AUDIO CAPTIONING IN FINNISH AND ENGLISH WITH TASK-DEPENDENT OUTPUT

Irene Martín-Morató, Manu Harju, Annamaria Mesaros

Signal Processing Research Centre, Tampere University, Finland
{irene.martinmorato, manu.harju, annamaria.mesaros}@tuni.fi

ABSTRACT

Describing audio content is a complex task for an annotator; the resulting caption depends on the annotator's language, culture and expertise. In addition, physiological factors like vision impairment may affect on how the sound is perceived and interpreted. In this work, we explore bilingual audio captioning in Finnish and English. In connection with this study, we release the SiVi-CAFE dataset, a small-size dataset of Sighted and Visually-impaired Captions for Audio in Finnish and English, with a collection of parallel annotations for the same clips. We analyze briefly the differences between captions produced by sighted and visually-impaired annotators, and train a system to produce captions in both languages that also mimics the style of different annotator groups. Obtaining a CIDEr score of 34.75% and 28.75% on the English and Finnish datasets, respectively. Furthermore, the system is able to perform a tagging task, obtaining F-score of 79.73%.

Index Terms— audio captioning, visually-impaired users, captioning dataset, tagging, Finnish language

1. INTRODUCTION

Automated Audio Captioning (AAC) is a relatively recent researched topic [1], with potential applications that include accessibility aids [2] and content indexing for search engines [3]. While AAC systems have primarily focused on generating captions in English, there is a general growing demand for personalized content in other languages. Recent years have seen development of multilingual methods for image captioning [4], and also a few studies on multilingual AAC [5, 6]. The mentioned multilingual AAC works use translated captions, in this case between Chinese and English [5] and French, German and English [6].

Multilingual AAC can be obtained by generating captions directly in the target language, or generating captions in English and automatically translating them to the target language. However, while generating captions in English and then translating them to other languages can be faster and more straightforward, some nuances, idiomatic expressions, or cultural references may not translate accurately. Authors of [6] show that direct captioning in the target language may capture specific language nuances better. However, this requires language-specific training data, which is not easily available. Instead, there is the option of translating training data from English to the target language, though the disadvantages remain as pointed out above. Creating training data for AAC is a complex problem. Each annotator brings their unique style, influenced by factors such as age, culture, and language. In general, native speakers tend to use more precise and expansive language compared to non-native speakers [7]. One complicating factor is that humans are used to using language to describe visual rather than other sensory information; this is evident in the fact that languages often have a more extensive vocabulary for describing visual experiences compared to auditory ones [8], and this may affect the quality and diversity of captions, particularly when produced in a second language. Moreover, the annotation procedure affects the reference data: providing additional hints to annotators who can strongly bias their wording, as shown in [9]. Other factors can also influence the way we describe sounds. For example, individuals with visual impairments naturally pay more attention to auditory cues in their daily lives, as they need to rely on different sensory cues to understand their surroundings. Studies show that there are differences in the assessment of soundscape between visually-impaired people (ViP) and non-visually impaired ones, in terms of soundscape pleasantness or quietness [10]. As a special category of users with a heightened awareness of auditory cues, we would expect that visually-impaired annotators create richer audio captions than normal sighted individuals.

Considering the potential applications for captioning, and in particular accessibility, we expect that the need for more personalized output will become an important driving factor in development of captioning systems. To understand the possibility of creating a single universal captioning system that can produce outputs of different styles and in different languages for different categories of users, we adopt the approach proposed in [11] that used a *task embedding* for training an AAC system with different datasets and conditioned it to produce an output in the style of the dataset. In this work, we investigate a multitask training and conditioning across different languages and captioning styles, including ViP users.

The main contributions of this work are as follows: (1) a study of differences in captioning between visually-impaired and normal sighted users, in Finnish language, and a comparison from a linguistic point of view to parallel data in English; and (2) a multitask system trained with different languages and styles: Finnish, English, visually-impaired, biased, and non-biased captions.

