BiVert: Bidirectional Vocabulary Evaluation using Relations for Machine Translation

Carinne Cherf Yuval Pinter
Department of Computer Science, Ben Gurion University
Beer Sheva, Israel
carinnecherf@gmail.com uvp@cs.bgu.ac.il

Abstract

Neural machine translation (NMT) has progressed rapidly in the past few years, promising improvements and quality translations for different languages. Evaluation of this task is crucial to determine the quality of the translation. Overall, insufficient emphasis is placed on the actual sense of the translation in traditional methods. We propose a bidirectional semantic-based evaluation method designed to assess the sense distance of the translation from the source text. This approach employs the comprehensive multilingual encyclopedic dictionary BabelNet. Through the calculation of the semantic distance between the source and its back translation of the output, our method introduces a quantifiable approach that empowers sentence comparison on the same linguistic level. Factual analysis shows a strong correlation between the average evaluation scores generated by our method and the human assessments across various machine translation systems for English-German language pair. Finally, our method proposes a new multilingual approach to rank MT systems without the need for parallel corpora.

Keywords: Machine Translation, Graph Sense, Multilingual, Quality Estimation

1. Introduction

Automatic evaluation of machine translation (MT) is crucial to determine the quality and performance of translation systems. It is an important step in the development and improvement of MT models, as it sheds light on the models' strengths and weaknesses. As the demand expands for high-quality translations, spanning a variety of languages, also the need for efficient and reliable evaluation techniques grows rapidly. The major goal of these evaluation methods is to approximate the semantic similarity between the target text and some generated text. Standard techniques rely on comparing the machine translation's output with the desired true reference. Common methods such as BLEU (Papineni et al., 2002) and ROUGE (Lin, 2004) rate the translation based on n-gram intersections. Many of these methods are effective at capturing aspects of text similarity, but fall short on the actual meaning difference. Advanced techniques using word-embedding based approaches like BERTScore (Zhang et al., 2020) have marked significant progress in assessing translation quality. With this in mind, a significant shift has occurred in recent years towards the need for accurate reference-less evaluation metrics.

Our goal is to introduce a different strategy for machine translation evaluation, one that does not require an aligned parallel test set. BiVert is a simple bidirectional and self-supervised method constructed from a multilingual encyclopedia. In essence, BiVert evaluates a translation between the source sentence s and the target sentence t by scoring the semantic similarity between the s and its back-translated sentence s', as illustrated in Figure 1. We refer to the former translation as the *direct* action and the latter as the *back* action. For the first step, we generate the back sentence s' using a standard machine translation system, which we commonly label as a state-of-the-art MT system. This way we form a single-language platform for comparing the meanings between the original text and the back translated text. With the help of contextualized embeddings, extracted by the model to be evaluated, we pair the words between the sentences and compare them. At this point, we can estimate the semantic distance between s and s' making use of the word pairs, resulting in an indirect estimation of the direct translation quality. We train BiVert features on the WMT Metrics Task 2021 dataset, and experiment on the WMT Metrics Task 2022 dataset, comparing our average results to existing methods (Freitag et al., 2022). Our experiments show that BiVert obtains strong correlation with the human scores for the English–German language pair, with promising potential on Chinese to English and English to Russian.

2. Related Work

Numerous methods measure the resemblance between generated text and human text such as classic n-grams techniques and word embeddings strategies, some of which rely on a predefined reference. Previous research findings (Novikova et al., 2017) cast doubt on the alignment between predicted outcomes and human judgments

LREC-COLING 2024, pages 2410–2420

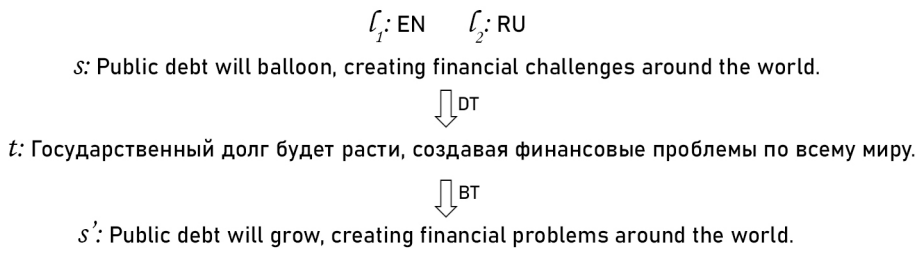

Figure 1: Example of a direct translation from English to Russian using the system we wish to evaluate, and its back-translation using a state-of-the-art translation system suitable for BiVert.

for known methods. Recent advancements in the field of quality estimation have introduced techniques that offer a more accessible solution as they do not require collecting human references or obtaining parallel alignments. Moreover, Previous research (Dyvik, 1998) introduced a knowledge discovery technique known as Semantic Mirroring, which relies on identifying semantic relationships between words in a source language and their counterparts in a target language. They emphasize that by mirroring source words and target words back and forth they are able to provide insights into cross-lingual semantic relations.

