GLOSS ALIGNMENT USING WORD EMBEDDINGS

Harry Walsh, Ozge Mercanoglu Sincan, Ben Saunders, Richard Bowden

CVSSP, University of Surrey
Guildford, United Kingdom
{harry.walsh, o.mercanoglusincan, b.saunders, r.bowden}@surrey.ac.uk

ABSTRACT

Capturing and annotating Sign language datasets is a time consuming and costly process. Current datasets are orders of magnitude too small to successfully train unconstrained Sign Language Translation (SLT) models. As a result, research has turned to TV broadcast content as a source of large-scale training data, consisting of both the sign language interpreter and the associated audio subtitle. However, lack of sign language annotation limits the usability of this data and has led to the development of automatic annotation techniques such as sign spotting. These spottings are aligned to the video rather than the subtitle, which often results in a misalignment between the subtitle and spotted signs. In this paper we propose a method for aligning spottings with their corresponding subtitles using large spoken language models. Using a single modality means our method is computationally inexpensive and can be utilized in conjunction with existing alignment techniques. We quantitatively demonstrate the effectiveness of our method on the Meine DGS-Annotated (MeineDGS) and BBC-Oxford British Sign Language (BOBSL) datasets, recovering up to a 33.22 BLEU-1 score in word alignment.

Index Terms— Sign Language, Gloss Alignment, Natural Language Processing (NLP), Automatic Dataset Construction

1. INTRODUCTION

Sign languages are the primary form of communication for the Deaf. Signs are expressed through the articulation of manual and non-manual features including body language, facial expressions, mouthing, hand shape, and motion [1]. Despite the recent successes of large language models, Sign Language Translation (SLT) between continuous sign language videos and spoken language remains a challenging task [2]. Even though results have been achieved within a constrained setting and a limited vocabulary [3, 4], progress towards unconstrained translation still requires larger-scale datasets. The visual nature of sign language has restricted the availability of high quality datasets, due to the difficulty of capturing and labeling a visual medium. The publicly available MeineDGS dataset [5] attempts to fully capture the details of the language using gloss[1] annotations, non-manual mouthing and the Hamburg Notation System (HamNoSys). However, the curation of such a high quality dataset is both time consuming and costly, which has restricted its size to only 50k parallel text gloss sequences [5]. This scarcity of data has motivated the research community to automate the collection and annotation of large-scale public datasets.

We also thank the SNSF Sinergia project 'SMILE II' (CRSII5 193686) and the European Union's Horizon2020 research project EASIER (101016982). This work reflects only the authors view and the Commission is not responsible for any use that may be made of the information it contains.

[1]Gloss is the written word associated with a sign

Broadcast content has repeatedly been used as the source of sign language datasets to assist with tasks such as sign language recognition, alignment, and translation [6, 7, 8]. Under the European Accessibility Act, all EU countries are obligated to make content accessible [9]. Specifically, UK broadcasters must supply 5% of their content with British Sign Language (BSL) translations, which leads to the generation of a steady stream of sign language translation data. However, the raw data only contains the spoken language subtitles and the video of the sign interpreter, who, although conducting translations from the subtitles, is often misaligned. In order to make use of this data for tasks such as SLT, the data needs to be curated, and subsequently aligned.

As shown in Fig. 1, we have identified two types of alignment error; 1) Glosses that correspond to the preceding sentence are aligned to the current, shown by the gloss POPULAR and PRAISE that are misaligned to sentence t_1. 2) Glosses are aligned to the following sentence, shown by the gloss INSECT that is misaligned to t_5. There are several factors that lead to the misalignment of the sign to the spoken language subtitle. Firstly, there is a weak correlation between the number of words in a sentence and the number of signs contained in the translation. Additionally, the time taken to speak a word is not related to the time taken to perform a sign. Finally, the ordering of spoken language words is different from the gloss order [1]. All these factors result in the sign language lagging or preceding the corresponding subtitle.

Previous work has attempted to align the subtitles with the sign language video by finding a correspondence between the glosses of the spotted isolated signs and words in the subtitle with a similar lexical form [10, 11, 12]. However, these works all require multimodal inputs and are expensive to compute.