There are a few unique elements in this study. Firstly, the use of an agglutinative language, in this case Finnish, as a typologically distant language from English, brings an element of novelty and difficulty to both the system vocabulary and its evaluation. Secondly, to the best of our knowledge, this is the first study using visually-impaired subjects in captioning as a category of annotators. The work aims to understand if such captions bring any advantage for training AAC, assuming they are more detailed. In conjunction with the study, we have published a multi-way annotated dataset that includes captions in English and Finnish: two sets of Finnish captions, ViP and sighted, and two sets of English captions, biased and non-biased in terms of vocabulary. Additionally, the dataset provides

This work was supported by Jane and Aatos Erkko Foundation under grant number 200061, "Guided captioning for complex acoustic environments". The authors wish to thank CSC-IT Centre of Science Ltd., Finland, for providing computational resources.

translations (automatically translated) between Finnish and English for the different captions sets.

The paper is organized as follows: in Section 2 we present shortly the data collection process and we analyze the differences between different types of annotations, focused on the use of language between Finnish and English and ViP and non-ViP. In Section 3 we introduce the multitask model training procedure, while in Section 4 we present the experimental results and discussion. Finally, Section 5 presents the conclusions and future work.

2. CAPTIONS WITH DIFFERENT ANNOTATOR PROFILE

The aim of the data collection process for this study was to obtain a variety of textual description for the same audio clips, in order to study how inter-cultural and linguistic differences between users produce different captions. In addition, we collected data from ViP users to study how visual impairment affects the descriptions. We started from the existing MACS dataset and proceeded with additional annotation tasks that have different annotator profile. The annotation task was similar for everyone, and followed the methodology presented in [9]. Audio clips are 10 seconds long, and the annotation was completed using a web-based interface that provided the clips one by one to be played back and annotated. The annotation process could be paused and continued later by logging in to the web platform. The complete collection of captions is published under the name SiVi-CAFE (Sighted and Visually-impaired CAptions in Finnish and English)[1].

2.1. Four-way data annotation

MACS dataset contains 10-second clips of audio from everyday environments (airport, public square and park) that were annotated by university students in a way that facilitated introducing bias in the captions. Namely, annotators were first given a tagging task, being asked to indicate what sounds from a given list of 10 classes they can hear [9]; after this, they were asked to describe the clip in one sentence. The sentences were found to contain the exact wording of the tags for 41.78% of the sentences [9]. In the SiVi-CAFE collection, this set is referred to as *English-bias*.

The same setup was repeated with another pool of students, this time without the tagging task. In contrast with the observations on biasing, the captions produced in this setup have a larger vocabulary and longer average caption length. In the SiVi-CAFE collection, this set is referred to as *English-nobias*.

Finnish language data collection has focused on obtaining captions from visually-impaired users. The annotation was performed using a company that employs visually-impaired workers for various tasks. We recruited 25 persons, native Finnish speakers, through Aarnikukko Oy[2] and provided them with an accessible web-based tool for the annotation process. Of the 25 annotators with visual disability, 14 reported themselves as blind, 9 as partially sighted, and two did not answer. In addition, 11 participants announced to have some environmental perception via vision, including 3 of the blind individuals. Each worker annotated 180 clips, resulting in 900 clips each having 5 captions. We refer to this set as *Finnish-ViP*.

The same 900 clips were used to collect parallel data in Finnish from normal-sighted people using volunteers who were native speakers; this set of captions is referred to as *Finnish*. This subset is also incomplete, i.e. not all 900 clips have 5 captions.

[1]https://doi.org/10.5281/zenodo.11505823

[2]https://www.aarnikukko.fi/

Dataset	Audio clips	Unique sentences	Sentence length (std)	Vocab. size	MATTR (std)
English-bias	3930	16262	9.5 (3.89)	2717	0.26 (0.02)
English-nobias	2050	9679	10.2 (3.78)	2685	0.27 (0.02)
Finnish-ViP	900	4458	8.3 (3.25)	4518	0.39 (0.03)
Finnish	900	3592	7.5 (2.79)	3540	0.37 (0.03)

Table 1: Statistics of the collected datasets.

2.2. Analysis of the annotations

The main difficulty in data collection was recruiting sufficiently many annotators. Some tasks were implemented with student volunteers that received various rewards for their time (e.g. movie tickets). There was added difficulty in recruiting digitally fluent visually-impaired workers; for this reason the *Finnish-ViP* data is relatively small. Moreover, while the Finnish annotators are native speakers, the ones providing English annotations are international students using English in their studies, hence very likely not English native. As discussed earlier, this probably affects their use of language for describing the sounds. The *English-bias* data was produced by 133 annotators, *English-nobias* by 89, *Finnish-ViP* by 25, and *Finnish* by 42. The sets are each somewhat incomplete, but there are 3612 captions provided for 900 clips which were annotated by all categories of users, and can be considered as parallel data. For completeness, all original data was translated into the other language using the DeepL translation API[3], following [6].

