Semantically Meaningful Metrics for Norwegian ASR Systems

Janine Rugayan, Torbjørn Svendsen, Giampiero Salvi

Department of Electronic Systems, NTNU

{janine.rugayan, torbjorn.svendsen, giampiero.salvi}@ntnu.no

Abstract

Evaluation metrics are important for quanitfying the performance of Automatic Speech Recognition (ASR) systems. However, the widely used word error rate (WER) captures errors at the word-level only and weighs each error equally, which makes it insufficient to discern ASR system performance for downstream tasks such as Natural Language Understanding (NLU) or information retrieval. We explore in this paper a more robust and discriminative evaluation metric for Norwegian ASR systems through the use of semantic information modeled by a transformer-based language model. We propose Aligned Semantic Distance (ASD) which employs dynamic programming to quantify the similarity between the reference and hypothesis text. First, embedding vectors are generated using the NorBERT model. Afterwards, the minimum global distance of the optimal alignment between these vectors is obtained and normalized by the sequence length of the reference embedding vector. In addition, we present results using Semantic Distance (SemDist), and compare them with ASD. Results show that for the same WER, ASD and SemDist values can vary significantly, thus, exemplifying that not all recognition errors can be considered equally important. We investigate the resulting data, and present examples which demonstrate the nuances of both metrics in evaluating various transcription errors.

Index Terms: ASR evaluation metric, semantic context

1. Introduction

The most widely used evaluation metric for automatic speech recognition (ASR) systems is the word error rate (WER). It is computed as the edit distance between the reference word sequence and the transcription word sequence normalized by the total number of words in the reference word sequence [1]. It, therefore, provides an indication of the performance of the system at the literal word level. However, if we are to consider the ASR system's usability for an end application where the overall meaning of the transcription matters, then the standard WER may be insufficient.

To give an example, it is important for a broadcast news transcription service to produce transcriptions that relay the exact same message as what has been spoken. Given that you have two different ASR systems, both of them transcribed an utterance with the same resulting WER. However, one of them may actually be semantically closer to the reference sentence. A single word error alone can cause the transcription to completely change its meaning. For example, the ASR system transcribed "He was eight" as "He was late". As such, evaluating the ASR system's performance using WER does not guarantee that the system would perform well on downstream tasks related to Natural Language Understanding (NLU) or information retrieval. Wang *et al.* [2] demonstrated that better WER does not necessarily correlate to better spoken language understanding accuracy. Their results show that transcriptions with lower WER may correspond to better understanding accuracy. It signified the importance of understanding as an optimization objective over reduction of WER.

A number of metrics have been proposed in the past to overcome the limitations posed by WER. Morris *et al.* [3] introduced word information lost (WIL) as a new performance measure. It is defined as $1 - \text{WIP}$, where word information preserved (WIP) is approximated as the mutual information between the reference and the automatic transcription. McCowan *et al.* [1] presented the evaluation of ASR systems as an information retrieval task wherein each word is treated as an information unit. Precision and recall were the measures proposed. The aforementioned methods focus on the word level accuracy of the ASR system and do not account for any semantic information.

Roy [4] presented Semantic-WER, wherein specific weights are assigned for substitution, deletion and insertion. An importance weight is applied to entities and sentiment words which when incorrectly transcribed can entirely change the meaning of the sentence. Its major drawback is that the reference must be labeled with the entities for evaluation. In addition, due to the coupling of ASR systems with various natural language processing systems, [5]–[7] presented domain-specific metrics that can better reflect the system performance on specific downstream tasks. The limitation of these approaches is that they do not easily generalize for various NLU applications.

The development of language models using transformers [8] such as Bidirectional Encoder Representation from Transformers (BERT) [9] and Robustly optimized BERT approach, RoBERTa [10], allowed the representation of text in the form of embedding vectors that contain semantic information. This was demonstrated in [11] on its successful application to semantic textual similarity decision. Notably, Kim *et al.* [12] introduced an implementation of a semantics-based evaluation metric called Semantic Distance (SemDist). SemDist first computes token-level embedding vectors of the reference and the hypothesis using a pre-trained RoBERTa model. Then it generates sentence-level embedding vectors by averaging the token embeddings along the utterance. Finally, the distance between these sentence-level embedding vectors in the embedding space is quantified by the cosine distance. SemDist has the potential drawback to blur the effect of local discrepancies due to the averaging of the token embeddings. In order to overcome this limitation, we propose a new method based on dynamic programming that we call Aligned Semantic Distance (ASD). Like SemDist, ASD uses BERT-based token embeddings. However, our method computes the distance between the reference and the hypothesis by performing token-wise comparison and finding the best semantic alignment between them.