In this paper, we propose an approach to align glosses to the corresponding spoken language sentence by leveraging the power of large spoken language models, such as BERT [13] and Word2Vec [14]. We make the following 2 contributions: 1) A novel alignment approach that can be used in conjunction with previous methods. 2) Quantitative evaluation of our approach on two different datasets from both BSL and German Sign Language - Deutsche Gebärdensprache (DGS), demonstrating our approach is language agnostic.

2. RELATED WORK

2.1. Spoken Language Alignment

Word alignment techniques have been researched since the late 20$^{\text{th}}$ century, where Brown et al. developed the IBM models to assist in statistical machine translation [15]. Since then a number of statistical word alignment techniques have been proposed, such as GIZA++ [16, 17] or alignment via a hidden Markov model [18]. More re-

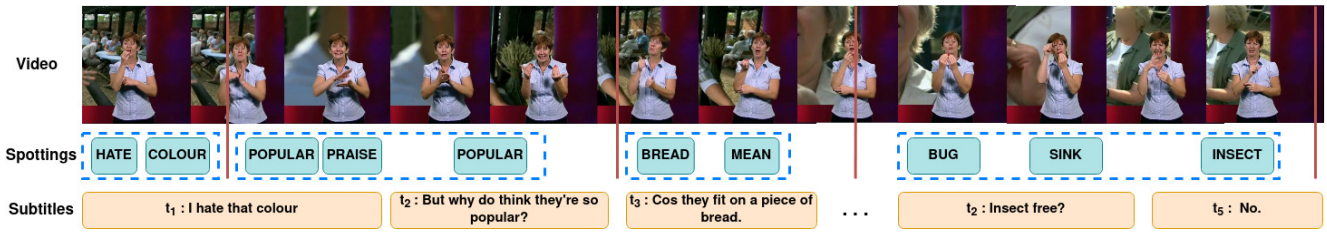

Fig. 1. A visualisation of the alignment from the BBC-Oxford British Sign Language (BOBSL) dataset (vertical red lines indicate the correct boundaries).

cently, deep learning based methods have demonstrated superior performance [19]. A variety of supervised methods have been created, some using statistical supervision [20], while others make use of the attention mechanism from a transformer [21]. Most similar to our approach, Stengel-Eskin et al. used the dot product distance between learnt embeddings to make an alignment prediction [19].

2.2. Sign Language Spotting

Sign spotting is the task of locating isolated instances of signs in a continuous video. Several methods have been suggested to tackle this task, from early techniques that use hand crafted features [22, 23], to methods that employ subtitles as a form of weak supervision [7, 8]. More recent methods have employed multiple modalities to improve performance e.g. visual dictionaries [11] and mouthings [10]. However, all these methods still result in the misalignment of spotted signs and the subtitles, as shown in Fig. 1.

2.3. Subtitle Alignment

Subtitle alignment attempts to align a continuous sequence of signing to the corresponding subtitles. Early attempts to solve the alignment issues used 3D pose in a multi step approach, but assumed a similar ordering between the spoken and signed languages [24]. To overcome this assumption, Bull et al. trained a Bidirectional Long Short-Term Memory (BLSTM) with 2D keypoints using manually aligned subtitles as ground truth to segment a continuous video into candidate signs [25]. However, without a strong language model, such approaches tend to over segment the video. In subsequent works, the subtitles were incorporated into the input of the model along with the video and shifted temporal boundaries, to align broadcast footage [26]. In contrast, in this work we attempt to align spotted glosses to the spoken language subtitles using only word embeddings. Note our approach can be used in conjunction with these existing methods.

3. METHODOLOGY

In this section we explain our methodology for aligning glosses with their corresponding subtitles. In Section 3.1 we explain how we use the embeddings from large spoken language models such as BERT and Word2Vec to create a mapping between a sequence of glosses and spoken language words. Then in Section 3.2 we show how to use the mapping to re-align glosses to the correct spoken language sentence.

3.1. Text Gloss Mapping

Our alignment approach relies on the lexical overlap that exists between the spoken language words and the signed glosses. Therefore, the gloss notation needs to be semantically motivated. Given the following example text, “where do you live?” and the following sequence of glosses, “YOU LIVE WHERE. ME LONDON”. It is clear to see that the first three glosses correspond to the given text and the last two glosses potentially correspond to the next sentence.