Reference-based measures assess the output of an MT system by comparing it to a limited set of reference text samples. Traditional methods, such as BLEU and ROUGE which search for matching n-grams, primarily aim to capture prominent similarities between the generated text and the true reference. To compare generated data against human text, Self-BLEU (Zhu et al., 2018) treats one sentence as a hypothesis and those remaining as references. It calculates the BLEU score for each generated sentence in comparison to the collection, as the average BLEU score is then defined as the document's Self-BLEU mark. Moreover, BERTScore (Zhang et al., 2020), an advanced evaluation technique, measures the similarity of two sentences as the sum of their cosine similarities between their pre-trained BERT contextual embeddings (Devlin et al., 2019). Although contextual embeddings are trained to capture long-range relationships effectively, they can still struggle with distinguishing between similar senses or meanings. BERTScore is affected by the antonymy problem (Saadany and Orasan, 2021), where antonyms usually have similar contextual values and are closer in vector space. As a result, a translation of one word to its exact opposite is not sufficiently captured as erroneous by the metric. Another issue is that BERTScore struggles to distinguish between the mistranslation of a critical word that could significantly alter the intended meaning. Occasionally, a word may have multiple interpretations depending on the context, whereas BERTScore may fail to capture the error that affects the actual sentence intention. However, MoverScore (Zhao et al., 2019) takes into account the Euclidean distances between the vector representations and tries to find the minimum effort to transform between both texts. This captures more effectively the degree of resemblance between the texts. An alternative approach, MAUVE (Pillutla et al., 2021), compares characteristics of the source and the target distributions using the Kullback-Leibler (KL) method. It creates a divergence curve that represents two types of errors: false positives (unlikely text) and false negatives (missing plausible text). By analyzing this curve and calculating the area under it, MAUVE provides a scalar value that quantifies the overall gap between both texts. We note that although evaluation of individual sentence-level texts against references is beneficial, corpus-based metrics provide a more comprehensive and meaningful assessment of machine translation systems.

Quality estimation (QE) for machine translation, also known as reference-less evaluation, presents an approach for assessing text, in particular relevant for authentic text, such as social media. Moreover, it can also drastically decrease the cost of developing effective machine translation systems. These methods value the quality of the translation without any information about aligned referenced text. For instance, CometKiwi (Rei et al., 2022) implements this manner by combining qualities of two frameworks, Comet (Rei et al., 2020) for the training process and OpenKiwi (Kepler et al., 2019) for prediction. Their architecture feeds a trained network with both the source and target sentence resulting a score for the task, thus not requiring a reference text for the evaluation. DeepQuest (Ive et al., 2018), a sophisticated neural-based sentence-level architecture for document-level quality estimation, achieves impressive performance compared to previous methods. The results of quality estimation can be either represented by standard metrics like F-measure or

by determining the correlation between the evaluation score and the state-of-the-art gold standard. In contrast, BiVert does not require a neural network training, as it is based on a multilingual connected sense-based network of words and only requires tuning of seven parameters.

Semantic Graphs provide a structured illustration of relationships between associated objects. These graphs represent a network of words and senses, connected based on a relationship between both sides. Word-sense disambiguation (WSD), a task of identifying the accurate sense of a word within a context, can be approached through graph-based algorithms. Many words have multiple senses, and the challenge of determining the correct sense of a word often relies on the surrounding context. In WSD, given a document represented as a sequence of words $W = \{w_1, w_2, ..., w_n\}$, the goal is to establish connections with the correct sense(s) for $w_i \in W$. Specifically, the objective is to find a mapping f from the searched words to their senses, such that $f(w_i; W) \in S(w_i)$, where $S(w_i)$ is the set of senses for the word $w_i \in W$. By forming semantic graphs assembled from words as nodes connected by edges representing semantic relationships, graph-based algorithms can resolve the obscure puzzle of connections between words. WordNet (Miller, 1992) is a prime example of semantic graphs, being a comprehensive lexical database that bridges semantic relationships among different concepts. Various approaches such as MetaGraph2Vec (Zhang et al., 2018) and Edge2vec (Wang et al., 2020) benefit from sense networks for learning embeddings.