In this work, we present a qualitative and quantitative evaluation of our method and compare it to SemDist on a Norwegian ASR system. Norwegian poses some interesting challenges. The language has two written standards, Nynorsk and Bokmål,

which to some extent are overlapping: 30-40% of the entries in full-form lexica for the two written standards are identical; a significant part of other entries have only minor differences; and both standards allow for variants in orthography and inflections. Moreover, the use of dialect is common even in formal settings in Norway, and since dialect speech does not usually conform to any of the written standards, transcription is at best non-trivial. In our study we use NorBERT [13], a BERT model trained on Norwegian corpora containing both Nynorsk and Bokmål.

2. Methods

The goal of the method is to define a semantic distance metric to compare a reference transcription to an ASR hypothesis with possibly different lengths. Both methods described below are based on first segmenting the text into tokens by a pre-trained tokenizer, and, then, computing token-level embedding vectors. Both the tokens and the embeddings are obtained by means of a contextualized language model based on transformers [8]. The use of positional encoding and attention mechanism in the model architecture allows us to produce unique embedding vectors that depend mainly on the combination of words in the sequence and their position in it. Moreover, BERT [9] with its Masked Language Modeling (MLM) pre-training task allows us to encode sentences such that each word is represented by tokens that are based on both the left and right context around the word.

Token embeddings $e_{\text{ref}}[i], i \in [1, N]$ and $e_{\text{hyp}}[j], j \in [1, M]$ for the reference transcription and ASR hypothesis are obtained by stacking the output of all the layers in the BERT model. Although we set a maximum number of tokens for practical reasons in the model, the length of the resulting representations will depend on the length of the input transcription:

$$\begin{aligned} E_{\text{ref}} &= \{e_{\text{ref}}[1], e_{\text{ref}}[2], ..., e_{\text{ref}}[N]\}, \\ E_{\text{hyp}} &= \{e_{\text{hyp}}[1], e_{\text{hyp}}[2], ..., e_{\text{hyp}}[M]\}. \end{aligned} \tag{1}$$

In the next subsections, we briefly describe the existing metric SemDist [12], and define our proposed new evaluation metric called Aligned Semantic Distance.

2.1. Semantic Distance (SemDist)

In [12], sentence-level embedding vectors e_{ref} and e_{hyp} for reference and ASR hypothesis are calculated by averaging the vectors in Equation 1 over all the tokens in the sequences. The SemDist is, then calculated by cosine distance between e_{ref} and e_{hyp}:

$$\text{SemDist}(e_{\text{ref}}, e_{\text{hyp}}) = 1 - \frac{(e_{\text{ref}})^T \cdot e_{\text{hyp}}}{\|e_{\text{ref}}\| \cdot \|e_{\text{hyp}}\|} \tag{2}$$

As shown in [12], this metric is bounded by the range $[0, 1]$. Cosine distance is chosen due to its successful application to determining semantic textual similarity in [11].

2.2. Aligned Semantic Distance (ASD)

We believe that some information is lost when we take the average of all the token embedding vectors in the sequence. Furthermore, as will be shown in our experiments, averaging over longer sequences leads to blurring of the representation, and a consequent reduction of the SemDist metric. To overcome these problems, we propose to perform token-wise comparison in our evaluation metric called Aligned Semantic Distance (ASD). ASD adopts dynamic programming (DP) to quantify the similarity between the reference and the hypothesis. Given the two sequences of token-level embeddings defined in Equation 1, the ASD is calculated as

$$\text{ASD}(E_{\text{ref}}, E_{\text{hyp}}) \triangleq \min_{\phi} \frac{1}{N} D_{\phi}(E_{\text{ref}}, E_{\text{hyp}}), \tag{3}$$

where ϕ is the alignment path between the two embedding sequences,

$$\phi = \{(i, \phi(i)); 1 \leq \phi(i) \leq M\}$$

and the distance along a specific path is

$$D_{\phi}(E_{\text{ref}}, E_{\text{hyp}}) = \sum_{i=1}^{N} d(e_{\text{ref}}[i], e_{\text{hyp}}[\phi[i]]) \tag{4}$$

with $d(x, y)$ as the local distance metric, which in our experiments is the cosine distance.