Following this intuition, we use two different word embedding techniques to find which glosses best correspond to a given spoken language sentence. Firstly, we use Word2Vec [13] to find connections between words and glosses that have a similar lexical form. Secondly, we use BERT [14] to find connections based on meaning. We find BERT embeddings capture the meaning of words allowing us to find connections between words and glosses that have a different lexical form, e.g. ”supermarket” and ”SHOP”. Note when we apply our approach to DGS we first apply a compound splitting algorithm to improve the performance [27].

To find a mapping between a spoken language sequence $X = (x_1, x_2, ..., x_W)$ with W words, and a sequence of glosses, $Y = (y_1, y_2, ..., y_G)$ with G glosses, we first apply Word2Vec;

$$X_{Vec} = Word2Vec(X) \tag{1}$$

$$Y_{Vec} = Word2Vec(Y) \tag{2}$$

where $X_{Vec} \in \mathbb{R}^{W \times E_{vec}}$ and $Y_{Vec} \in \mathbb{R}^{G \times E_{vec}}$. Calculating the outer product between the two embeddings produces the Word2Vec alignment;

$$A_{Vec} = Y_{Vec} \otimes X_{Vec} \tag{3}$$

where $A_{Vec} \in \mathbb{R}^{G \times W}$. We repeat the above using BERT, to find connections based on meaning;

$$X_{BERT} = BERT(X) \tag{4}$$

$$Y_{BERT} = BERT(Y) \tag{5}$$

$$A_{BERT} = Y_{BERT} \otimes X_{BERT} \tag{6}$$

where $X_{BERT} \in \mathbb{R}^{W \times E_{BERT}}$, $Y_{BERT} \in \mathbb{R}^{G \times E_{BERT}}$ and $A_{BERT} \in \mathbb{R}^{G \times W}$. The BERT model we apply uses a word-piece tokenizer, therefore to find an alignment on the word level we average the embeddings of the sub-units. The final alignment is found by joining the alignments from BERT and Word2Vec. We filter the Word2Vec alignment scores by α, only keeping strong connections. Thus, the final alignment is defined as;

$$Align(X, Y) = A_{BERT} + (\alpha * A_{Vec}) \tag{7}$$

where $A \in \mathbb{R}^{G \times W}$. A visualization of two alignments, A, is shown in Fig. 2. Here we find the alignment between two sequential text

sentences from the BOBSL dataset, and the four glosses that correspond to the two sentences. FOOD and FACTORY belong to the first sentence "I've set up my own food factory inside this barn", and as shown by Fig. 2 (left) a strong alignment is found between the spoken language and it's corresponding glosses, FOOD, and FACTORY. The same applies to Fig. 2 (right) with the gloss BREAD.

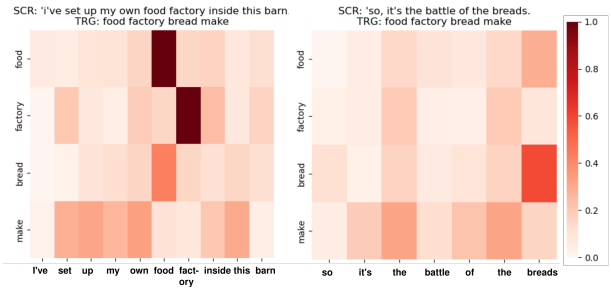

Fig. 2. An example of two alignments, A, found between the TRG: "FOOD FACTORY BREAD MAKE" and sentence one (left): "I've set up my own food factory inside this barn", and sentence two (right): "So, it's the battle of the breads.".

3.2. Gloss Alignment

We use the alignment found above for two sequential sentences to re-align glosses to their corresponding subtitles. Given a dataset that consists of N sequences of spoken language sentences $T = (X_1, X_2, ..., X_N)$ and N sequences of glosses $S = (Y_1, Y_2, ..., Y_N)$, we take two sequential sentences, X_i, X_{i+1}, and we concatenate their corresponding glosses $Y_{i:i+1} = Y_i + Y_{i+1}$.

