Extension Multi30K: Multimodal Dataset for Integrated Vision and Language Research in Ukrainian

Nataliia Saichyshyna, Daniil Maksymenko, Oleksii Turuta, Andriy Yerokhin, Olena Turuta and Andrii Babii
Kharkiv National University of Radio Electronics / Nauky Ave. 14, Kharkiv, Ukraine
{nataliia.saichyshyna, daniil.maksymenko, oleksii.turuta,
andriy.yerokhin, olena.turuta, andrii.babii}@nure.ua

Abstract

We share the results of the project within the well-known Multi30k dataset dedicated to improving machine translation of text from English into Ukrainian. The main task was to manually prepare the dataset and improve the translation of texts. The importance of collecting such datasets for low-resource languages for improving the quality of machine translation has been discussed. We also studied the features of translations of words and sentences with ambiguous meanings.

The collection of multimodal datasets is essential for natural language processing tasks because it allows the development of more complex and comprehensive machine learning models that can understand and analyze different types of data. These models can learn from a variety of data types, including images, text, and audio, for more accurate and meaningful results.

1 Introduction

Creating and processing high-quality datasets for such low-resource languages as Ukrainian is incredibly important for solving machine learning tasks. The task of machine translation, like other tasks, requires large amounts of data to effectively learn and understand the nuances and complexities of language. However, for low-resource languages, the available data may be limited, which can make it difficult to develop accurate and efficient models. Datasets directly affect the performance of machine learning models. This can lead to higher accuracy and better generalization for tasks such as language translation, speech recognition, and natural language processing.

A multimodal dataset refers to a dataset that consists of various data types, each representing a different modality. These modalities can include images, text, audio, and other types of data, each with its own unique meaning. By including multiple domains, a dataset can collect a wider range of information, allowing the development of more complex machine learning models. For our task we decided to choose Multi30k (Elliott et al., 2016) dataset which consists of two modalities: images and their descriptions.

Multi30K is a modification of the Flickr30K dataset (Young et al., 2014) with 31,014 German translations of English annotations and more than 150,000 collected annotations in German. This dataset was edited by professional translators, and one picture corresponds to one annotation in English and German, which is the reason for choosing this dataset for further adaptation into Ukrainian.

The problem under consideration consists of three parts:

- **Task 1: Machine translation** involves translating a source language description of an image into a target language. The training data is made up of pairs of sentences in different languages. We published some results in our previous articles (Maksymenko et al., 2022) covering this topic. Here we want to extend some explanations and conclusions.
- **Task 2: Multilingual multimodal semantic search** is a task with great demand considering how much unstructured multimodal data is stored nowadays. We need methods to search it quickly not only for English but for other even low-resource languages. We wanted to check some available models with the support of Ukrainian language using samples from our translated version of Multi30k.
- **Task 3: Usage of multilingual text embedding models to measure translation quality** which should in theory allow us to check model performance without any target language ground truth text. Some hard cases like phrases with a figurative sense should be considered to either prove or disprove the efficiency of this approach.

Proceedings of the Second Ukrainian Natural Language Processing Workshop (UNLP), pages 54–61
May 5, 2023 ©2023 Association for Computational Linguistics