ASD uses dynamic programming to perform the minimization of Equation 3. $\text{ASD}(E_{\text{ref}}, E_{\text{hyp}})$ is equivalent to the accumulated distance between the reference and ASR tokens aligned along the optimal path ϕ normalized by the length of the reference embedding.

3. Experimental Setup

The ASR transcriptions evaluated in this paper are generated by a Norwegian ASR system developed by our project partners from the Technical University of Liberec. The lexicon contains 573k words, with a mixture of Nynorsk and Bokmål. The language model is a bigram trained on a text corpus comprised of publicly available text sources, mostly in Bokmål. The acoustic model has a hybrid HMM/DNN architecture, and it is trained on approximately 250 hours of data.

The speech is from various sources, all having manual reference transcriptions. A part of the test set of Rundkast [14], a database of Norwegian radio broadcast news shows, and the free speech test set of NB Tale [15] were used. In addition, the Norwegian Parliamentary plenary meeting from May 10, 2021 [16] was transcribed as well based on the minutes of the meeting that are available online. The final test set contained reference and ASR generated transcriptions of approximately 800 sentences or sentence-like segments.

Since we want to develop an evaluation metric for Norwegian ASR systems, we utilized a pre-trained large-scale monolingual language model for Norwegian called NorBERT [13]. This model employs the same architecture and pre-training tasks as the standard BERT model. The training corpus consists of a combination of Nynorsk and Bokmål news text and Wikipedia dumps, and contains roughly 2 billion words. It uses customized WordPiece [17] embeddings with a vocabulary size of 30,000. The model has 12 layers with a hidden size of 768, and 12 self-attention heads. We used the pre-trained model available from the Nordic Language Processing Laboratory word embeddings repository[1]. To run the model and its corresponding tokenizer, we used the Hugging Face Transformers API[2].

To compute the proposed metric we used the dtw-python[3] package [18] to implement the DP algorithm with cosine distance as the pointwise distance function.

[1] http://vectors.nlpl.eu/repository/

[2] https://huggingface.co/docs/transformers/v4.17.0/en/index

[3] https://pypi.org/project/dtw-python/

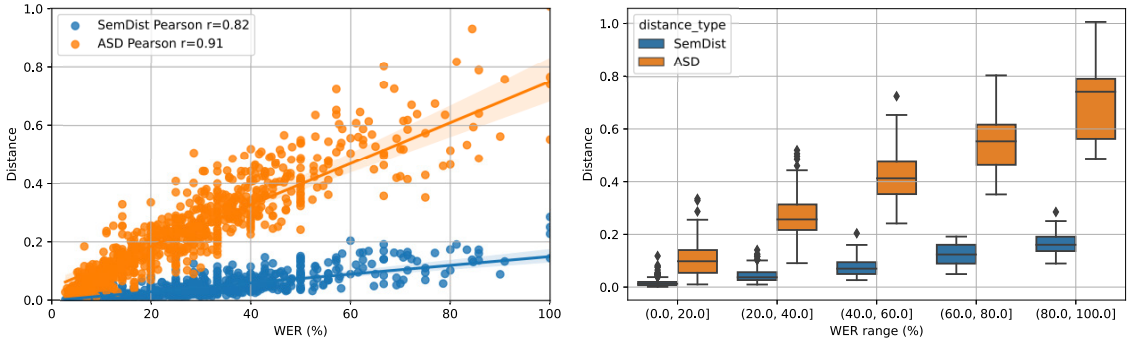

Figure 1: *Comparison of SemDist and ASD as a function of sentence WER for the test set. Left: scatter plot, linear regression and correlation. Right: boxplot. The two metrics span different ranges and should only be compared in relative terms.*

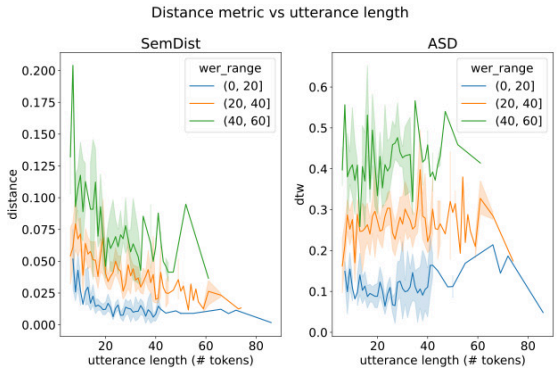

Figure 2: *Average and standard deviation of distance as a function of utterance length (# of tokens) for SemDist (left) and ASD (right). Colors correspond to data points in different ranges of the sentence WER. WERs higher than 60% are not displayed to reduce the clutter in the figure.*