We want to find the index to split $Y_{i:i+1}$ back into two sequences that result in the best alignment. The number of possible splits is $N_{split} = length(Y_{i:i+1}) + 1$. Therefore, for each possible split we sum the max alignment score for each gloss in $Align(X_i, Y_{i:i+1})$ and $Align(X_{i+1}, Y_{i:i+1})$ as in Equation 7. Fig. 3 shows the alignment score for each possible split of $Y_{i:i+1}$. We take the argmax of the alignment score to determine the optimal index to split $Y_{i:i+1}$ back to two sequences.

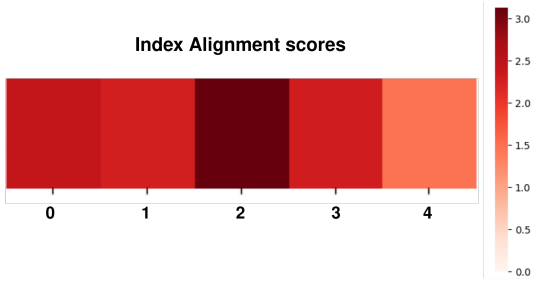

Fig. 3. The alignment scores found for the two sentences in Fig. 2

Fig. 3 shows the alignment score for the two sentences in Fig. 2, showing our approach is able to find the optimal alignment.

The proposed algorithm is auto-regressive, meaning the output of the first split affects the next iteration. This introduces a bias that favours earlier sentences in the dataset. Therefore, to counter this effect we iterate through the data from $i = [0, 1, 2, ...N]$ and then for each subsequent iteration we reverse the order, such that iteration two is $i = [N, N-1, N-2, ..., 0]$. In the next section we show that multiple iterations (forwards then backwards) of the algorithm increase the alignment score, but quickly converges.

4. EXPERIMENTAL SETUP

In this section we outline the experimental setup, detailing the pre-trained models that we use to create the word embeddings for both English and German. In Section 4.1 we describe how we corrupt the MeineDGS dataset to simulate a spotting misalignment. Finally, in Section 4.2 we explain how we gather the spottings from [26] and process them to create parallel text gloss sequences for our alignment algorithm.

We use the Fasttext implementation of Word2Vec that supports 157 languages [28]. The models are trained on the Common Crawl and Wikipedia datasets and have an output dimension of 300. For the following experiments we use the English implementation when testing our approach on the BOBSL dataset and the German version when testing on MeineDGS. Note, we filter the Word2Vec embeddings by setting α to 0.9.

When creating embeddings with BERT we use Huggingface's python library transformers to load the models. When testing on MeineDGS we use Deepsets implementation of German BERT [29], which is trained on approximately 12GB of data from the Wiki, OpenLegalData, and News datasets. Finally, when testing on the English BOBSL dataset we use GoogleAI's implementation of BERT [30], which is trained on the Bookcorpus and Wikipedia datasets. To evaluate the performance of our algorithm on all datasets we use BLEU-1 score. We do not present results using higher n-gram BLEU scores as these metrics are used to measure the order accuracy, that is unnecessary for this task.

4.1. MeineDGS Dataset

All results on the MeineDGS dataset are computed against the original ground truth. The MeineDGS dataset contains 50k parallel sequences [5] and we follow the translation protocol set in [31]. The dataset has a source vocabulary of 18,457 with 330 deaf participants performing free form signing. Note we reorder the sequences sequentially as in the original videos.

To evaluate our approach we corrupt the MeineDGS dataset. This allows us to simulate an alignment error created when using previously mentioned sign spotting techniques to automatically spot glosses in a sequence of continuous signing. We create two versions of the dataset to simulate;

4.1.1. Sequence misalignment

A worst-case scenario, a total misalignment of all sequence pairs. For this, we offset all the gloss sequences by one. We add an empty sequence to the start Y_{empty} and remove the last sequence Y_N to maintain an equal number, N, of text gloss pairs. Therefore, we apply our alignment approach to $T = (X_1, X_2, ..., X_N)$ and $S = (Y_{empty}, Y_1, ..., Y_{N-1})$.