3. BiVert: A Semantic Evaluation

BiVert, or **Bi**directional **V**ocabulary **E**valuation using **R**elations for machine **T**ranslation, is an evaluation method for multilingual translation that concentrates on identifying the actual senses of the source sentence s and the target sentence t. This is achieved through comparing the source sentence s and its back-translated sentence s', both of whom share a common language l_1, allowing to calculate the semantic distance between them using only monolingual resources. The first step is to generate the back-translated sentence s' using a state-of-the-art translation system. An alternative use case could be to rely on the evaluated system itself for the back-translation. Any translation system of adequate quality can be employed for this task. This is followed by matching word pairs between both sentences using a pairing algorithm on the words embeddings. The words embeddings might be split by sub-words and need to be aggregated. We then identify the relation of each pair and assign a score accordingly (see section 3.3). We sum the scores achieved by the word pairs for each category. Finally, the assessment of the translation's quality is accomplished by aggregating the summed scores of all categories using trained weights for each relation type, which are tuned for each language pair, as detailed in section 3.4.

3.1. BabelNet

One of BiVert's objectives is to identify the correct sense connection between a pair of words. To this end, we make use of BabelNet (Navigli and Ponzetto, 2012), a consistently updated multilingual encyclopedic dictionary that connects named entities in a very large network of semantic relations. BabelNet follows the WordNet model, consisting of *synsets*, each representing a set of synonyms which encode the same concept. Synsets are linked to each other using semantic relation edges of types such as *hypernym*, *hyponym*, and *antonym*. BabelNet is unique in providing extensive coverage of words and their meanings across multiple languages. Moreover, BabelNet aggregates data from a variety of resources: Wikipedia, Wiktionary, Wikidata, VerbAtlas, WordNet, GeoNames and OmegaWiki.

3.2. Word Alignment

Following the action of back-translating the target sentence t into s', we proceed to align the words between s and s', thereby generating pairs of matching words as demonstrated in Figure 2. To ensure accurate alignment of word pairs, we calculate the cosine similarity score between the embeddings corresponding to the aligned elements in both sentences. We match element pairs using the linear sum assignment problem (LSAP), implemented using a modified Jonker-Volgenant algorithm (Crouse, 2016). LSAP is equivalent to minimum weight matching problem in bipartite graphs. The objective is to pair each row with a distinct column in a manner that minimizes the sum of the corresponding entries. In other words, we want to select n tokens (rows) from s and find their corresponding matches (columns) in s' while maximizing the sum of cosine similarities. Since the systems we evaluate on and with employ subword token embeddings, we require a way for pooling multiple tokens that correspond to a single word when such a segmentation occurs. In one approach, the overall sentence-level alignment is performed over the token sequence, obviating the need for word-level aggregation. Other methods encourage subword pooling as a preliminary step for word-level

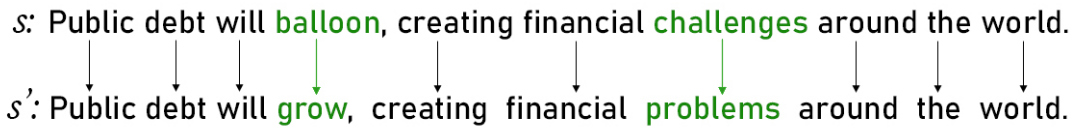

Figure 2: An example of final words alignment using the linear sum assignment problem algorithm.

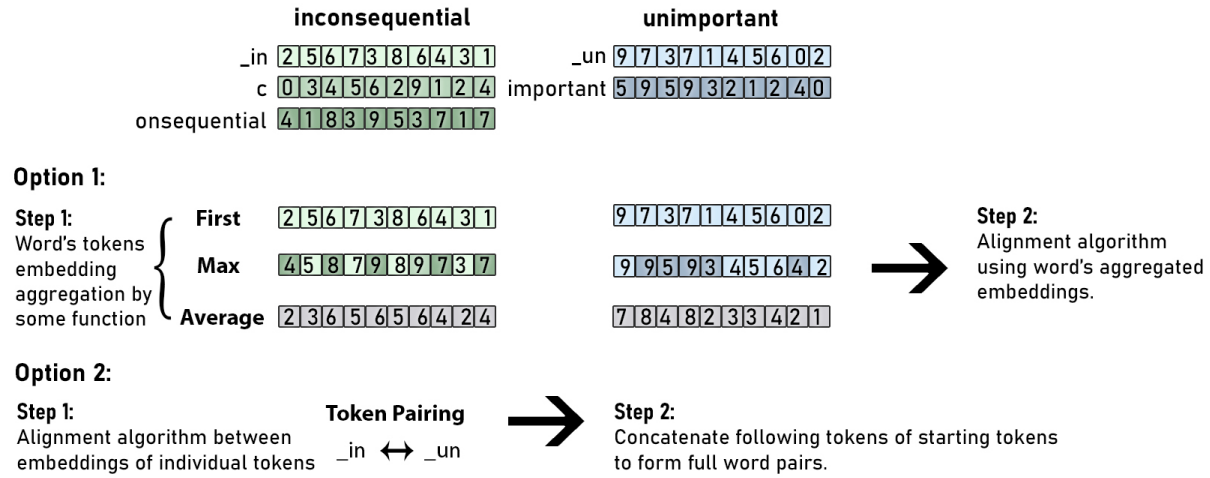

Figure 3: Example of words *inconsequential* and *unimportant* with illustrative embedding values, demonstrating different subword pooling strategies for word alignment. The word alignment algorithm calculates the cosine similarity between the embeddings representing the words chosen via option 1 or 2.