4. Results and Discussion

We first analyze the correlation between WER and the two semantic metrics (SemDist and ASD). Figure 1 illustrates the dependency in two different ways: scatter plot (left) and boxplot(right). In both cases, we restrict the WER to be less than or equal to 100%. We observe that similar to SemDist, ASD is highly positively correlated with the WER. Furthermore, ASD has a higher Pearson correlation coefficient of 0.91 compared to 0.82 for SemDist.

Figure 2 shows the relationship between the semantic metrics and the length of the reference transcription (in # of tokens). The figure depicts the mean and standard deviation of the measures for specific ranges of the WER, as a function of # of tokens. We can observe that SemDist tends to fade with the utterance length, whereas ASD is stable with respect to the length. This is confirmed by the Pearson correlation coefficient that is -0.34 for SemDist and -0.04 for ASD. This finding can be attributed to SemDist performing mean pooling over all the token embeddings prior to calculating the cosine distance. As the utterance length increases, it further blurs the effect of each individual token, and results in sentence-level embedding vectors that tend to be more similar (lower distances). In ASD, on the other hand, token embeddings are first realigned and the global distance is computed as a combination of the local distances, thus preserving the discrimination.

Turning now to the distribution of the values at different ranges of WER, we observe from Figure 1 that SemDist values have an increasing interquartile range (IQR) as the WER increases, with the exception of values at the 80-100% WER range. Meanwhile, we only observe significant increase in IQR for ASD at WER values 60% and higher. The figure also shows that ASD maximum and minimum values (excluding outliers) have a linear relationship with increasing WER, while the relationship for SemDist seems to be non-linear. We also observe that there are more outliers at lower WERs. This may be due to the higher amount of samples that lie within this range. Because we are interested to study the cases where the semantic metrics differ from WER, we will focus our attention to the outliers in Figure 1 and analyze some examples in details. Before we present the examples, however, its important to point out that the two metrics operate on rather different ranges of values. Although SemDist is theoretically confined within the range [0,1], Figure 1 shows that for our particular dataset, the maximum SemDist value attained is around 0.25 only. On the other hand, ASD spans the whole range [0,1]. For this reason, the two metrics should not be compared directly, but rather in relative terms.

Furthermore, we observe some representative examples from the outliers of the boxplot in Figure 1. Transcribing proper names is one common shortcoming of ASR systems. In Table 1, we demonstrate the ability of the semantics-based metrics to distinguish this type of errors. In each example, we present the transcription from the actual Norwegian ASR system and a hypothesis we constructed for illustrative purposes. The constructed hypothesis has errors that are also typical for ASR systems: insertion of short words (i.e. *"og"* (and)) and substitution with phonetically close words (i.e. *"nye"* (new) and *"mye"* (much)). The first example in Table 1 shows the ASR system transcribed *"tregt"* (slow) as *"Trek"* which is a proper name, while in the constructed hypothesis *"for"* (for) is transcribed as *"før"* (before). Although the WER is the same in the two cases, we can see that SemDist and ASD values are higher when the ASR system incorrectly transcribed the adjective into a proper name. In the second example the proper name "Mugabe" is incorrectly transcribed by the ASR system, while in the constructed hypothesis other, less semantically important parts, were wrong. Again, the WER remains unchanged, whereas ASD and SemDist give much more weight to mistakes of proper names as opposed to less meaningful mistakes.

In Table 2, we show examples demonstrating the disagreement between the metrics when the WER is considerably high.