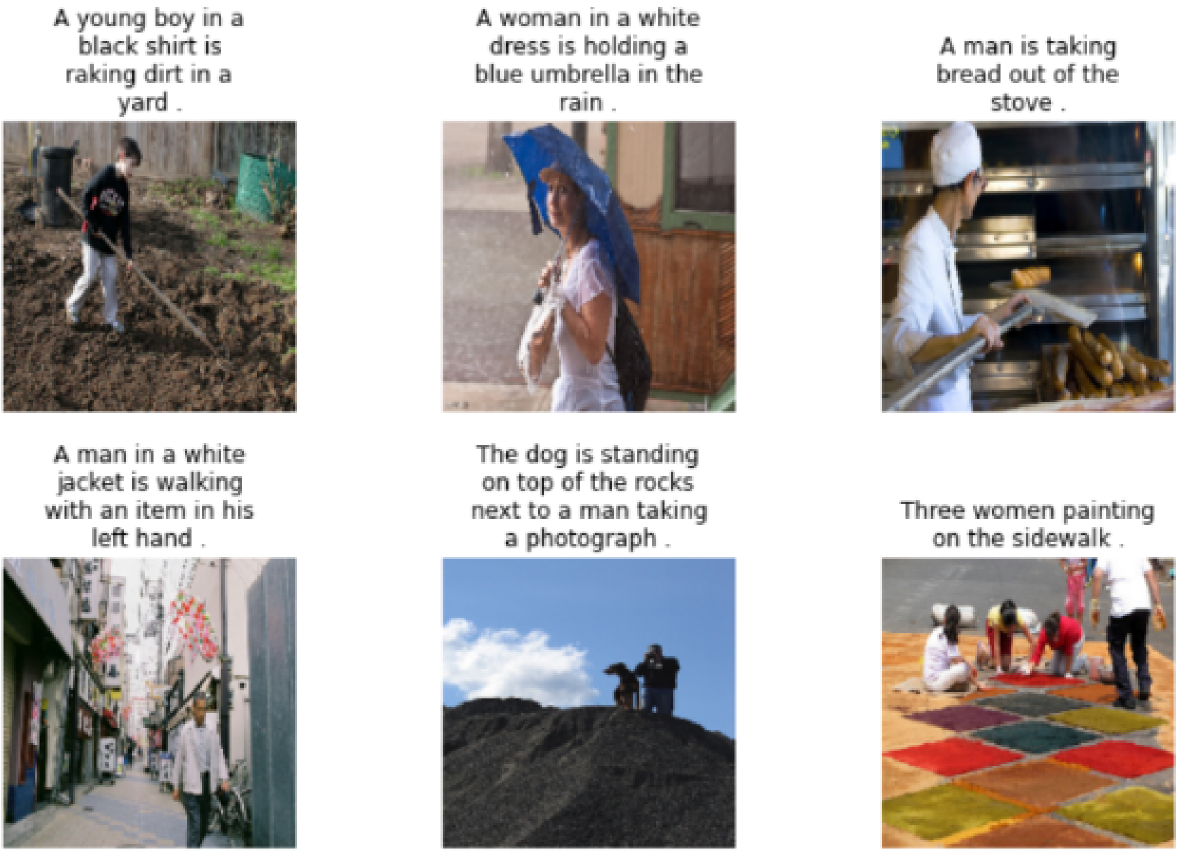

A young boy in a black shirt is raking dirt in a yard .
A woman in a white dress is holding a blue umbrella in the rain .
A man is taking bread out of the stove .
A man in a white jacket is walking with an item in his left hand .
The dog is standing on top of the rocks next to a man taking a photograph .
Three women painting on the sidewalk .

Figure 1: Annotations in English from Multi30k dataset

The first and main step of this research was to collect datasets for low-resource languages such as Ukrainian. We provided the necessary data to develop accurate and efficient machine-learning models. This may include datasets for tasks such as language translation, image captioning, text generation (Erdem et al., 2022), visual Q&A, sentiment analysis, and others.

2 Datasets and Tasks

The main dataset used for the tasks described above is the Multi30K dataset, which includes 31,000 images originally described in English. The presented dataset also includes translations created by professional German translators.

In the next iterations, this dataset was also translated into French (Elliott et al., 2017), Czech (Barrault et al., 2018) and Turkish (Citamak et al., 2020). Dataset overview is presented on Figure 1.

The dataset can be used to boost performance of some existing multilingual multimodal models for various machine learning tasks, such as multimodal machine translation, image captioning, image semantic search, cross-lingual transfer learning, and multilingual text summarization etc.

As a result, we managed to process 31,014 sentences for Ukrainian and English, and the number of words that are in this dataset was also counted. We prepared a Ukrainian version of Multi30k dataset with the following features. Comparison of the number of tokens for the languages from the original article (English and German) and Ukrainian can be seen in Table 1.

This number for the English language exceeds the given number for the Ukrainian language, due to its linguistic features, for example, there are no articles in Ukrainian.

Descriptions	Sentences	Tokens
English	31 014	357 172
German	31 014	333 833
Ukrainian	31 014	276 520

Table 1: Number of tokens in the dataset

3 Dataset preparation process

The first step was to load the selected dataset and conduct an initial inspection, determine the columns and data type, image format, count the number of sentences, words and images in order to select a further strategy for its processing.

