differences in BLEU performance between informal and formal Korean translations. To do so, we manually curated a corpus of 500 Polish sentences of varying sentence types and topics into 500 Korean informal translations and 500 Korean formal translations. Our results show that while direct translation performs rather well for some models, pivot translation through English can lead to improved translation quality for this language pair. We also find that the models perform drastically differently on informal and formal Korean translations, highlighting the importance of considering the specific nuances of the different language registers when evaluating machine translation performance. Finally, we investigated whether the Korean machine translations showed a preference for formal or informal translations across various sentence topics. This analysis revealed that all Korean machine translations favored formal translations across all sentence topics, which justified why all three machine translation models achieved higher BLEU scores on the formal translations than informal translations. Our research adds value in this field as these models lack default specification and / or documentation of formal or informal Korean used in their training and generated translations.

1. Introduction

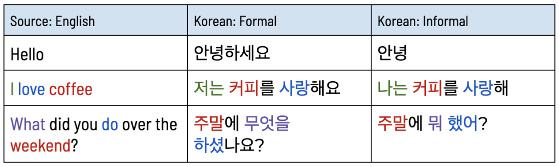
The development of precise and accurate machine translation is a significant field of research as there is a demanding need for cross-language communication.The research question we are investigating is how effective a machine translation system can be in accurately translating Polish sentences to Korean. Currently, translating from Polish to English and then from English to Korean is more common-practice than a direct translation system between Polish and Korean. But because English acts as an intermediary between the two languages, there can be several translation errors and loss of meaning, especially for complex sentences, cultural nuances, and idiomatic expressions.

Machine translation still has room for improvement, especially when it comes to informal and formal language translations. Formal and informal language can differ significantly in terms of grammar, vocabulary, and syntax, which makes it difficult for MT models to accurately capture the intended meaning of a sentence. This issue is particularly evident in Korean, a language in which handling of the translations of honorifics has historically received little attention in machine translation (Yongkeun Hwang et al., 2021).[1] Hence, this project aims to compare the performance of three pre-trained models on Polish-Korean translations, considering both formal and informal registers. This is a valuable topic as most existing pre-trained models lack default specifications or documentation for if they utilize formal or informal Korean in their training corpus, making it difficult to assess their effectiveness accurately (Yongkeun Hwang

[1] See https://www.mdpi.com/2079-9292/10/13/1589 for further reference.

et al., 2021). In fact, there is not much documentation / information on what sets the two Korean registers apart on the English web alone at least from what we observed during our research. As such, it is critical to conduct more research on the effect of machine translation on Korean in regards to honorifics.

In the Korean language, as displaying deference and courtesy towards elders and superiors is ingrained within Korean culture and continues to be considered of utmost importance, honorifics play a crucial role in communication. The form of language, informal or formal, affects the meaning, tone, and structure of the formed sentences significantly. An example of such a case can be seen even with the simplest of dialogues. As can be seen in Figure 1, the English word "Hello" can have two distinct variations; in Korean, a translation to formal language would be "안녕하세요," while in the informal language, it would be "안녕".



Source: English	Korean: Formal	Korean: Informal
Hello	안녕하세요	안녕
I love coffee	저는 커피를 사랑해요	나는 커피를 사랑해
What did you do over the weekend?	주말에 무엇을 하셨나요?	주말에 뭐 했어?

Figure 1: Example Translations from English to Formal/Informal Korean

Evidently, when it comes to translations, the differentiation between the usages of formal / informal language and judgment can and will alter the product notably. However, translation models are incapable of passing judgment on what form of language to use without context or additional information, which commonly leads to improper formatting and incorrect syntax in the course of translation (Yongkeun Hwang et al., 2021).[2] We seek to address this problem and generate a range of viable strategies with this project.

Hence, the native Polish speaker, Sabina Prochowski, of the group curated a dataset of 500 Polish sentences of distinct sentence structures and topics and then the two native Korean speakers, Lara Kim and Austin Lee, conducted the human translations. To ensure quality and consistency of the annotations, IAA was measured for both, informal and formal translations. Three existing pre-trained models (Facebook Hugging Model, Yandex.Translate, & OpenNMT) were used to translate the dataset. Then, the models' performances were measured based on the evaluation metric, BLEU, which is widely adopted by the machine translation community, to be discussed below.

Additionally, while direct translation performs well for some models, pivot translation can lead to improved translation quality, especially for under-resourced languages (Bogdan Babych et al., 2017).[3] By highlighting the limitations and capabilities of existing pre-trained models in this area, this study can inform the development of more accurate and effective machine translation systems for these languages.

Then, the model that performed best, i.e. attained the highest bleu score, was “fine-tuned” by using adaptation techniques, which involved adding additional parameters and code. While it is standard to stick to either form of speech in one sentence, formal or informal in the Korean

[2] This study delves into the language-specific problems that Korean poses, including translation of honorifics and advanced methods to approach them. The authors develop a context-aware NMT to enhance machine translations.