operations (Ács et al., 2021) For instance, the maximum element-wise approach aggregates the tokens embeddings into single word representations by selecting the maximum value at each position of the embedding. Another strategy settles on the first token for the word representation. We chose to operate over the token level, selecting word alignments based on tokens they contain as "representatives" for the full word: as soon as a token inside a word is aligned, the word in its entirety is paired with the corresponding token's word from the other sequence. Follow the example in Figure 3.

3.3. Word Pair Relations

After pairing the words from the source and back-translated text we define each pair's relationship. We identify the following categories of possible word relations: *Same*, *Extra*, *Missing*, *Stopwords*, *Inflection*, *Derivation*, and *Sense*. Each match receives a value according to its type as described below.

1. **Same**: This category refers to word pairs in which both words are identical. Since this pair does not cause any variation between the sentences, it is not taken into account within the final score decision. Their presence does not affect the evaluation hence the score assigned is zero.

2. **Extra**: The extra category suggests a word has been added to the translation sentence and has no match in the source counterpart. This relation costs $1/\text{len}(s)$, to account for its relative a-priori weight in the sentence.

3. **Missing**: Missing word pair indicates a word from the source sentence s lacks a parallel match in the back-translated sentence s'. A missing pair costs $1/\text{len}(s)$ as well.

4. **Stop words**: Non-identical paired words which are both contained in a list of language-specific stopwords are treated as one half of a replacement operation and cost $1/len(s)$, since they are often interchangeable (e.g., 'at' $\leftrightarrow$ 'on').

5. **Inflection**: Inflection refers to a process of word formation to signal differences in grammatical attributes like tense, person, number, and gender. Two words are categorized as an inflection if their lemmas are identical. We

weigh this relation by calculating the cosine similarity between both words' embeddings.

6. **Derivation**: Derivation is the process of varying a word's part of speech while retaining its core semantic content. For instance, "happy" and "happiness" have a derivation relationship. We assess these pairs by computing their cosine similarity.

7. **Sense**: Sense-related words are different words which have been chosen by the alignment algorithm due to their close embedding distance. These words may be synonyms, hypernyms, or antonyms. We aim to grade the actual distance of their intentional sense in the given context, using the multilingual encyclopedia BabelNet. For this issue we assemble a semantic graph described in section 3.3.1.

3.3.1. Sense Relation Type

The BiVert evaluation method is focused on finding the differences between words' true senses in order to correctly estimate the direct translation. For each word pair found to exemplify the *sense* relation, we form a semantic subgraph using BabelNet. To construct the graph we pass both words, $x \in s$ and $y \in s'$, through a lemmatizer, if available in language l_1, and extract their senses. The graph now has two roots, x and y, and nodes connected to each root representing their senses. We locate the shortest path from root x and root y using Dijkstra's algorithm (Dijkstra, 1959). As long as a path between the two roots has not been found, we continue expanding the graph by extracting each sense's hypernyms and iteratively searching for a connected path, as illustrated in Figure 4. After marking the route, we score it as described in the remainder of the section. If a path is not found according to a pre-specified max search depth threshold, we revert to scoring the relation as the cosine similarity between the roots. We note that BabelNet's resources restrict us to scoring relations between nouns and between verbs.

The **sense score** for a matching pair is calculated using the semantic graph G, constructed from nodes V representing the root words and their senses, and edges E consistent of the relations between the nodes. Each edge receives a score by the type of lexical connection it represents according to research done by Michael Sussna (Sussna, 1993). Each edge weight consists of type weights defined by the relation of the words (1). The type weight (2) is defined by minimum and maximum values chosen for word relations of types hypernymy, hyponymy, holonymy, and meronymy. In practice, all of these relations have weights ranging from 1 to 2. In contrast, the weight used for all antonymy arches is constantly valued at 2.5. The edge weight is then averaged by the two inverse weights and divided by the depth of the edge within the graph. Together, the weight between node a and b is defined as:

$$w(a,b) = \frac{w\left(a \rightarrow_r b\right) + w\left(b \rightarrow_{r^{-1}} a\right)}{2d}, \quad (1)$$

$$w\left(x \rightarrow_r y\right) = \max_r - \frac{\max_r - \min_r}{n_r(X)}, \quad (2)$$

where $\rightarrow_r$ is a relation of type r and r^{-1} is its inverse; d is the depth of the deeper of the two nodes; $\max_r$ and $\min_r$ are the maximum and minimum weights possible for a relation of type r; and $n_r(X)$ is the number of relations of type r leaving node X.