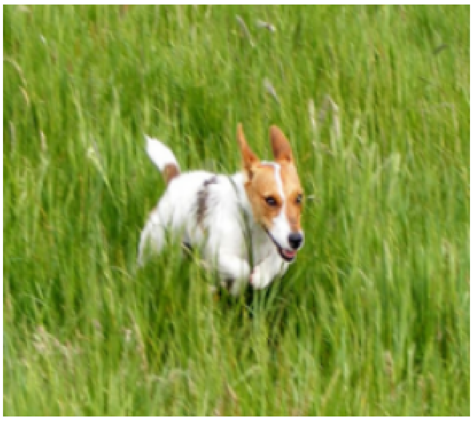
At the next stage, we performed the translation of English descriptions of images into Ukrainian using Google Cloud Translator in order to provide a further basis for manual verification and correction of texts.

An example of the annotation we received after translation with the help of Google translator can be seen in Figure 2. This example clearly shows that the resulting translation is not accurate and correct in this case. Thus, here is an adjective that is incorrect in meaning and an incorrect declension, since the word dog is masculine in Ukrainian.

Further this translation was the basis on which our team, which was engaged in corrections, relied. Our team consisted of 8 people: students and teachers of our university. For this work, an English text, an image and a Ukrainian text translated by Google Cloud Translator were provided.

It is important to note that in the process of correcting the text, the person who did it had access to both the image and the pictures. Ukrainian translation turned out to be dependent on these two sources, sometimes the picture helped to recognize what exactly was meant by the English description.

During preparation for translations, the data set was cleaned of incorrect characters and punctuation in order to be able to be used for training.



en: A white and tan dog runs through the tall green grass
uk: Біло-засмагла собака біжить крізь високу зелену траву

Figure 2: Sample from dataset

Cosine similarity value	Initial text	Manually translated text
0.9	1516	1546
0.8	3700	3763
0.7	4616	4675
0.6	4912	4919
0.5	4997	4977
0.4	4999	4999
0.3	5000	5000
0.2	5000	5000
0.1	5000	5000

Table 2: Cosine similarity count

The dataset is able on public repository https://huggingface.co/datasets/turuta/Multi30k-uk and https://github.com/researchlabs/multi30k-uk.

It is worth noting that the Google Cloud Translator translated simple sentences (which contain simple actions like "walk", "look", referring to a certain person) correctly and without comments. However, when faced with complex sentences and atypical actions, manual correction is required. Therefore, about 51% of the proposals were manually corrected.

4 Cosine similarity

The data obtained after translation were analyzed. We decided to calculate the cosine similarity using a multilingual model distiluse-base-multilingual-cased-v2 (Reimers and Gurevych, 2019) for the original translation using Google Cloud Translator and for the translation obtained after manual correction.

We got a high value of cosine similarity for all sentences for both languages. As a result, 4997 sentences have a high value, and 3 have a low value. Here, values above 0.4 are taken into account, which is considered sufficient for a general understanding of the meaning of a sentence or phrase.

Table 2 shows, using the example of 5000 records, the value of cosine similarity using the model described above.

The columns "Initial text" and "Manually translated text" indicate the number of sentences that exceed the corresponding cosine similarity value. Thus, as a result of our translation adjustment, an additional 30 values went out of range of 0.9, 63

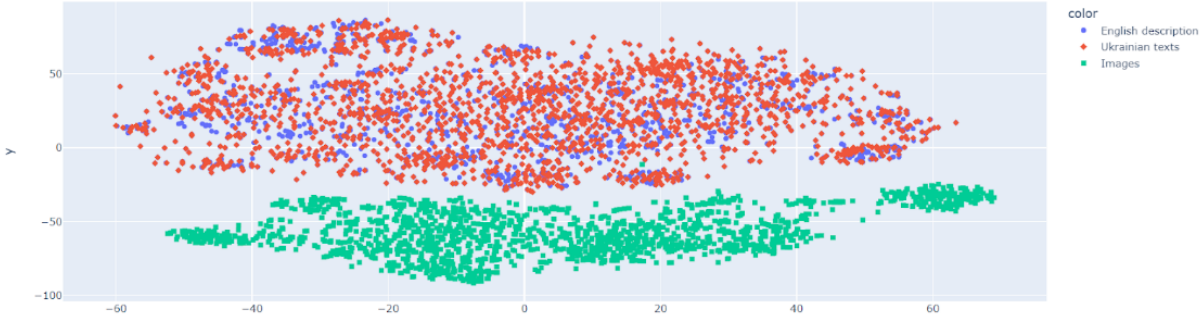


Figure 3: Texts and images embeddings projections

values went out of range 0.8, and so on.