[3] This study analyzes two translation methods of direct translation and a translation via a cognate language. The latter shows that it achieved better translation quality by leveraging available advanced dictionaries and grammars and syntactic / lexical similarities between the source and pivot languages.

language, we observed that even the highest achieving model was producing translated sentences that were mixing the two forms of language together within the sentence, as well as varying throughout the dataset without a constant factor. As such, although we created two separate versions of Korean translations, one in formal and one in informal language, we found that the Bleu score was still relatively low. In many languages, such as English, honorifics are not pertinent and do not affect conjugations or structure of a sentence. In fact, even Polish and Korean differ significantly in terms of their use of honorifics which may be a preliminary reason contributing to the poor outcomes. While honorifics play a crucial role in the Korean language and are used extensively to express social status and respect in daily interactions, the use of honorifics in Polish is more limited and generally used in formal or professional contexts. However, it is worth noting that the degree of honorific usage in both languages can vary depending on the speaker's age, gender, and social status, in addition to the context of the conversation. As such, we believed fine-tuning the model to homogenize form of speech would be an efficient approach to producing more accurate and consistent results that would better align with our data.[4]

2. Related Work

Despite several large-scale meta-evaluations (Callison-Burch et al., 2006; Koehn and Monz, 2006) revealing significant disparities between its results and human judgments of translation accuracy, BLEU (Papineni et al., 2002) continues to be the most commonly used metric for MT evaluation. The BLEU score is the proposed statistical definition of essentially "closeness to human translation." Human translation from source to target is still seen as superior to MT as it preserves meaning and finds idiomatic expressions with similar connotations, changes the order / structuring of a sentence to reflect natural order of target language, and produces a dynamic, intended rather than literal, word-for-word translation. The caveat is that it is a slow and laborious process for humans. Thus, human translation is a time-consuming craft that requires years of practice to perfect, which makes it impractical for large-scale translation projects. Human translation is also extensive because it can account for the cultural context of source & target. On the other hand, computers can translate at an expedited rate and can process large amounts of text quickly and efficiently, but the quality of their translations is often not up to par due to the lack of nuance and understanding of the cultural context.

Although machine translation technology has improved in recent years, it still struggles with producing translations that are on par with those of human translators. Since every language contains a vast amount of linguistic data, training can take a long time. Even then, the generated translations may still be literal and awkward, making them less desirable for applications that require a high degree of accuracy and naturalness. However, they are fit for tasks where some degree of error is to be expected and tolerated, like movie caption translations. This supports our motivation to manually curate a corpus of 500 Polish sentences translated into both informal and formal Korean, in order to evaluate the performance of pre-trained models on this

[4] This article (Lijie Wang et al., 2019) is very aligned with our intended research as it highlights how current neural MT models lack honorific generation. To address this issue, the paper discusses using honorific fusion training loss and a data labeling method to tag training data. The results are promising as they significantly improve honorific generation ratio by 34.35% and 45.59%, motivating us to use honorific fusion inspired fine-tuning for our project.

language pair. By providing additional data, our research can contribute to improving the accuracy and naturalness of machine translations for this language pair, thereby making them more useful for practical applications, such as in fields of business, diplomacy, or legal documents, where accuracy and fluency are crucial.[5]

The Facebook Hugging Model is considered to be a state-of-the-art MT pretrained model that has presented promising results on 100 language pairs, including Polish and Korean (Angela Fan et al., 2020). It is based on the Transformer architecture that permits direct translation between languages without pivoting through English, which has proven to be particularly effective in MT tasks. Specifically, the model referenced as M2M-100, is a many-to-many multilingual model that contains data directly through distinct directions whereas past MT systems are built based on an english-centric multilingual model that contains training data to and from English (Angela Fan et al., 2020). Since the model has already been fine-tuned specifically for MT tasks, it is a well equipped model for our project to handle the nuances and complexities of target to source translations. However, it does hold some limitations as the training for Polish and Korean are built off bridge languages which implies that it may not accurately capture the nuances and idiosyncrasies of the original languages. You can also see that Polish and Korean training is not as plentiful in comparison to other languages as seen in Figure 2.

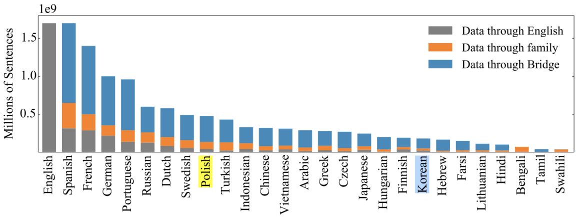


Figure 2: M2M-100 Training Data Statistics[6]

Yandex.Translate is another MT model that can be used to assess the accuracy of translating from Polish to Korean. Yandex.Translate is an online translation service developed by Yandex. Similar to “Google,” Yandex is a technology company based in Russia that provides various online services, including search engines, email, maps, news, and more. Yandex.Translate uses a combination of statistical and neural machine translation (NMT) technologies to provide translations between different languages. Statistical machine translation uses statistical models that are trained on large datasets of bilingual texts to automatically learn the patterns of how words and phrases in one language correspond to words and phrases in another language. This approach was widely used before the advent of NMT and is still used in some machine translation systems, including Yandex.Translate. Neural machine translation, on the other hand, uses neural networks to model the probability distribution of translations based on the context of the source text. This approach has been shown to outperform statistical machine translation.