The statistics of the different caption sets are provided in Table 1. To characterize the lexical diversity, we use the type-token ratio (TTR) the ratio between the unique words (types) and total words (tokens) in each set. To account for the difference in size, we calculate the moving average TTR (MATTR) [12] which calculates TTR every 500 words, hence MATTR allows comparing texts of different lengths. While the two languages are not comparable, the difference between *English-bias* and *English-nobias* shows a difference in lexical diversity, as does the *Finnish-ViP* compared to *Finnish*.

The 3612 captions that form a parallel corpus results in a vocabulary of 1132 and 1328 for the *English-bias* and *English-nobias* sets, respectively, while for *Finnish-ViP* and *Finnish* the vocabulary size is 2142 and 2498, respectively. The *Finnish-ViP* set has the richest vocabulary; this is also reflected in the high MATTR.

The most interesting detail is the way annotators describe the location of the sounds in the audio clips. While all groups indicated sounds as appearing in the background (Fi: taustalla), in the distance (etäällä), far away (kaukainen), or less often nearby (Fi: lähempänä, comp.), the visually impaired Finnish speakers described egocentric directions by indicating sounds being ‘on the right’ (oikealla) or ‘on the left’ (vasemmalla). In the *Finnish* set, ‘on the left’ appears 10 times and ‘from the left’ once, while in the *Finnish-ViP* set there are 443 variants for ’left’ (including ‘to the left’, ‘on the left’, ‘from the left’, ’front left’, ’back left’). Similarly, variations of ’right’ appear 397 times in the *Finnish-ViP* set, and only 8 times in the *Finnish* set. In the English data “on the left” appears 19 times and “on the right” only 14 times.

3. A UNIVERSAL CAPTIONING SYSTEM

A single model is trained using all the different annotation types, in order to create a universal captioning system. We employ a task embedding token as proposed in [11]; each different annotation type is seen as a task that is assigned a specific token. Figure 1 illustrates

[3]https://www.deepl.com/pro-api

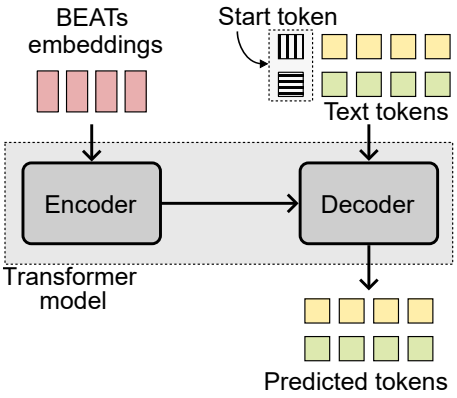

Figure 1: Block diagram of the AAC system with task tokens.

how a *start token* is concatenated at the beginning of the sentence for each of the different annotation types. The translated datasets are also considered as separate tasks, to provide the system with the ability of producing as large variety of styles as possible.

The model follows a standard transformer architecture and a pre-trained tokenizer. The tokenizer is based on Byte-Pair Encoding (BPE), and creates a list of unique words with their frequency; a vocabulary size parameter has to be selected beforehand. Before training the transformer, the first step is to fit the tokenizer. The role of the tokenizer is to split the sentences into words and then into subwords. Finally, those subwords are converted to ids using a look-up table; this will facilitate generation of words that have not been seen in the training vocabulary, achievable by breaking unknown words into smaller units that the tokenizer can recognize. The tokenizer is trained with the vocabulary of all the datasets, English and Finnish originals and the translated versions. The maximum vocabulary size is set to 5000, which was experimentally found to be sufficient to wrap English and Finnish language. For each dataset we use a *start token* as done in [11], indicating to which dataset the caption belongs to. The audio is fed to the model after a pre-processing step where a feature extractor is used. We use the pre-trained encoder BEATs [13] as feature extractor; the resulting embeddings are used as inputs to the transformer encoder. We chose BEATS as audio representation based on the system that achieved the best performance in the DCASE 2023 Challenge Audio Captioning task. However, to reduce the number of input tokens, we average pool the BEATs embeddings over the time dimension with a factor of 32.