This reflects the effectiveness of the work done on manual verification of the selected data.

5 Results

5.1 Machine translation

We published more detailed results of the machine translation task in our previous articles (Maksymenko et al., 2022). Those experiments involved not only obtained Multi30k translations, but also some other datasets, which we gathered like Ukrainian laws, scientific articles abstracts, programming documentation.

Machine translation was done using a fine-tuned MarianMT model, both on separate datasets and on all of them combined. We used Huggingface implementation, which is based on BART interface. This model was trained as a part of Helsinki NLP (Tiedemann and Thottingal, 2020) project with OPUS datasets.

Multi30k without any additional set did not drastically improve the performance of the MarianMT model (Junczys-Dowmunt et al., 2018), however it was able to improve generalization of previously tuned model as it provided some examples of new words and phrases, which were absent in the present checkpoint. Multi30k descriptions do not contain any domain specific words, the structure of sentences is easy to understand and capture. Such an effect was expected from it.

We used TATOEBA dataset with 5,000 texts to validate trained model and got METEOR score equal to 0.3810 and BERT F1 Score equal to 0.9232. METEOR was used as a classic token metric, which is more suitable for flexible languages as it uses synonyms matching and stemming to avoid extra penalties. BERT Score was used as an embedding metric to measure how well did the model capture meaning of the source text set.

5.2 Multilingual multimodal semantic search

We used a combination of CLIP and Siamese DistilBERT (Sanh et al., 2019) in our research as a multimodal semantic search model. “sentence-transformers/clip-ViT-B-32-multilingual-v1” model weights from Huggingface hub were used as an initial checkpoint for checked models. First of all, we tried to visualize embeddings of images, source English texts, and their manually fixed Ukrainian translations.

Models return vectors, which consist of 512 elements with values from -1 to 1. The first step here is to train a language embedding model, like original BERT, for a high-resource language like English. Then DistilBERT should try to replicate this embedding vector for translations of original texts to maximize cosine similarity between the same texts in different languages. The same process is applied to CLIP to replicate similar embedding space for corresponding images. So we encoded our images and texts in both languages and used TSNE to create 2D projections of original embedding vectors (figure 3). You can see how Ukrainian translations almost perfectly replicate the form of English text distribution, which proves that our fixed translations should be close to the original descriptions and can be further used for some real-world tasks. However, images fall into an absolutely different part of this embedding space. They try to replicate the form of those text clouds, but they are still far from texts and don’t really correspond to their descriptions judging by embeddings. Only 44.5% of images correspond mostly to their real description. This value is equal to 29.55% for Ukrainian versions.

Here are some examples of errors made by the semantic search model. We have the following Ukrainian image description: "Молодий бородатий чоловік у білій безрукавці сидить