[5] There is merit in utilizing machine translation and human translation in complementation, as the article highlights the benefits of combining machine and human translation to achieve better results, similar to our goal of this project. There is great importance in evaluating MT systems from the perspective of the translation task and the user’s needs (J Hutchins et al., 2001).

[6] This figure is taken from the paper that introduced the M2M100 418M model (Angela Fan et al., 2020).

Yandex.Translate is recognized for its ability to translate Eastern European languages with relatively high accuracy. Yandex Cloud claims that Translink, the fifth largest translation agency in Eastern Europe, was able to accelerate their work processes by 30% while also reducing routine operations costs after integrating machine translation with Yandex.Translate. Additionally, Yandex.Translate is set up to continuously train their data through resources on the Internet.[7]

OpenNMT is the third model that was used to evaluate the accuracy of machine translations. In order to learn the mapping between the source language and the target language, OpenNMT employs a deep learning architecture called a sequence-to-sequence model (Klein et al., 2017). Other methods used by OpenNMT to enhance translation quality include attention mechanisms, which allow for the model to prioritize more relevant portions of the source text while translating, and beam search algorithms, in which the model generates multiple translations and chooses the most probable. OpenNMT has demonstrated to be a powerful and adaptable tool for creating high-quality machine translations across a variety of languages and topics. OpenNMT also serves as a flexible platform that offers many features for researchers such as us working on low-resource language pair machine translation projects, and continues to be developed actively by a large community of contributors.

As deep neural networks have advanced, substantial improvements have been made in neural machine translation (NMT) capabilities. Despite these advancements, language-specific issues like honorifics have largely been overlooked and remain underexplored (Hwang et al., 2021). Given the current state of machine translation, there is ongoing research into the implementation of formal and informal language registers within machine translation. These include using heuristics, context-aware post-editing (CAPE) technique, and side-constraints (Weston et al., 2019, Sennrich et al., 2016).[8]

In machine translation research, the choice of register, or the degree of formality or informality within the source text has a significant effect on the accuracy and fluency, as well as the tone and style of resulting translations. Informal language registers can involve various complications such as colloquialisms, slang, or difficulties such as conjugations for formal language registers. The choice of register also has impacts on the quality and relevance of data used to train and evaluate MT models. To address these difficulties, MT researchers continue to develop various techniques and methods to optimize models for these different registers, including the application of specialized training data and domain adaptation (Niu et al., 2018).[9]

We acknowledge that English may not be the best fit pivot language, like another

[7] This study goes in depth with Yandex.Translate's API. It was found that this translating model is based on statistics derived from web sources, making the translation more flexible and almost limitless.

[8] The research that has gone into MT honorifics deal with translations on parallel texts. This approach allows a reliable reference to rely on motivating our research on using parallel texts as well.

[9] This study focuses on generating natural language with differing levels of formality. The authors propose a multi-learning approach that will be able to solve two tasks simultaneously: monolingual formality transfer and formality sensitive machine translation. The study demonstrates that their approach is capable of performing formality-sensitive translation without the need for explicit training on style-annotated examples.

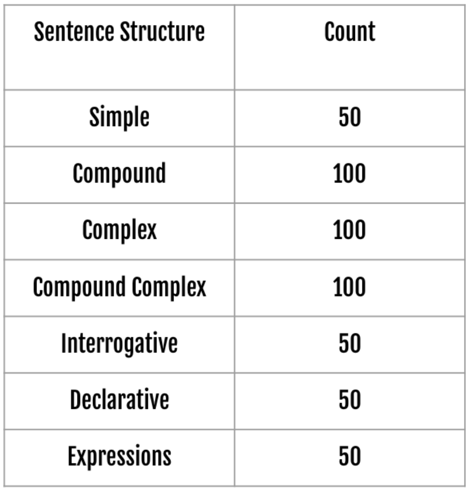
Asian language would have been for pivot translation from Polish to English to Korean. However, due to the lack of proficiency in other languages of all group members, English was utilized as the pivot language, as it would be the only communicable way for analysis. As the (Bogdan Babych et al., 2017) paper suggests, investing in machine translation for closely related languages can yield better outcomes than developing systems from scratch for new translation directions. In fact, past studies have shown that MT has revealed great success in producing high-quality output in languages that are closely related like Czech and Slovak (Hajic et al., 2000) or Ukrainian and Russian (Bogdan Babych et al., 2017).

3. Data Preparation

The data for this project was generated from diverse avenues to ensure the dataset was comprehensive and diverse. The data is a combination of AI-generated sentences, sentences from news sources, and original sentences created by Sabina specifically for the study. When using ChatGPT for the generation of some of these sentences, requests were made to confirm structuring and topics of sentences that were factual and cross–checked by Sabina. Others were taken from Polish news sources or American news sources translated into Polish directly by Sabina. Furthermore, some sentences were created by Sabina alone, especially the conversational / everyday expressions as they tend to be culturally specific and difficult for AI to capture efficiently. Please reference the master data sheet to see the sources of each sentence.