4.1.2. Gloss misalignment

To simulate the errors shown in Fig. 1 (glosses are misaligned to the preceding or succeeding sentence) we randomly shift up to 3 glosses to the previous or following sequence. We set probabilities of 15%, 20% and 10% of moving 1, 2 or 3 glosses, respectively. 10% of the time we do not alter the sequence. Note if the sequence has fewer

glosses than we wish to shift, then we do not alter it. In total we move 21,273 glosses to the preceding sequence and 21,359 to the next sequence.

4.2. BOBSL Dataset

The BOBSL dataset contains 1,193K sentences extracted from 1,962 videos from 426 different TV shows [6]. The dataset itself only contains the subtitle from the original TV show and the signer. The test set comes with three variants of the subtitles audio-aligned, audio-aligned shifted, and manually aligned. By calculating the average difference between the audio-aligned and signing-aligned sentences, Albanie et al. [6] found the signer lags the subtitle by approximately +2.7 seconds. Thus, the audio-shifted variant applies this 2.7 second delay to all time stamps. The manually aligned subtitles contain a subset of the original audio-aligned subtitles. Therefore, when comparing our alignment results against the manually aligned subtitles we are restricted to this subset of the data.

In order to perform alignment we use the automatically extracted spottings from [32]. We then process the data, aligning the dense spots with the three variants of the subtitle provided in the original BOBSL test set.

5. EXPERIMENTS - QUANTITATIVE EVALUATION

Fig. 4 shows the results of applying the alignment algorithm to two versions of the MeineDGS and BOBSL datasets. As can be seen, the algorithm has a positive effect on all variants, increasing the BLEU-1 scores by up to 33.22. Next we discuss the results in detail, starting with MeineDGS (sequence level misalignment, then gloss level misalignment), followed by the BOBSL dataset (audio aligned, then manually aligned).

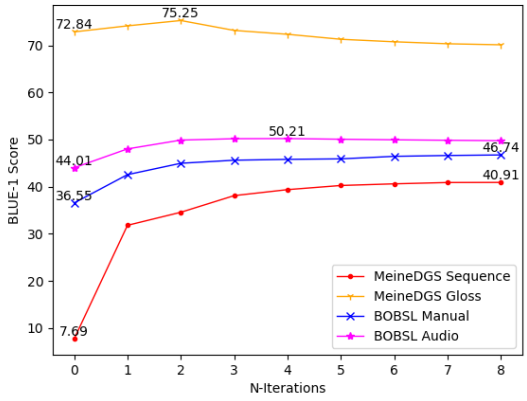

Fig. 4. Results of applying the alignment algorithm to the MeineDGS (Gloss and Sequence level misalignment) and BOBSL datasets (Audio aligned and Manually aligned)

5.1. MeineDGS Alignment

5.1.1. Sequence misalignment

In this experiment we offset the MeineDGS dataset by 1 sequence, to simulate the worst case where all glosses are misaligned. Fig. 4 - MeineDGS Sequence (orange line) shows there is a shared gloss vocabulary between sequential sentences, as the baseline score is not zero. Impressively, the approach is able to recover a large proportion of the glosses, increasing the BLEU-1 score from 7.69 to 40.91, a improvement of 432%.

5.1.2. Gloss misalignment

Fig. 4 - MeineDGS Gloss (yellow line) shows the results of applying our alignment approach to the corrupted dataset. By corrupting the data we decrease the BLEU-1 score from perfect alignment (100 BLEU-1) to 72.84. From this baseline we are able to recover 2.41 BLEU-1 score using a single forward and backward pass through the data, an improvement of 3.3%. However, further iterations are detrimental as can be expected from a greedy algorithm.

This shows that the approach is able to recover a portion of the corruption. However, it should be noted that the effectiveness of the approach is dataset dependent, as the similarity of sequential sentences will effect the reliability of the mapping found between words and glosses.

5.2. BOBSL Alignment

5.2.1. Audio aligned

Here we show that our approach is able to move the audio-aligned subtitle toward the improved audio-shifted subtitles. As shown in Fig. 4 the approach improves the audio alignment by 6.2 BLEU-1. It should be noted that the audio-shifted subtitles are not perfect ground truth. Additionally, the spottings are not perfect, which introduces an error to any alignment approach as we may be attempting to map glosses that do not align with any words in the spoken sentence. Thus, we could expect the performance to increase if the quality of the underlying spottings improves.