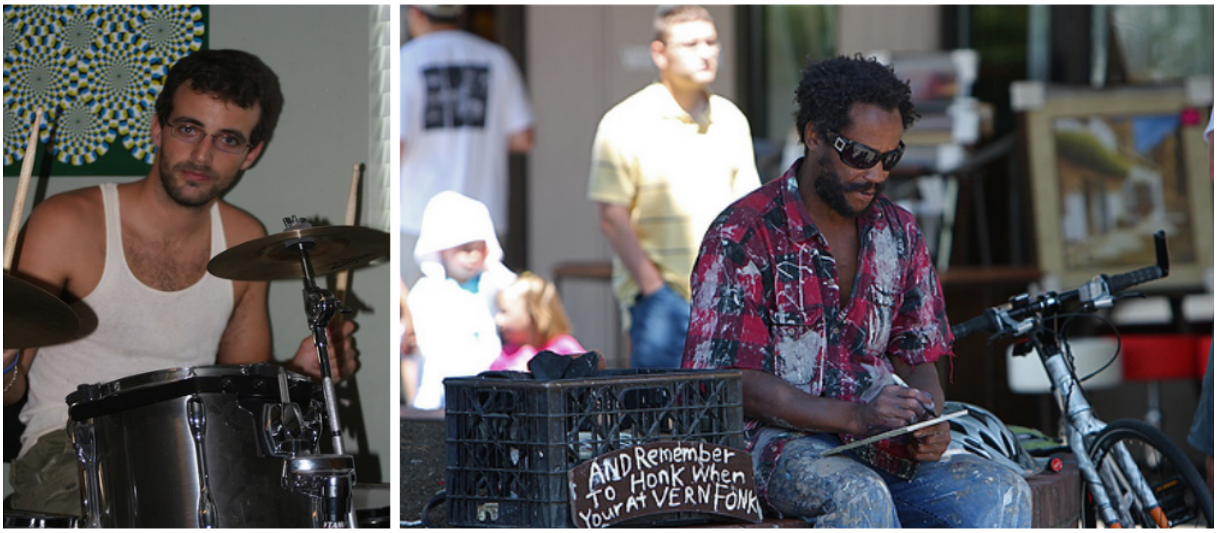


Figure 4: Original corresponding image and one proposed by model

за барабанною установкою" (Young bearded man wearing a white tank top sits behind a drum set.). The semantic search model returns a similar image, but with slightly different details. There is really a man with a beard who sits somewhere on figure 4, but he just draws something on his tablet and there are no drums.

Model finds subjects really well, as most errors we saw are related to the action or some environment details or background objects. We have some images where people swim by the river or some kind of lake using a canoe. These images can be distinguished by some small details like the number of people in the boat, the type of background (cave, rocks, some specific type of forest), description of the river. However, the algorithm usually catches only the most significant details like "people in a canoe". So it misses all those small details, like in a previous case with a bearded man.

So, we can not definitely recommend this model for some real-world tasks for Ukrainian language as it still makes some obvious mistakes, which can be fixed by further fine-tuning. That is where our proposed dataset can be useful, so we can try to drastically improve the performance of Ukrainian multimodal semantic search by using these combinations of images and their descriptions during further research.

5.3 Usage of multilingual text embedding models to measure translation quality

Every classic approach to measure the translation quality relies on some target language ground truth value. However, what if we need to check if translation is good to use and we do not have any previously checked sample? Modern multilingual language models can produce similar embeddings for the same text in multiple languages, so we can compare them in one shared space. We have shown it in the previous section of the article as Ukrainian texts distribute almost identically to English ones.

We calculated cosine similarity between English and Ukrainian embedding vectors to check how an external model (siamese DistilBERT in distiluse-base-multilingual-cased-v2 implementation) would score our fixed translations. On figure 5 is a histogram of cosine similarity scores distribution.

Most texts fall into 0.6 and higher bins, which is a really good result as it indicates that our translations capture original meaning. 98.38% of texts belong there. Such a result is a great achievement for this metric as it seems like it almost replicates human judgment. However, there are a few smaller beans, which are of interest to us. Let's start with the ones in the range [0.4, 0.6).

We checked texts which belong to these bins and mostly they consist of cases where an English phrase or word combination gets translated into a single Ukrainian word, which is also a rare and not commonly used word. Like for example phrase "horse shoes" gets translated into word "підковки", which is a correct translation, but it this word is not so common and can slightly misdirect the language model. Here is another similar example: English phrase "give high-fives" gets translated as "дає п'ять". Translation is correct and the phrase itself is similar to the English one, but the model gets confused a bit, because it does