The final graph score $S(a,b)$ from root a to b is given by the normalized sum of the edge weights along the path between them:

$$S(a,b) = 2 \times (0.5 - \frac{1}{\sum_{e \in P(a \rightsquigarrow b)} w(e)}). \quad (3)$$

3.4. Final Score

Training and testing evaluation strategies occasionally requires human interference with the desired output. Our procedure does not require any human resources as our method is completely automatic, comparing the source sentence with the generated backtranslated sentence. The score of each relation pair is summed by relation categories. The final score of BiVert is a trained combination of all relation types into a final score. We use gradient descent to train our method in order to achieve optimal predictions for each language pair.

4. Experiments

In this section we describe the experiments conducted for finding the optimal BiVert configurations. For each language we learn the optimized values for BiVert features as resulted in Table 1. For our self-supervised method, we start by applying a machine translation system on the source sentences to generate the back-translation. We use a state-of-the-art translation model, MarianNMT (Junczys-Dowmunt et al., 2018), for this task. This model is based on the Marian open-source tool for training and serving neural machine translation. It was trained on multiple sources from parallel data collected at OPUS (Tiedemann and Nygaard, 2004). The model used the SentencePiece Tokenizer, an unsupervised text tokenizer, along with pre-trained embeddings from Word2Vec vectors (Kudo and Richardson, 2018). Next, we make sure to apply

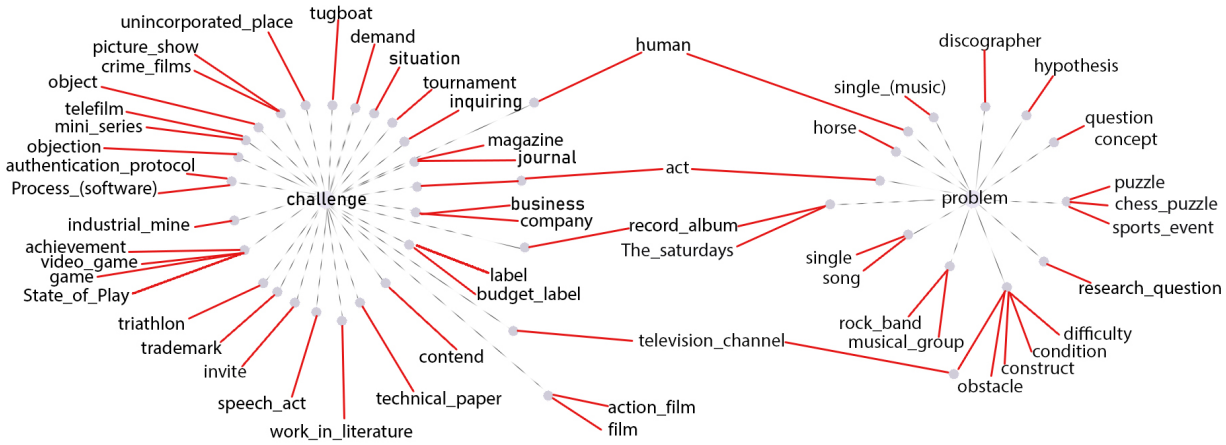

Figure 4: Fragment of a semantic graph between the two words *challenge* and *problem*. The hatched grey edges connect roots to their senses, and the red edges represent hypernym relations between the nodes contents.

	Extra	Missing	Stopword	Inflection	Derivation	Sense
English-German	0.121	0.134	0.188	0.101	0.092	0.360
English-Russian	0.112	0.164	0.196	0.087	0.063	0.375
Chinese-English	0.172	0.203	0.126	0.000	0.000	0.497

Table 1: Feature importance scores learned by a Gradient Boosting Regression model for BiVert language pairs.

Language pair Human Translation Included	**eng-deu** yes	**eng-deu** no	**eng-rus** no	**zho-eng** yes	**zho-eng** no
BERTScore	0.338	0.428	0.811	0.843	0.924
Cross-QE	**0.643**	0.661	**0.806**	**0.817**	0.870
COMETKiwi	0.592	**0.674**	0.763	0.795	0.866
MS-COMET-QE-22	0.417	0.539	0.672	0.799	**0.897**
UniTE-src	0.509	0.509	0.779	0.791	0.874
MATESE-QE	0.363	0.337	0.637	0.741	0.767
COMET-QE	0.480	0.502	0.468	0.544	0.569
KG-BERTScore	0.369	0.400	0.612	0.617	0.743
HWTSC-TLM	0.311	0.428	0.597	0.368	0.460
HWTSC-Teacher-Sim	0.290	0.385	0.675	0.294	0.356
BiVert	**0.694**	**0.703**	0.657	0.376	0.239