Table 1: *ASR transcriptions with errors involving proper names.*

Type	Transcription	WER	SemDist	ASD
REF	regjeringens opptrappingsplan for psykisk helse går for **tregt**			
Actual ASR	regjeringens opptrappingsplan for psykisk helse går for Trek	0.125	0.054	0.255
Constructed Hyp	regjeringens opptrappingsplan før psykisk helse går for tregt	0.125	0.019	0.102
REF	**Mugabe** ble i fjor gjenvalgt ved hjelp av massivt valgfusk for nye seks år			
Actual ASR	Vu grabbe ble i fjor gjenvalgt ved hjelp av massivt valgfusk for nye seks år	0.143	0.047	0.287
Constructed Hyp	Mugabe ble i fjor gjenvalgt og ved hjelp av massivt valgfusk for mye seks år	0.143	0.021	0.121

Table 2: *ASR transcriptions with WER greater than 20%.*

Type	Transcription	WER	SemDist	ASD
REF	**de pengene** bare forsvinner ut i dragsuget			
ASR	det penger bare forsvinner ut i dragsuget	0.286	0.021	0.091
REF	**Synnøve Finden er** tilbake i Rema hyllene kjeden har forstått at kundene bestemmer **mener markedsforsker**			
ASR	synet ved finnene tilbake i Rema hyllene kjeden har forstått at kundene bestemmer min markets forska	0.400	0.081	0.519
REF	**lurer på hvorfor** noen av kollegene må gå mens **Knut Grøholt** fremdeles **blir sittende**			
ASR	jo på hvor få noen av kollegene må gå mens søtning og holdt fremdeles lese	0.571	0.133	0.637

Table 3: *Nynorsk reference transcription.*

Type	Transcription (WER=0.444, SemDist=0.047, ASD=0.245)
REF	mange nye **kommunar** har derfor **teke i bruk** eigedomsskatt
ASR	mange nye kommuner har derfor innført eigedomsskatt

The first example shows that even though 30% WER would be deemed unacceptable, the low values of SemDist and ASD indicate that the ASR transcription may actually be useful. The erroneous words in the transcription did not impose high penalties on the ASD or SemDist values because the phrases *"de pengene"* and *"det penger"* both refer to the same concept, which is "money" (*"penger"*) .

On the other hand, the second example shows an inconsistency between SemDist and the other two metrics. The ASR system incorrectly transcribed the proper name *"Synnøve Finden"* and the last two words, the verb *"mener"* (to mean) and the compound word *"markedsforsker"* (market researcher). Interestingly, the SemDist value is not substantially high even if the subject of the sentence, *"Synnøve Finden"*, was completely missed out in addition to the other errors. This may be explained by the blurring phenomenon that we have described earlier that is due to taking the average over all the tokens to compute SemDist. As a consequence, SemDist seems to underestimate the errors if the utterance length is increased.

Lastly, the third example illustrates a combination of the most common mistakes the ASR system makes. The reference word *"hvorfor"* (why) is phonetically close to *"hvor få"* (how few) but has an entirely different meaning. Furthermore, the proper name *"Knut Grøholt"* and the phrase *"blir sittende"* (stays seated) were also transcribed incorrectly. These errors make the ASR transcription completely useless, and the high values of the ASD and SemDist reflect this.

Another significant observation is the special case wherein the reference transcriptions are not the true match for the actual spoken words. It can be remarked that this a common occurrence, especially for Norwegian due to its two written standards. In Table 3, we show an example case where the reference transcription used Nynorsk writing standard. Both the reference and ASR transcription have the same meaning: *many new municipalities have therefore introduced property taxes*. The word *"kommunar"* in the reference text is the Nynorsk word for municipality, while *"kommuner"* is the Bokmål version. Here, we also observe the interesting case where the reference transcription has chosen to use the Nynorsk phrasing *"teke i bruk"* (taken into use), while the ASR transcription is *"innført"* (introduced), which is what is actually spoken, and a legal word in both Bokmål and Nynorsk. As desired, the spelling variations and reduction of a phrase to a single word did not result in substantially high ASD or SemDist values. It exemplifies how both metrics do not impose high penalties if the words have similar contexts, since their vectors lie close to each other in the embedding space. Furthermore, while the ASR transcription would normally be rejected because its WER is above the acceptable range, the ASD and SemDist values prove that it is actually acceptable and has a context similar to the reference.

To conclude, our observations indicate that both SemDist and the proposed ASD provide more meaningful metrics to evaluate ASR results compared to WER. By incorporating semantic information, we may be able to distinguish between errors that lead to misunderstanding of the message and minor errors such as variants in orthography or inflections. Furthermore, ASD has the desired property to be more stable with respect to the utterance length when compared to SemDist.