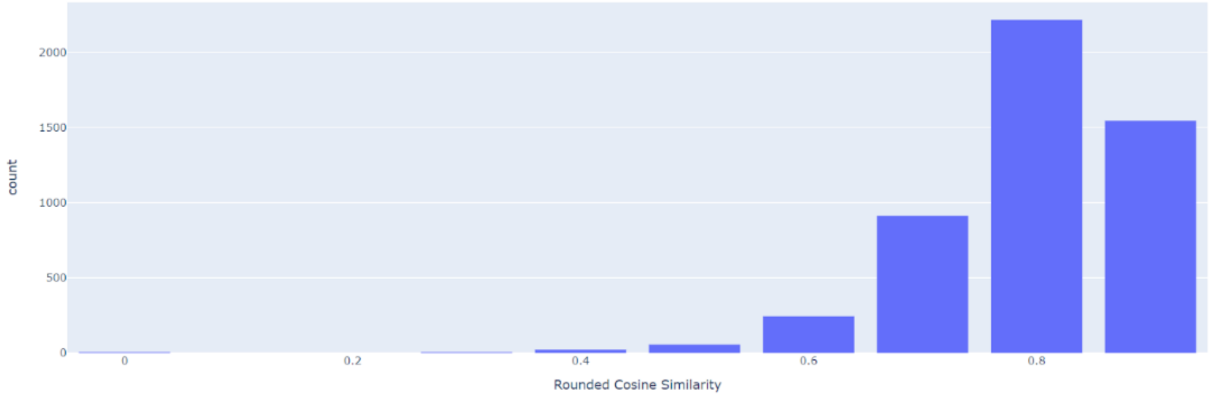


Figure 5: Histogram of cosine similarity scores distribution

not really understand the meaning of the phrase. That is an interesting case as it shows that even the correct translation of similar phrase in a figurative sense lowers the score.

Now let's move to some lower buckets in range [0, 0.4). There are only 3 texts in this buckets and all of them contain texts with some slangs or phrases with a figurative sense. For example a radio receiver was called "walkie talkie" in English description. Ukrainian version just used a word "рація". This text got a cosine similarity score 0.32 as model just was not able to capture this slang and connect it to Ukrainian analogue. Another example contains a rare word "волосінь", which stands for a fishing line. The model did not capture it as it probably did not encounter this word or some similar ones during the fit.

Also, we tried to do the same using a siamese DistilBERT aligned with image embeddings, which was used in task 2. We did not use images and just compared 2 texts. The results are drastically different as injection of visual information allowed neural network to better capture phrases like "high five" or "walkie talkey". It seems like an additional domain was able to give enough context for the network to compare these sentences in a more human-like way. Sentence which contained "walkie talkie" got 0.7456 score this time. There are no translations with a score lower than 0.5, if we measure the translation using this model. On figure 6 is the histogram we have built.

This area needs some further research, but from our tests and experiments it seems like such models can be used further to capture some figurative phrases or slangs in combination with some traditional metrics, like token-based ones. Usage of multidomain models to measure translation quality also has great potential as its results were much better than just text model. It fixed main problems, which we encountered in ordinary text model, but it definitely should be tested more before giving a recommendation to use it as a benchmark for machine translation.

We made some additional checks with some random phrases. English sources sound like "Murder will out" and "Keep the change". Here are Ukrainian translations: "Правди не сховаєш", "Здачі не треба". Only the textual model gave the following scores respectively: 0.2495 and 0.2599. Textual model tuned to resemble visual embeddings gave these scores: 0.9569 and 0.9497. Results are much better than we expected and outperform ones obtained from the only textual model. However, as we said before, the theory that visual embeddings were the main reason that boosted model performance still needs more proof and more research.

6 Conclusion

In conclusion, the importance of collecting high quality datasets for low-resource languages such as Ukrainian cannot be overestimated for machine learning tasks. An example of this was our project to improve the machine translation of text from English to Ukrainian by manually preparing the Multi30k dataset and examining translations of ambiguous words and sentences.

Collecting multimodal datasets that include different types of data such as images, text, and audio is especially important as they provide richer and more complex data for developing accurate and meaningful machine learning models. The results from our project demonstrate the impact of such datasets in improving the performance of machine

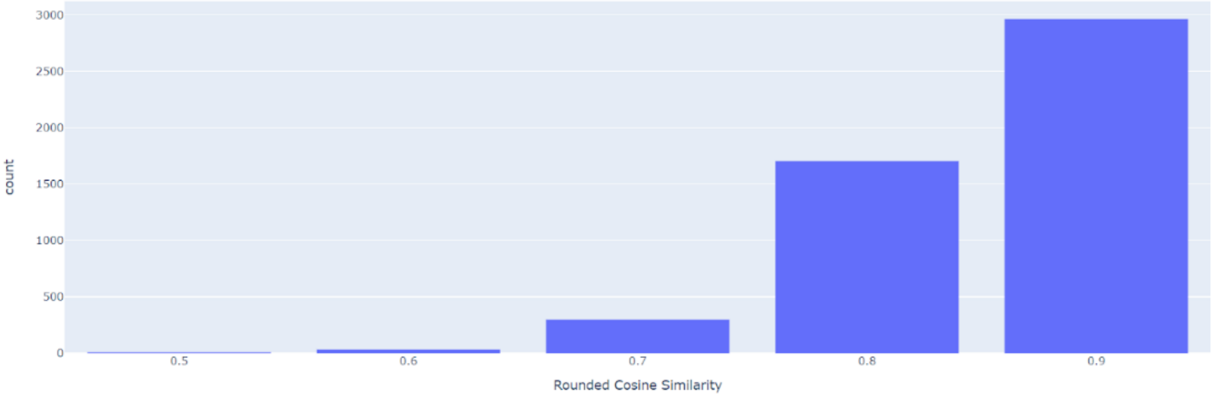


Figure 6: Histogram from a siamese DistilBERT aligned with CLIP image embeddings

learning models for tasks such as machine translation.