Table 2: System-level Pearson correlation between human scores and BiVert scores, compared to other evaluation metrics. "Human Translation Included" refers to refB system which may be included or excluded from the correlation calculation. See system-level scores in Table 4. Highest reference-free scores are **bolded**.

a pre-processing language-specified routine on all data for optimal results. For Chinese, we keep only Chinese characters in the text. For English, we lowercase the sentence and expand contractions, for example *don't* → *do not*. After cleaning the text we combine embeddings using the pair-wise token technique. The next step is aligning the words between the source sentence s and its back-translated counterpart s'. For this process we calculate the score between each word pair (w_1, w_2) as $similarity = cos\_sim(w_1, w_2)$, using their embedding representations summed before.

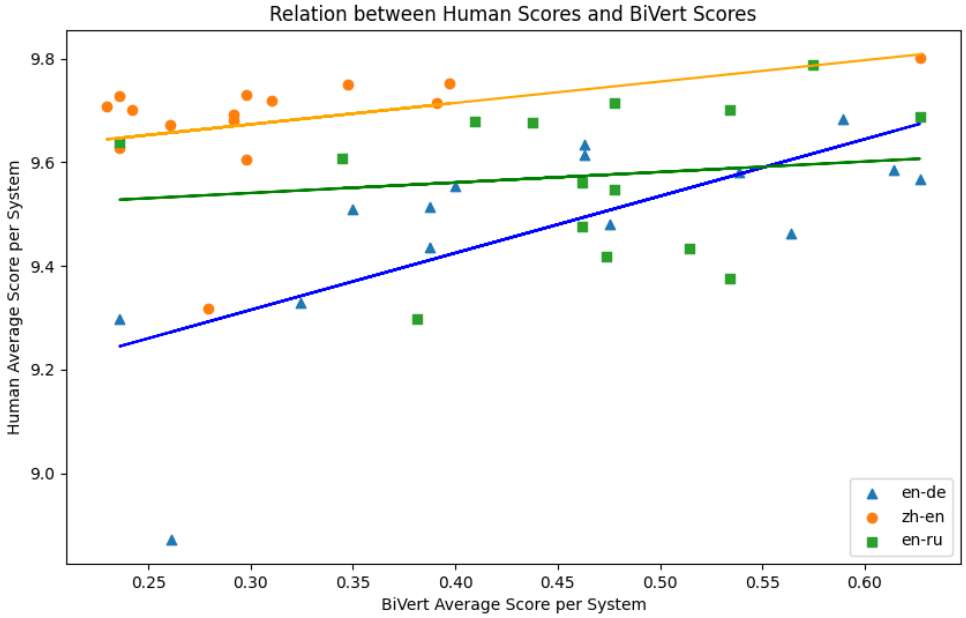

Figure 5: A comparison of average human scores and average BiVert scores for each language pair on all translation systems.

Since the algorithm searches for the minimum total cost we update each value to be $similarity = 1 - similarity$. For each aligned pair, we define the match relation and sum its value according to the category cost definition. Specifically for the *Sense* relation we apply Lemmatization[1] (currently only in English) prior looking up the words on BabelNet for accurate results. We restricted the sense connection edges to hypernym type only, and limited the graph depth to seven levels. We operate on the most recent version 5.2 of BabelNet as our multilingual encyclopedia resource for extracting words' senses. Finally, we learn the most optimal feature values to aggregate the summed up costs for each relation category, according to the original human reference scores in training data per sentence. We learn the feature values by training a Gradient Boosting Regression model for each language pair. Our training datasets are WMT Metrics Task MQM 2021 datasets in the following language pairs: English → German, English → Russian, and Chinese → English. After fine-tuning our final model, we test our new evaluation method on the WMT Metrics Task MQM 2022 datasets for the same languages. We compare our results to other evaluation techniques by calculating Pearson's correlation coefficient on the averaged human scores and averaged BiVert scores by translation system detailed in Appendix A.

[1]Simplemma: a simple multilingual lemmatizer for Python at https://github.com/adbar/simplemma

4.1. Training

We trained our model using Gradient Boosting Regression (Friedman, 2002), with different hyperparameters for each language pair. For both English → German and English → Russian we set the learning rate to 0.1; For English→German we used 100 estimators and max depth 6; For English→Russian we used 550 estimators and max depth 7. Both data set labels, the human scores per sentence, are normalized for optimal training. Specifically in Russian training data, we normalized negative human scores to zero, as explained in (Fonseca et al., 2019) section 2.2. For Chinese→English, we use 1000 estimators, max depth of 6, and set the learning rate to 0.05. The English stopwords list is provided by NLTK,[2] and the Chinese stopwords list is from the Stopwords-iso library.[3] Table 3 displays the number of sentences used for training and predicting.