5. Conclusions

In this paper, we propose a new evaluation metric called Aligned Semantic Distance (ASD). ASD uses token embedding vectors from a transformer-based contextualized language model in combination with dynamic programming to measure the similarity between a reference transcription and an ASR hypothesis. We apply ASD and SemDist [12] to evaluate transcriptions from a Norwegian ASR system, and compare the results obtained from both. We find that ASD and SemDist are both highly positively correlated with WER. However, we demonstrate by examples that both metrics put more weight onto errors with high semantic significance, such as proper names. At the same time, they give less emphasis to minor errors related, for example, to spelling variants between Nynorsk and Bokmål. We also show how the proposed ASD metric is more stable than SemDist with respect to the utterance length. For our future work, we plan to perform a correlation study between user-rated transcriptions and ASD.

6. Acknowledgements

This work was carried out within the EEA and Norway Grants project NORDTRANS - Technology for automatic speech transcription in selected Nordic languages.

7. References

[1] I. A. McCowan *et al.*, "On the use of information retrieval measures for speech recognition evaluation," IDIAP, Tech. Rep., 2004.

[2] Y.-Y. Wang, A. Acero, and C. Chelba, "Is Word Error Rate a Good Indicator for Spoken Language Understanding Accuracy," *2003 IEEE Workshop on Automatic Speech Recognition and Understanding (IEEE Cat. No.03EX721)*, pp. 577–582, 2003.

[3] A. C. Morris, V. Maier, and P. Green, "From WER and RIL to MER and WIL: improved evaluation measures for connected speech recognition," *Interspeech 2004*, pp. 2765–2768, 2004.

[4] S. Roy, "Semantic-WER: A unified metric for the evaluation of ASR transcript for end usability," *CoRR*, vol. abs/2106.02016, 2021.

[5] L. van der Werff and W. Heeren, "Evaluating ASR output for information retrieval," *Searching Spontaneous Conversational Speech*, pp. 13–20, 2007.

[6] M. Levit, S. Chang, B. Buntschuh, and N. Kibre, "End-to-end speech recognition accuracy metric for voice-search tasks," *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5141–5144, 2012.

[7] M. A. B. Jannet, O. Galibert, M. Adda-Decker, and S. Rosset, "How to evaluate ASR output for named entity recognition?" In *Proc. Interspeech 2015*, 2015, pp. 1289–1293.

[8] A. Vaswani *et al.*, "Attention is all you need," in *Advances in Neural Information Processing Systems*, I. Guyon *et al.*, Eds., vol. 30, Curran Associates, Inc., 2017.

[9] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," *CoRR*, vol. abs/1810.04805, 2018.

[10] Y. Liu *et al.*, "RoBERTa: A robustly optimized BERT pretraining approach," *CoRR*, vol. abs/1907.11692, 2019.

[11] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence embeddings using siamese BERT-networks," *CoRR*, vol. abs/1908.10084, 2019.

[12] S. Kim *et al.*, "Semantic Distance: A New Metric for ASR Performance Analysis Towards Spoken Language Understanding," *Interspeech 2021*, pp. 1977–1981, 2021.

[13] A. Kutuzov, J. Barnes, E. Velldal, L. Øvrelid, and S. Oepen, "Large-scale contextualised language modelling for Norwegian," in *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, Reykjavik, Iceland (Online): Linköping University Electronic Press, Sweden, 2021, pp. 30–40.

[14] I. Amdal, O. M. Strand, J. Almberg, and T. Svendsen, "RUNDKAST: An annotated Norwegian broadcast news speech corpus," in *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco: European Language Resources Association (ELRA), May 2008.

[15] National Library of Norway. "NB Tale - speech database for Norwegian." (2015), [Online]. Available: https://www.nb.no/sprakbanken/ressurskatalog/oai-nb-no-sbr-31/ (visited on 03/17/2022).

[16] Stortinget. "Stortinget - møte mandag den 10. mai 2021." (2021), [Online]. Available: https://www.stortinget.no/no/Saker-og-publikasjoner/Publikasjoner/Referater/Stortinget/2020-2021/refs-202021-05-10/ (visited on 05/10/2021).

[17] Y. Wu *et al.*, "Google's neural machine translation system: Bridging the gap between human and machine translation," *CoRR*, vol. abs/1609.08144, 2016.

[18] T. Giorgino, "Computing and visualizing dynamic time warping alignments in r: The dtw package," *Journal of Statistical Software*, vol. 31, no. 7, pp. 1–24, 2009.