3.1. Experimental setup and evaluation

The system is evaluated in a 10-fold manner, because the distribution of the data is unbalanced among datasets; the smaller dataset (*Finnish*) is used as norm for splitting the data into folds based on the 10 cities where the data has been recorded. We report results using BLEU [14] (a measure of n-gram overlap between generated and reference captions) and CIDEr-D [15] (consensus-based measure from image captioning), as language-agnostic measures. We also use sentence-BERT cosine similarity (sBERT$_{sim}$) [16] as a more meaning-oriented metric that compares the captions at sentence level. For the Finnish captions we calculate this metric using TurkuNLP_sbert [17], shown to perform better on the Finnish language tasks than a multilingual version; for English we use paraphrase-multilingual-mpnet-base-v2 as used in [6].

Dataset	**BLEU$_1$**	**CIDEr**	**sBERT$_{sim}$**
English-bias	45.65	20.95	60.15
English-nobias	48.40	21.61	60.13
Finnish-ViP	29.98	9.97	72.88
Finnish	26.80	12.38	74.64

Table 2: Human-to-human evaluation of captions. One caption is randomly selected as predicted and compared with the other captions available for the same clip.

3.2. Human-to-human evaluation

To analyze the connection between system predictions and human-produced annotations, we calculate the human-to-human comparison for the datasets using the same metrics. Their values for the original (annotated) data are presented in Table 2. Unigram overlaps, shown by BLEU$_1$, are strong for the English datasets and less for Finnish; based on CIDEr, *Finnish-ViP* has the least consensus in descriptions between annotators. $sBERT_{sim}$ is very similar for the English sets, while *Finnish-ViP* has a somewhat higher $sBERT_{sim}$ than Finnish, indicating that ViP annotations are more similar in meaning, even though their wording differs.

4. EXPERIMENTAL RESULTS

4.1. Captioning results

Table 3 shows the AAC metrics for the multitask model, including cross-testing in which we generate the caption with a specific task token, and evaluate against a different reference set of the same language. For the *English-bias* data, the model achieves a CIDEr score of 34.72%, which is, surprisingly, almost 14 points higher than the human performance. This can be attributed to the fact that the models typically generate rather repetitive captions, more so than the human annotators. We verify this by inspecting the top 3-grams: "in the background" appears 607 and 887 times in *English-bias* and *English-nobias*, respectively, while in the predicted outputs they appear 167 times and 372 times for the *English-bias* and *English-nobias* captioning style, respectively. The next most common 3-grams in the predicted captions are "talking and walking", appearing 210 and 154 times, respectively; and "people are talking", 71 and 235 times. For the *Finnish* dataset we achieve a CIDEr of 17.31%, while for *Finnish-ViP* we achieve a CIDEr of 13.88%, both higher than the human-to-human evaluation.

For comparison, we trained monolingual models as multitask models but using only the data from a single language, including the automatically translated captions from the other language. In general the monolingual models had a slightly worse performance, being trained with less data. For *English-bias* we achieve BLEU$_1$ 63.80%; CIDEr 33.95% and sBERT$_{sim}$ 60.36%, while for *Finnish*, we achieve BLEU$_1$ 42.97%; CIDEr 15.99% and sBERT$_{sim}$ 74.18%.

Comparing the predicted captions against reference captions with a different style produces lower scores, with a few exceptions: *Finnish* vs *Finnish-ViP* has a good BLEU$_1$, showing a high overlap in unigrams; all cross-evaluations for Finnish language have a very similar sBERT$_{sim}$, showing that the descriptions are similar in meaning, although not in the exact wording. English sets always score much lower when evaluated against another style.

Prediction	Reference	BLEU$_1$	BLEU$_4$	METEOR	CIDEr	sBERT$_{sim}$
English-bias	English-bias	67.07	18.32	20.60	34.72	61.64
English-bias	English-nobias	62.03	14.58	17.86	25.01	59.46
English-nobias	English-nobias	69.16	21.40	21.69	33.66	59.28
English-nobias	English-bias	58.64	13.80	18.66	26.04	57.63
Finnish-ViP	Finnish-ViP	51.30	4.85	14.63	13.88	73.73
Finnish-ViP	Finnish	43.64	4.54	13.50	15.17	74.01
Finnish	Finnish	46.29	5.28	14.02	17.31	75.04
Finnish	Finnish-ViP	50.30	5.55	13.90	13.63	74.40

Table 3: Results on all the datasets using the multitask model, with evaluation across same language reference sets.