4.2. Results

We evaluate how BiVert's quality judgments fare in comparison to human scores on the full WMT Metrics Task 2022 Dataset. The feature importance scores by language pair are presented in Table 1

[2]Natural Language Toolkit https://www.nltk.org/index.html

[3]A collection of stopwords for multiple languages. https://github.com/stopwords-iso/stopwords-iso

	Train	Predict
English-German	19,501	19,725
English-Russian	12,000	19,725
Chinese-English	16,124	28,124

Table 3: Number of sentences used for training and prediction for each language pair. Prediction is for the whole WMT Metrics Dataset 2022 provided.

by language pair. We evaluated our results by calculating Pearson's correlation between source-language BiVert average scores per system and the human gold-standard aligned scores. We notice that for Chinese the Inflection and Derivation weights are zero, as these processes do not occur at the word level in Chinese. We see that the Sense category is identified with the highest importance value in all language pairs. Thus indicating the success of BabelNet's sense network in assisting with the evaluation quality of the direct translation. In Table 2, we compare our method scores with the correlation scores of other methods mirrored from the WMT22 Metrics Task findings. Morever, Figure 5 represents a graphical display of the correlations calculated for BiVert in Table 2. BiVert achieves the highest score for the English–German language pair among reference-less methods, as well as higher than BERTScore's. For English–Russian BiVert achieves a middle-ranked score, and in Chinese–English the rank is lower. This is perhaps due to the existing categories in use. It's possible that revising these categories to align better with the unique linguistic properties of the Chinese language could improve the results. Furthermore, the quality and coverage of BabelNet data for Russian and Chinese might play a significant role in the challenges we're observing.

5. Conclusion and Future Work

In this paper we present BiVert, a new multilingual reference-less method for evaluating machine translation. This technique introduces an aspect of evaluation using graph senses extracted from semantic graphs, offering an untapped use case for these resources that is simple to implement and has immediate potential to achieve high results compared with human evaluation. Its reference-free application mode allows high-quality evaluation of translation without need for parallel corpora, which can greatly lower the barrier for development of MT systems for low-resource languages and language pairs.

In the future, we aim to assess BiVert's potential to be implemented in other generative NLP tasks. An additional avenue involves the potential role switch between the evaluated system and the state-of-the-art system, where the evaluated system would back-translate the target sentence, thereby enchancing the consistensy of evaluations accross different systems. Moreover, we plan to expand our language categories to cover linguistically diverse languages, and also expand our graph knowledge of senses using resources other than BabelNet, such as Wikionary. Finally, our word alignment algorithm does not currently deal with phrases or idioms, a fascinating avenue for future development.

Acknowledgments

We thank the reviewers for their valuable comments. This research was supported by grant no. 2022215 from the United States—Israel Binational Science Foundation (BSF), Jerusalem, Israel.

Bibliographical References

Judit Ács, Ákos Kádár, and Andras Kornai. 2021. Subword pooling makes a difference. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2284–2295, Online. Association for Computational Linguistics.

David F. Crouse. 2016. On implementing 2d rectangular assignment algorithms. *IEEE Transactions on Aerospace and Electronic Systems*, 52(4):1679–1696.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. 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 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Edsger W Dijkstra. 1959. A note on two problems in connexion with graphs. *Numerische mathematik*, 1(1):269–271.

Helge Dyvik. 1998. *A translational basis for semantics*, pages 51 – 86. Brill, Leiden, The Netherlands.

Erick Fonseca, Lisa Yankovskaya, André F. T. Martins, Mark Fishel, and Christian Federmann.

2019. Findings of the WMT 2019 shared tasks on quality estimation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pages 1–10, Florence, Italy. Association for Computational Linguistics.

Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, Eleftherios Avramidis, Tom Kocmi, George Foster, Alon Lavie, and André F. T. Martins. 2022. Results of WMT22 metrics shared task: Stop using BLEU – neural metrics are better and more robust. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 46–68, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Jerome H Friedman. 2002. Stochastic gradient boosting. *Computational statistics & data analysis*, 38(4):367–378.

Julia Ive, Frédéric Blain, and Lucia Specia. 2018. deepQuest: A framework for neural-based quality estimation. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3146–3157, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

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.

Fabio Kepler, Jonay Trénous, Marcos Treviso, Miguel Vera, and André F. T. Martins. 2019. OpenKiwi: An open source framework for quality estimation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 117–122, Florence, Italy. Association for Computational Linguistics.

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.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

George A. Miller. 1992. WordNet: A lexical database for English. In *Speech and Natural Language: Proceedings of a Workshop Held at Harriman, New York, February 23-26, 1992*.