The sentence structure was broken down into simple, compound, complex, compound complex, interrogative, declarative, and expressions while the sentence topics included conversational / everyday expressions, politics, sports, business, health & wellness, travel, and food & cuisine. This diversity of sentence structure and topics was utilized to ensure that the corpus was representative of a wide range of language data, which is crucial in providing a reliable dataset to the field as it is attended to be used to improve accuracy and naturalness of MT, especially if it were to be used for training or fine-tuning by us or others in the MT industry.

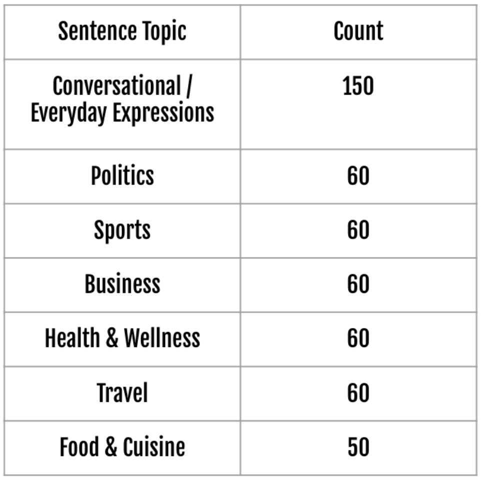
Below, you can find the amount of data allocated to each category within sentence structure and sentence topic. In Table 1, you can see that simple, interrogative, declarative, and expressions all each have 50 sentences whereas compound, complex, and compound-complex sentences have 100 sentences each. The allocation of more sentences to the latter is because these sentences are often used in written language and therefore, more prevalent in news articles, documents, etc. This exposure helps to improve the naturalness of machine translations. Nonetheless, it is important to include sentences of different varieties so that the model can be tested on different cases that it may encounter in real-world applications.



Sentence Structure	Count
Simple	50
Compound	100
Complex	100
Compound Complex	100
Interrogative	50
Declarative	50
Expressions	50

Table 1: Polish Sentence Structures Breakdown

Below, you can find that politics, sports, business, health & wellness, and travel topics each hold 60 sentences along with 50 food & expressions and 150 conversational / everyday expressions. By including a higher number of conversational phrases, the project ensures that the MT models are better equipped for handling cultural specifics of common everyday language, which MT models may not be as familiar with. There is also allocation of fewer sentences to food & cuisines as food-related language data may be less present in written materials than the other topics, which are often used to train MT models as this source acknowledges (Takayuki Sato et al., 2016). In addition, the diverse sentence topics enabled the analysis of possible impact of sentence subject matter on the models' capacity for precise translation across different areas of interest. Thereby, uncovering any discrepancies in translation quality in regards to formal and informal registers that may be dependent on the topic.



Sentence Topic	Count
Conversational / Everyday Expressions	150
Politics	60
Sports	60
Business	60
Health & Wellness	60
Travel	60
Food & Cuisine	50

Table 2: Polish Sentence Topics Breakdown

After the completion of 500 Polish sentences, Austin and Lara individually translated the 500 Polish sentences into Korean, totalling in 1000 Korean sentences in total. The IAA was then calculated to gauge the agreement between the two Korean translations. This step was done to ensure a reliable reference to fall back on when comparing the annotator's translations to the machine translations. Given that the IAA produced 47 differing translations, Austin and Lara worked on homogenizing the two sets of Korean translations to create one set of 500 Korean sentences that were suitable for the research. The translations were done with cultural intentions, colloquialisms, and intended meaning (behind certain phrases) in mind. With the homogenized data set, Austin and Lara carefully worked on the translations to create a set of 500 formal and informal Korean sentences.

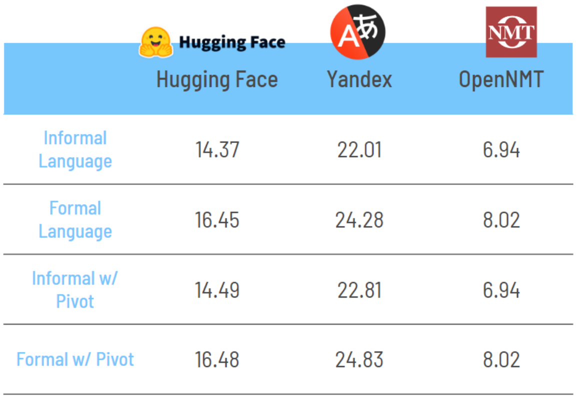
4. Methodology & Experiments

Each member of the group took on their role and divided their attention to a specific model. Sabina was responsible for serving as the primary writer of the project proposal and final write-up as well, preliminary researcher for the related work of the group, as well as the programming evaluator of the results for the group. Austin was responsible for translations and research in regards to approaching the evaluation process of the MT models for the group. Lara also partook in creating and evaluating manual translations as well as machine translations for one of the models, developed the fine tuned model, and assessed the Inter-Annotator Agreement for the data.

First, the Inter-Annotator Agreement (IAA) using statistical measure Cohen's Kappa was used in order to assess the precision and agreement between the two independent annotators, Austin and Lara. The IAA score was computed through our system.The IAA score was computed to ensure a reliable reference to fall back on.