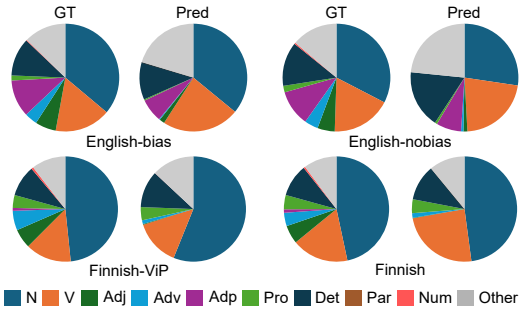

Figure 2: Pie charts showing the POS analysis of the English and Finnish datasets. N (Noun), V (Verb), Adj (Adjective), Adv (Adverb), Adp (Adposition), Pro (Pronoun), Par (Particle), Num (Numeral). GT stands for Ground truth; Pred stands for predicted captions from the multitask model.

4.2. Generated language analysis

We investigate the sentence structure in the reference and predicted captions by performing a Part-Of-Speech (POS) analysis and visualize the proportions of POS as pie charts in Fig. 2; for English we use the spaCy[4] toolbox, for Finnish the Finnish-tagtools software[5]. We easily notice that captions are mainly formed using nouns and verbs, with nouns dominating the sentences; the reference annotations also contain a non-negligible percentage of adverbs and adjectives. The charts show a clear difference between the languages: English datasets make more use of verbs, prepositions and determinants, while the Finnish datasets use more nouns, adverbs and pronouns. The predicted captions on the other hand contain almost no adverbs or adjectives, which is an interesting observation that holds for both languages. The system produces a good proportion of adpositions for English and pronouns for Finnish, but overall the model is mostly generating nouns and verbs. The difference between *English-bias* and *English-nobias* is reflected in the predicted captions: "adults talking", "children voices" and "birds singing" are mentioned 154, 58 and 183 respectively for *English-bias* style, while they do not appear in this exact form at all in *English-nobias*. This comes from the training data, where they appear 2179, 443 and 1241 times, and only 20, 1 and 113 times in *English-bias* and *English-nobias*, respectively.

4.3. Tagging system

As a different annotation type indicated by the *start token*, it is also possible to use the multitask model as a tagging system. In this case, instead of the caption, the system receives in training the concatenated tags, seen as a sentence, though it is not a grammatically correct one. The *English-bias* dataset has tags available that were collected during the same annotation process as the caption, as explained in [9]. With the task token we indicate that we require similar "sentences". The model achieves an overall micro-F1 score of of 79.73% (Precision 77.01% and Recall 82.64%) for tagging.

GT tags	Predicted caption
adults talking, traffic noise, music	"**music** is playing and people are talking"
children voices, footsteps	"birds singing and **children voices**"
birds singing, traffic noise	"**traffic noise** and **birds singing**"
adults talking, footsteps	"people are talking and walking"

Table 4: Examples of *English-bias* predicted captions and the reference tags for the respective clips; tags exact matches are in bold.

Tags also allow evaluating if the predicted captions match the sound events tagged in the reference for each clip. If we evaluate the predicted *English-bias*-style captions against the reference tags as captions, we obtain BLEU$_1$ 27.66%, CIDEr 17.07% and sBERT$_{sim}$ 67.14%; for *English-nobias*-style captions BLEU$_1$ is 10.40%, CIDEr is 4.16% and sBERT$_{sim}$ is 57.40%. This evaluation setup illustrates well the induced bias, i.e. the annotators being hinted the tags while listening the clip for recognizing the sounds.

Finally, we calculate to what extent the reference tags are present in the predicted captions, obtaining that 51.3% of the predicted captions with the *English-bias* task token have at least one correct n-gram. A few examples are shown in Table 4. Only exact matches can be easily identified; however, we can observe that captions may contain very similar words to the tags, e.g."people are talking" matching in meaning "adults talking" in the provided examples.

5. CONCLUSIONS

This paper presented a more linguistically-oriented study to AAC, focusing on a parallel corpus of linguistically-different references. The work introduced a dataset comprised of captions in English and Finnish, including annotations provided by visually-impaired users. We designed a multitask system that can produce captions in all required styles, including tags. The dataset analysis shows differences between languages and user types, which were well modeled by the proposed method. Most importantly, the proposed captioning system was capable to learn from a collection of tasks that share some information, i.e. the audio content, but are at the same time very different, i.e. the language or style. We have also successfully shown that the system can be combined with more simplified tasks, in this case audio tagging, paving the way for developing linguistically-mixed systems that can handle multiple languages and multiple sentence styles.