Roberto Navigli and Simone Paolo Ponzetto. 2012. Babelnet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217–250.

Jekaterina Novikova, Ondřej Dušek, Amanda Cercas Curry, and Verena Rieser. 2017. Why we need new evaluation metrics for NLG. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2241–2252, Copenhagen, Denmark. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Krishna Pillutla, Swabha Swayamdipta, Rowan Zellers, John Thickstun, Sean Welleck, Yejin Choi, and Zaid Harchaoui. 2021. Mauve: Measuring the gap between neural text and human text using divergence frontiers.

Ricardo Rei, Craig Stewart, Ana C. Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. *CoRR*, abs/2009.09025.

Ricardo Rei, Marcos Treviso, Nuno M. Guerreiro, Chrysoula Zerva, Ana C Farinha, Christine Maroti, José G. C. de Souza, Taisiya Glushkova, Duarte Alves, Luisa Coheur, Alon Lavie, and André F. T. Martins. 2022. CometKiwi: IST-unbabel 2022 submission for the quality estimation shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 634–645, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Hadeel Saadany and Constantin Orasan. 2021. BLEU, METEOR, BERTScore: Evaluation of metrics performance in assessing critical translation errors in sentiment-oriented text. In *Proceedings of the Translation and Interpreting Technology Online Conference*, pages 48–56, Held Online. INCOMA Ltd.

Michael Sussna. 1993. Word sense disambiguation for free-text indexing using a massive semantic network. In *Proceedings of the Second International Conference on Information and*

Knowledge Management, CIKM '93, page 67–74, New York, NY, USA. Association for Computing Machinery.

Jörg Tiedemann and Lars Nygaard. 2004. The opus corpus-parallel and free: http://logos. uio. no/opus. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*.

Changping Wang, Chaokun Wang, Zheng Wang, Egsg Dvd, and Philip Yu. 2020. Edge2vec: Edge-based social network embedding. *ACM Transactions on Knowledge Discovery from Data*, 14:1–24.

Daokun Zhang, Jie Yin, Xingquan Zhu, and Chengqi Zhang. 2018. Metagraph2vec: Complex semantic path augmented heterogeneous network embedding.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert.

Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M. Meyer, and Steffen Eger. 2019. MoverScore: Text generation evaluating with contextualized embeddings and earth mover distance. 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 563–578, Hong Kong, China. Association for Computational Linguistics.

Yaoming Zhu, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, and Yong Yu. 2018. Texygen: A benchmarking platform for text generation models. In *The 41st international ACM SIGIR conference on research & development in information retrieval*, pages 1097–1100.

A. Individual evaluation of translation systems

We present the full evaluation scores for the systems in Table 4.

	System	Human	BiVert
English → German	bleu_bestmbr	9.615	0.614
	bleurt_bestmbr	9.555	0.609
	comet_bestmbr	9.567	0.627
	JDExploreAcademy	9.581	0.620
	Lan-Bridge	9.435	0.608
	M2M100_1.2B-B4	8.872	0.598
	Online-A	9.514	0.608
	Online-B	9.585	0.626
	Online-G	9.510	0.605
	Online-W	9.684	0.624
	Online-Y	9.480	0.615
	OpenNMT	9.329	0.603
	PROMT	9.297	0.596
	QUARTZ_TuneReranking	9.462	0.622
	refB	9.634	0.614
English → Russian	bleu_bestmbr	9.715	0.296
	comet_bestmbr	9.677	0.286
	eTranslation	9.417	0.295
	HuaweiTSC	9.476	0.292
	JDExploreAcademy	9.679	0.279
	Lan-Bridge	9.639	0.236
	M2M100_1.2B-B4	9.298	0.272
	Online-A	9.561	0.292
	Online-B	9.701	0.310
	Online-G	9.687	0.333
	Online-W	9.789	0.320
	Online-Y	9.608	0.263
	PROMT	9.548	0.296
	QUARTZ_TuneReranking	9.375	0.310
	SRPOL	9.434	0.305
Chinese → English	AISP-SJTU	9.682	0.443
	bleu_bestmbr	9.701	0.435
	bleurt_bestmbr	9.749	0.452
	comet_bestmbr	9.714	0.459
	HuaweiTSC	9.692	0.443
	JDExploreAcademy	9.718	0.446
	Lan-Bridge	9.753	0.460
	LanguageX	9.727	0.434
	M2M100_1.2B-B4	9.318	0.441
	Online-A	9.627	0.434
	Online-B	9.729	0.444
	Online-G	9.707	0.433
	Online-W	9.605	0.444
	Online-Y	9.672	0.438
	refB	9.801	0.497

Table 4: Individual system scores from WMT on human evaluation and through BiVert.