Every member of the group used the 500 polish sentences to obtain 500 Korean direct machine translations using their designated models. As well as this, an additional 500 Korean machine translations were created by each member (on their designated models) with the implementation of a pivot translation in between, that being English. Each member used their two sets of machine translations to calculate BLEU scores against the manual 500 formal and 500 informal Korean translations. The BLEU scores were calculated using the SacreBleu module. As such, each translation model had four BLEU scores that were able to be analyzed. In addition to this, Austin and Lara programmed a script that counts the number of formal and informal translations for each sentence topic. The results of the formal and informal count are presented in the results below.

5. Results



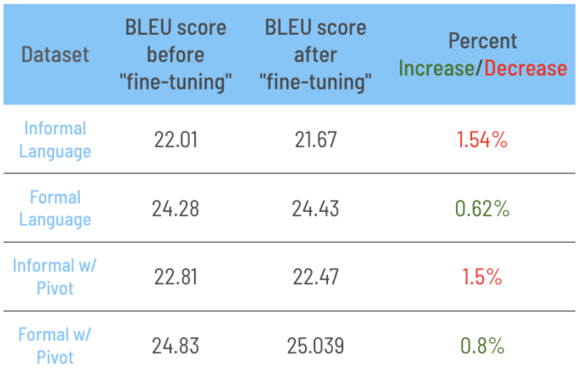
	Hugging Face	Yandex	OpenNMT
Informal Language	14.37	22.01	6.94
Formal Language	16.45	24.28	8.02
Informal w/ Pivot	14.49	22.81	6.94
Formal w/ Pivot	16.48	24.83	8.02

Table 3: Results of BLEU Scores for all Models on Different Datasets

Yandex.Translate yielded significantly higher BLEU scores than the other two models in all areas which we can justify as Yandex.Translate is recognized for translating Eastern European languages exceptionally for a MT model. The Yandex.Tranlsate model yielded the highest BLEU scores out of the three models, with a score of 22.01 for the dataset with informal language and 24.28 for the dataset with formal language. As for the implementation of the pivot language, the BLEU scores went up for each respective Korean Language register. For the data sets with informal language, the BLEU score increased by 0.8 with a new total of 22.81. For the data sets with formal language, the BLEU score increased by 0.55 with a new total of 24.83.

As for the Hugging Face model, it obtained a result of 14.37 for the informal Korean translation dataset and 14.49 for the formal Korean translation dataset. We noticed an increase in BLEU score of .12 on the informal dataset with pivot vs. without the pivot and an increase of .03 in the run of the model on the formal dataset with the pivot from the BLEU score received on the formal dataset without the pivot.

The OpenNMT model yielded the lowest BLEU scores out of the three models, with a score of 6.94 for the datasets with informal language and 8.02 for the datasets with formal language. The results remain unchanged after using the pivot with the formal and informal data.



Dataset	BLEU score before "fine-tuning"	BLEU score after "fine-tuning"	Percent Increase/Decrease
Informal Language	22.01	21.67	1.54%
Formal Language	24.28	24.43	0.62%
Informal w/ Pivot	22.81	22.47	1.5%
Formal w/ Pivot	24.83	25.039	0.8%

Table 4: Results of BLEU Scores After Fine-tuning Yandex.Translate Model on Different Datasets

Since Yandex.Translate produced the highest BLEU scores, we aimed to achieve a higher score with fine-tuning by modifying the Yandex.Translate model to produce more accurate Korean translations in only the formal language to avoid the intermixing of formal / informal forms of speech in a line, we observed improvements in the BLEU score with the formal language datasets. We did this by adding code that would detect informal suffixes or prefixes in the Korean language and convert them to the formal language. We were able to add a glossary of these prefixes / suffixes and integrate it into the model. However, after this modification, there were also decreases in the BLEU scores for the datasets with informal language. Although this was the case, even if misaligned with our dataset, when going through the machine translations after the finetuning manually, the results proved to be both more consistent and more accurate than before.

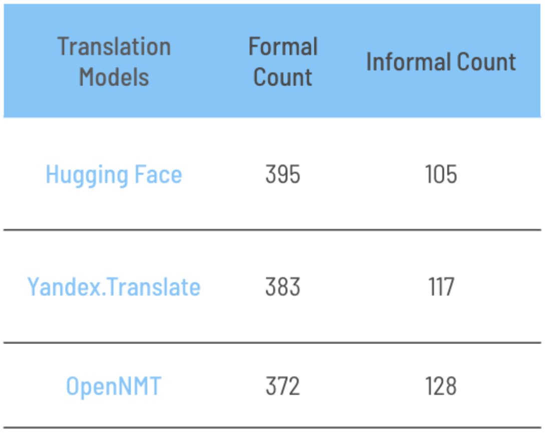
Our system also analyzed the amount of formal and informal translations each translation model outputted. This was to gauge whether sentence topics had an impact on the honorifics of the machine translations themselves.