[4]https://spacy.io/
[5]http://urn.fi/urn:nbn:fi:lb-2021101101

6. REFERENCES

[1] X. Mei, X. Liu, M. D. Plumbley, and W. Wang, "Automated audio captioning: an overview of recent progress and new challenges," *EURASIP journal on audio, speech, and music processing*, vol. 2022, no. 1, p. 26, 2022.

[2] O. Alonzo, H. V. Shin, and D. Li, "Beyond subtitles: Captioning and visualizing non-speech sounds to improve accessibility of user-generated videos," in *Proceedings of the 24th International ACM SIGACCESS Conference on Computers and Accessibility*, ser. ASSETS '22. New York, NY, USA: Association for Computing Machinery, 2022.

[3] A. S. Koepke, A.-M. Oncescu, J. F. Henriques, Z. Akata, and S. Albanie, "Audio retrieval with natural language queries: A benchmark study," *IEEE Transactions on Multimedia*, vol. 25, pp. 2675–2685, 2022.

[4] R. Ramos, B. Martins, and D. Elliott, "LMCap: Few-shot multilingual image captioning by retrieval augmented language model prompting," in *Findings of the Association for Computational Linguistics: ACL 2023*, A. Rogers, J. Boyd-Graber, and N. Okazaki, Eds. Toronto, Canada: Association for Computational Linguistics, July 2023, pp. 1635–1651.

[5] M. Wu, H. Dinkel, and K. Yu, "Audio caption: Listen and tell," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 830–834.

[6] M. Cousin, E. Labbé, and T. Pellegrini, "Multilingual Audio Captioning using machine translated data," Sept. 2023, working paper or preprint. [Online]. Available: https://hal.science/hal-04220315

[7] C. Bentz, A. Verkerk, D. Kiela, F. Hill, and P. Buttery, "Adaptive communication: Languages with more non-native speakers tend to have fewer word forms," *PloS one*, vol. 10, no. 6, p. e0128254, 2015.

[8] B. Winter, "Sensory linguistics," *Converging Evidence in Language and Communication Research*, 2019.

[9] I. Martn-Morató and A. Mesaros, "Diversity and bias in audio captioning datasets," in *Proceedings of the 6th Detection and Classification of Acoustic Scenes and Events 2021 Workshop (DCASE2021)*, Barcelona, Spain, November 2021, pp. 90–94.

[10] J. Vida, J. A. Almagro, R. García-Quesada, F. Aletta, T. Oberman, A. Mitchell, and J. Kang, "Urban soundscape assessment by visually impaired people: First methodological approach in granada (spain)," *Sustainability*, vol. 13, no. 24, 2021.

[11] E. Labbé, T. Pellegrini, and J. Pinquier, "CoNeTTE: An efficient Audio Captioning system leveraging multiple datasets with Task Embedding," Sept. 2023, working paper or preprint. [Online]. Available: https://ut3-toulouseinp.hal.science/hal-04193791

[12] M. Covington and J. McFall, "Cutting the gordian knot: The moving-average type-token ratio (MATTR)," *Journal of Quantitative Linguistics*, vol. 17, pp. 94–100, 05 2010.

[13] S. Chen, Y. Wu, C. Wang, S. Liu, D. Tompkins, Z. Chen, W. Che, X. Yu, and F. Wei, "BEATs: Audio pre-training with acoustic tokenizers," in *International Conference on Machine Learning*. PMLR, 2023, pp. 5178–5193.

[14] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: A method for automatic evaluation of machine translation," in *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ser. ACL '02. USA: Association for Computational Linguistics, 2002, p. 311–318.

[15] R. Vedantam, C. Lawrence Zitnick, and D. Parikh, "CIDEr: Consensus-based image description evaluation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 4566–4575.

[16] N. Reimers and I. Gurevych, "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)*, K. Inui, J. Jiang, V. Ng, and X. Wan, Eds. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 3982–3992.

[17] J. Kanerva, F. Ginter, L.-H. Chang, I. Rastas, V. Skantsi, J. Kilpeläinen, H.-M. Kupari, J. Saarni, M. Sevón, and O. Tarkka, "Finnish paraphrase corpus," in *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa'21)*. Linköping University Electronic Press, Sweden, 2021, pp. 288–298.