As a result, the creation and processing of such data sets will lead to a significant improvement in the solution of the problem of machine translation and many other tasks.

The project involved loading and validating a selected data set to determine the data type, image format, and word count. The data set was translated from English into Ukrainian using Google Translate, which served as the basis for manual verification and correction by a team of 8 people

As a further development we want to research siamese language models and cosine similarity of their embeddings even more to finally either prove or disprove that they can be used as benchmarks for machine translation. Also, we want to check how our gathered dataset will affect the performance of existing multimodal multilingual semantic search models by finetuning them using Ukrainian Multi30k. Another area for further research is to combine token metrics and text embeddings from a multilingual semantic search network to capture figurative meaning and some professional or just domain specific words and phrases.

Limitations

In the process of working on the study, we encountered a number of limitations and an unsuccessful experiment that did not give results. For example, different machine learning models sometimes showed different results, so it would be wise to explore more for our calculations. The images that are part of the considered datasets also require the necessary attention and refinement. We plan to integrate them more closely with textual information, thus improving the quality of the resulting machine translation. At some points in our study, we ran into a lack of computing power.

Ethics Statement

In creating this study, we are fully guided by generally accepted ethical principles towards the community of authors and organizers. We respect scientific developments and works and study them with interest for our further research and communication.

References

Loïc Barrault, Fethi Bougares, Lucia Specia, Chiraag Lala, Desmond Elliott, and Stella Frank. 2018. Findings of the third shared task on multimodal machine translation. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 304–323, Belgium, Brussels. Association for Computational Linguistics.

Begum Citamak, Ozan Caglayan, Menekse Kuyu, Erkut Erdem, Aykut Erdem, Pranava Madhyastha, and Lucia Specia. 2020. Msvd-turkish: A comprehensive multimodal dataset for integrated vision and language research in turkish.

Desmond Elliott, Stella Frank, Loïc Barrault, Fethi Bougares, and Lucia Specia. 2017. Findings of the second shared task on multimodal machine translation and multilingual image description. In *Proceedings of the Second Conference on Machine Translation*, pages 215–233, Copenhagen, Denmark. Association for Computational Linguistics.

Desmond Elliott, Stella Frank, Khalil Sima'an, and Lucia Specia. 2016. Multi30K: Multilingual English-German image descriptions. In *Proceedings of the 5th Workshop on Vision and Language*, pages 70–74, Berlin, Germany. Association for Computational Linguistics.

Erkut Erdem, Menekse Kuyu, Semih Yagcioglu, Anette Frank, Letitia Parcalabescu, Barbara Plank, Andrii Babii, Oleksii Turuta, Aykut Erdem, Iacer

Calixto, Elena Lloret, Elena-Simona Apostol, Ciprian-Octavian Truică, Branislava Šandrih, Sanda Martinčić-Ipšić, Gábor Berend, Albert Gatt, and Grăzina Korvel. 2022. Neural natural language generation: A survey on multilinguality, multimodality, controllability and learning. *J. Artif. Int. Res.*, 73.

Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins, and Alexandra Birch. 2018. Marian: Fast neural machine translation in C++. In *Proceedings of ACL 2018, System Demonstrations*, pages 116–121, Melbourne, Australia. Association for Computational Linguistics.

Daniil Maksymenko, Nataliia Saichyshyna, Oleksii Turuta, Olena Turuta, Andriy Yerokhin, and Andrii Babii. 2022. Improving the machine translation model in specific domains for the ukrainian language. In *2022 IEEE 17th International Conference on Computer Sciences and Information Technologies (CSIT)*, pages 123–129.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *ArXiv*, abs/1910.01108.

Jörg Tiedemann and Santhosh Thottingal. 2020. OPUS-MT – building open translation services for the world. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 479–480, Lisboa, Portugal. European Association for Machine Translation.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78.