6. Discussion / Evaluation

The metric used to calculate BLEU score ranges from 0-100 in which a higher score indicates better translation quality. Results with a BLEU score under 10 are generally considered poor (like the OpenNMT results), with a high number of errors and poor fluency, whereas results with a BLEU score of ~25 are considered moderate quality (Yandex.Translate's results), with fewer errors and better fluency, but may still have some inaccuracies and lack of naturalness (Google Cloud, 2022). However, according to past research with other relatively low resource language pairs, our findings align with comparable results (Marcely Boito et. al, 2022).[10]

Upon further research, the availability of pre-trained models varies depending on the language pair. As demonstrated by the consistent BLEU scores for the datasets with and without English as a pivot language, for low resource language pairs such as Polish and Korean, the model depends on an intermediary language, English, in order to try to improve the quality of the translations. Due to this, we observed a lot of loss in meaning and accuracy in the outputs of the models, especially OpenNMT. OpenNMT is designed to be a customizable platform that can be employed as a framework and baseline for users to develop their own model (Ranathunga et al., 2021). Consequently, rather than using OpenNMT to create machine translations reliant on pre-trained models, it can be more useful as a starting point when attaining tools and resources to build a model.

Regarding the counts of formality within the machine translations, all models received a higher BLEU score when tested against the formal Korean translations dataset. This finding is consistent with our observation that all system outputs tended to generate more formal translations than informal ones as shown on table 5.

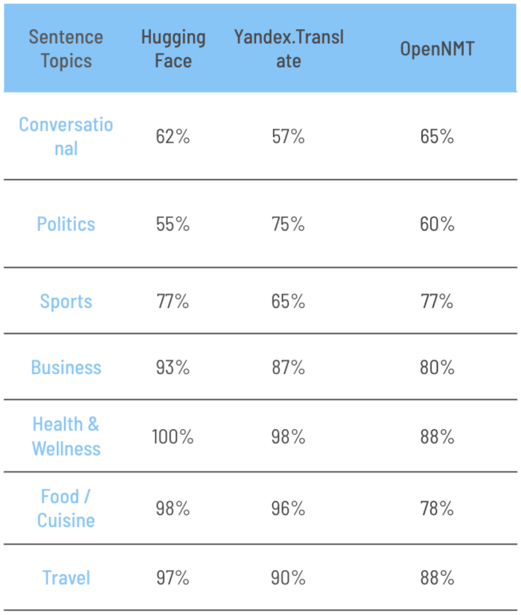


Translation Models	Formal Count	Informal Count
Hugging Face	395	105
Yandex.Translate	383	117
OpenNMT	372	128

Table 5: Formal to Informal Counts for each Translation Model

[10] There are comparable results of a low resource language in this paper of Tunisian Arabic to English translation systems that score in the range of 12-16.

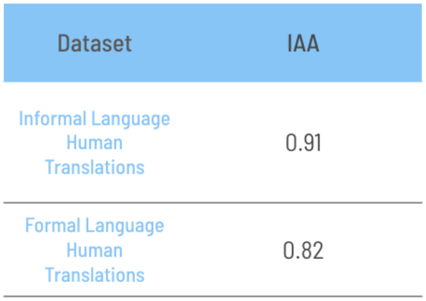
In addition to the analysis of formal and informal counts on each translation model, the effect of sentence topic on formal and informal translations were analyzed as well, as shown on table 6.



Sentence Topics	Hugging Face	Yandex.Translate	OpenNMT
Conversational	62%	57%	65%
Politics	55%	75%	60%
Sports	77%	65%	77%
Business	93%	87%	80%
Health & Wellness	100%	98%	88%
Food / Cuisine	98%	96%	78%
Travel	97%	90%	88%

Table 6: Formal Count Percentages for each Sentence Topic

As shown in the data, there is little correlation that goes into whether or not sentence topic has an effect on the translation. Possible explanations as to why the models chose to translate to either formal or informal could be from training data, model architecture, tokenization, and pre-methodology processing.



Dataset	IAA
Informal Language Human Translations	0.91
Formal Language Human Translations	0.82

Table 7: IAA Scores for Datasets

$$\kappa = \frac{p_o - p_e}{1 - p_e}$$

$$0.91 = \frac{0.908 - 0.001824}{0.84 - 0.111}$$

$$0.82 = \frac{0.84 - 0.111}{1 - 0.111}$$

Equation 1: IAA Calculations

The IAA was found to measure the level of agreement of the annotators tasked with annotating the same set of data to ensure quality and consistency. The annotations of Korean translations generated an IAA score of 0.91 for the informal dataset and 0.82 for the formal dataset as shown on table 7. Both of these IAAs indicate a good to very high level of agreement, indicating that the annotation guidelines were well-defined and the annotators had similar levels of linguistic expertise in Korean to ensure consistent translating. These scores were computed using a program that took the two annotation files as inputs and defined a cohen_kappa() method (as shown in equation1) that compares each token in the text file, obtains a count, computes the expected agreement, and finally outputs the IAA value. The lower IAA score for the formal language dataset can be attributed to greater variation and complexity associated with this form of speech. Additionally, differences in beliefs / perspectives, as well as ambiguity were factors in the disagreement between our annotators (Abercrombie et al., 2023). Nevertheless, these IAA scores demonstrated a high level of agreement and consistency between Austin and Lara. This was ensured by our thorough process, which involved carefully going through every one of the 500 lines of corpora to properly understand and consider each of the translations to minimize loss of meaning or structure.

The effort to finetune improved the Yandex.Translate model's bleu score by .2 points. This was a relatively small improvement, as this did not solve the issue of inaccurate structure, incorrect expressions or conjugations. However, it was a successful approach to confronting issues regarding the usage of both formal/informal language within one sentence by the model before the fine-tuning process.

7. Conclusion

This study provides concrete data on the performance of three pre-trained translation models (Facebook Hugging Model, Yandex.Translate, and OpenNMT) for Polish-Korean translations, exploring both direct and pivot translation methods as well as the impact of formal and informal language registers. Our results indicate that while direct translations prove to be effective, the pivot translation approach through English provides an improvement, with an average improvement of 0.28 points across the three models for this particular language pair.

Furthermore, the significant disparities in translation quality between informal and formal Korean emphasize the importance of considering the nuances of the target language and its registers when evaluating machine translation performance. In the Korean language, the appropriate use of honorifics is of paramount importance, as they play a crucial role in conveying politeness, respect, and social hierarchy, making them a vital aspect to consider in the development and evaluation of machine translation models.

This research contributes to the broader understanding of the limitations and capabilities of pre-trained NLP models in handling various language pairs and translation methods, and highlights the need for further investigation in this area. The absence of default specifications or documentation for formal or informal Korean in these models calls for future research to address this gap, ultimately leading to better translation tools and resources. By addressing these challenges, the NLP community can continue to improve machine translation models and better serve the diverse needs of users across different languages and linguistic contexts.

8. Future Work

If more time was permitted for this project, certain modifications and additions would be implemented to improve its outcome. Firstly, a greater amount of time would be allocated towards incorporating more data into the corpus. 500 Polish sentences is a small corpus that may not be sufficient for effective testing of the models. A bigger corpus would also serve to be beneficial for fine-tuning as fine-tuning on more data can help enhance the performance of the pretrained models, resulting in higher BLEU scores. We would also explore another corpus which includes a mix of both formal and informal translations, which would be to be agreed and discussed with more speakers for reliability.

To ensure the accuracy of the Polish sentences and Korean translations, we would incorporate the assistance of other Polish and Korean proficient speakers. This would ensure that the manual sentences and translations are accurate and reliable, which is crucial for the success of the MT.

Additionally, while BLEU score is a commonly used metric for machine translation, exploring other metrics such as TER, METEOR, and ROUGE could provide a more comprehensive understanding of the models' performance. Thus, considering these metrics in addition to the BLEU score could potentially provide greater insight into the effectiveness of the models.

We would also explore using a different high resource pivot language, which would more likely achieve a higher score, but pursuing this option would take the project into a different direction. An alternative shows us that pivot-based transfer learning works best when both, the source to pivot and pivot to target are high resource language pairs and the source to target is a low resource language pair. However, in cases like Indic languages, the pivot to target language pair is considered to not be a high resource pair, which implies that we should use multiple related languages to pivot languages to assist the source to target with a study showing that using multiple pivot language resulted in a 2.03 BLEU increase over a baseline model (Shivam Mhaskar, 2022).[11]

We originally intended to fine-tune all of the pretrained models according to our datasets with appropriate training, validation, and test splits. However, after running into consistent errors, whether that was with Yandex.Translate's or OpenMT's lack of resources on fine-tuning and Facebook Hugging Model's deprecations and memory capacity issues with training even with a v100 or a100, as well as the time constraints with this project, we decided to follow an alternate approach of pivot translating and then "fine-tuning" the data for the best system. However, if we were to have more time and technical assistance, fine-tuning would be worth exploring as this should have a positive impact on BLEU score. It would also be interesting to develop a model of our own using these existing models and the data we've gathered in order to create one that could handle these informal/formal registers and produce more accurate translations. As we noted our concern in the beginning in not having a partner with strengths in programming, this would be a challenge to do on our own within this time constraint.

[11] This study presented a task to improve the performance of the English to Marathi NMT model which originally used Hindi as a closely-related high resource pivot language. However, since the English-Hindi pair is a high resource language pair, but the Hindi-Marathi pair is not, multiple Indic languages were used as pivot languages to attain better translations.

References

Abercrombie, G., Rieser, V., &amp; Hovy, D. (2023, January 25). Consistency is key: Disentangling label variation in Natural Language Processing with intra-annotator agreement. arXiv.org. Retrieved April 4, 2023, from https://arxiv.org/abs/2301.10684

Artstein, R. (n.d.). Inter-Annotator Agreement. Retrieved April 4, 2023, from http://ron.artstein.org/publications/inter-annotator-preprint.pdf

Babych, B., Hartley, A., &amp; Sharoff, S. (2017, January). Translating from under-resourced languages: Comparing direct transfer against pivot translation. ACL Anthology. Retrieved April 6, 2023, from https://aclanthology.org/2007.mtsummit-papers.5/

Boito, M. Z., Ortega, J., Riguidel, H., Laurent, A., Barrault, L., Bougares, F., Chaabani, F., Nguyen, H., Barbier, F., Gahbiche, S., &amp; Estève, Y. (n.d.). On-trac consortium systems for the IWSLT 2022 dialect and low-resource speech translation tasks. ACL Anthology. Retrieved April 4, 2023, from https://aclanthology.org/2022.iwslt-1.28/

Callison-Burch, C., Osborne, M., &amp; Koehn, P. (n.d.). Re-evaluating the role of Bleu in Machine Translation Research. ACL Anthology. Retrieved April 16, 2023, from https://aclanthology.org/E06-1032/

Fan, A., Bhosale, S., Schwenk, H., Ma, Z., El-Kishky, A., Goyal, S., Baines, M., Celebi, O., Wenzek, G., Chaudhary, V., Goyal, N., Birch, T., Liptchinsky, V., Edunov, S., Grave, E., Auli, M., &amp; Joulin, A. (2020, October 21). Beyond english-centric multilingual machine translation. arXiv.org. Retrieved May 1, 2023, from https://arxiv.org/abs/2010.11125

Hwang, Y., Kim, Y., &amp; Jung, K. (2021, June 30). Context-aware neural machine translation for Korean honorific expressions. MDPI. Retrieved May 4, 2023, from https://www.mdpi.com/2079-9292/10/13/1589

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., &amp; Amodei, D. (2020, January 23). Scaling laws for neural language models. arXiv.org. Retrieved April 4, 2023, from https://arxiv.org/abs/2001.08361

Kishore Papineni IBM T. J. Watson Research Center, Salim Roukos IBM T. J. Watson Research Center, Roukos, S., Todd Ward IBM T. J. Watson Research Center, Ward, T., Wei-Jing Zhu IBM T. J. Watson Research Center, Zhu, W.-J., &amp; Metrics, O. M. V. A. (2002, July 1). Bleu: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics. DL Hosted proceedings. Retrieved April 2, 2023, from https://dl.acm.org/doi/10.3115/1073083.1073135

Klein, G., Kim, Y., Deng, Y., Nguyen V., Senellart J., Rush, A. M., (2018, May). The OpenNMT neural Machine Translation Toolkit. Retrieved April 5, 2023, from https://arxiv.org/abs/1805.11462

Klein, G., Zhang, D., Chouteau, C., Crego, J., Senellart, J. (2020, July). Efficient and High-Quality Neural Machine Translation with OpenNMT. In Proceedings of the Fourth Workshop on Neural Generation and Translation, pages 211-217. Association for Computational Linguistics. Retrieved April

3, 2023, from https://aclanthology.org/2020.ngt-1.25/

Meyers, A. (n.d.). Machine Translation Lecture Notes. Retrieved April 14, 2023, from https://cs.nyu.edu/courses/spring23/CSCI-UA.0480-057/

Mhaskar, S., &amp; Bhattacharyya, P. (n.d.). Multiple pivot languages and strategic decoder initialization helps Neural Machine Translation. Aclanthology. Retrieved April 18, 2023, from https://aclanthology.org/2022.loresmt-1.2.pdf

Niu, X., Rao, S., Carpuat, M. Multi-Task Neural Models for Translating Between Styles Within and Across Languages. Retrieved April 18, 2023, from https://arxiv.org/pdf/1806.04357.pdf

Papineni, K., Roukos, S., Ward, T., &amp; Zhu, W.-J. (n.d.). Bleu: A method for automatic evaluation of Machine Translation. Aclanthology.org. Retrieved April 19, 2023, from https://aclanthology.org/P02-1040.pdf

Park, C., Yang, Y., Park, K., &amp; Lim, H. (2020, September 24). Decoding strategies for improving low-resource machine translation. MDPI. Retrieved April 2, 2023, from https://www.mdpi.com/2079-9292/9/10/1562

The prague dependency treebank: A three-level annotation scenario. (n.d.). Retrieved May 2, 2023, from https://www.researchgate.net/publication/251347641_The_Prague_Dependency_Treebank_A_Three-Level_Annotation_Scenario

Ranathunga, S., Lee, A. E., Skenduli, P. M., Shekhar, R., Mehreen, A., Kaur, R. (2021, June). Neural Machine Translation for Low-Resource Languages: A Survey. Retrieved April 4, 2023, from https://arxiv.org/pdf/2106.15115.pdf

Richards, L. (2019, October 24). Yandex: Beating google in Europe's biggest internet market. Search Engine Watch. Retrieved April 17, 2023, from https://www.searchenginewatch.com/2018/10/24/yandex-beating-google-in-europes-biggest-internet-market/

Hutchins, J. (n.d.). Machine translation over fifty years - ACL anthology. Retrieved May 1, 2023, from https://aclanthology.org/www.mt-archive.info/00/HEL-2001-Hutchins.pdf

Wang, L., &amp; Zhai, M. (n.d.). Deep Learning Enabled Semantic Communication Systems | IEEE journals ... Neural Machine Translation Strategies for Generating Honorific-style Korean. Retrieved April 17, 2023, from https://ieeexplore.ieee.org/abstract/document/9037681

Japanese-English machine translation of recipe texts - ACL anthology. Aclanthology. (n.d.). Retrieved April 12, 2023, from https://aclanthology.org/W16-4603.pdf

Google. (n.d.). Evaluating models &nbsp;|&nbsp; automl translation documentation &nbsp;|&nbsp; google cloud. Google. Retrieved May 1, 2023, from https://cloud.google.com/translate/automl/docs/evaluate