5.2.2. Manually aligned

The manually aligned subtitles and their timings affect how we collect the spottings, which leads to a variation in the number of glosses. Hence, why the baseline score at $N = 0$ is lower compared to the previous audio-aligned experiment. Despite this limitation, the approach is able to improve the alignment by 10.19 BLEU-1.

6. CONCLUSION

Sign Language alignment is an essential step in creating large-scale datasets from raw broadcast data. Improving the alignment between the subtitles and the associated signed translation would have positive effects on tasks such as translation, recognition, and production. In this paper we have demonstrated that embeddings from large spoken language models can be used to align glosses with their corresponding subtitles. Our approach can be run in addition to existing multi-model methods and is computationally inexpensive in comparison. We have shown the approach is capable of recovering up to a 33.22 BLEU-1 score in word alignment.

7. ACKNOWLEDGMENT

We thank Adam Munder, Mariam Rahmani, and Marina Lovell from OmniBridge, an Intel Venture, for supporting this project. We also thank Thomas Hanke and the University of Hamburg for use of the MeineDGS data.

8. REFERENCES

[1] Rachel Sutton-Spence and Bencie Woll, *The linguistics of British Sign Language: an introduction*, Cambridge University Press, 1999.

[2] Danielle Bragg, Oscar Koller, Mary Bellard, Larwan Berke, Patrick Boudreault, Annelies Braffort, Naomi Caselli, Matt Huenerfauth, Hernisa Kacorri, Tessa Verhoef, et al., "Sign language recognition, generation, and translation: An interdisciplinary perspective," in *ACM SIGACCESS*, 2019, pp. 16–31.

[3] Necati Cihan Camgoz, Simon Hadfield, Oscar Koller, Hermann Ney, and Richard Bowden, "Neural sign language translation," in *CVPR*, 2018, pp. 7784–7793.

[4] Necati Cihan Camgoz, Oscar Koller, Simon Hadfield, and Richard Bowden, "Sign language transformers: Joint end-to-end sign language recognition and translation," in *Computer Vision and Pattern Recognition*, 2020, pp. 10023–10033.

[5] Reiner Konrad, Thomas Hanke, Gabriele Langer, Dolly Blanck, Julian Bleicken, Ilona Hofmann, Olga Jeziorski, Lutz König, Susanne König, Rie Nishio, Anja Regen, Uta Salden, Sven Wagner, Satu Worseck, Oliver Böse, Elena Jahn, and Marc Schulder, "Meine dgs – annotiert. öffentliches korpus der deutschen gebärdensprache, 3. release / my dgs – annotated. public corpus of german sign language, 3rd release," 2020.

[6] Samuel Albanie, Gül Varol, Liliane Momeni, Hannah Bull, Triantafyllos Afouras, Himel Chowdhury, Neil Fox, Bencie Woll, Rob Cooper, Andrew McParland, and Andrew Zisserman, "BOBSL: BBC-Oxford British Sign Language Dataset," 2021.

[7] Helen Cooper and Richard Bowden, "Learning signs from subtitles: A weakly supervised approach to sign language recognition," in *CVPR*. IEEE, 2009, pp. 2568–2574.

[8] Patrick Buehler, Andrew Zisserman, and Mark Everingham, "Learning sign language by watching tv (using weakly aligned subtitles)," in *CVPR*. IEEE, 2009, pp. 2961–2968.

[9] "Directive (eu) 2019/882 of the european parliament and of the council of 17 april 2019 on the accessibility requirements for products and services (text with eea relevance) l 151/70," *OJ*, vol. L 151, pp. 70–115, 7.6.2019.

[10] Samuel Albanie, Gül Varol, Liliane Momeni, Triantafyllos Afouras, Joon Son Chung, Neil Fox, and Andrew Zisserman, "Bsl-1k: Scaling up co-articulated sign language recognition using mouthing cues," in *ECCV*. Springer, 2020, pp. 35–53.

[11] Liliane Momeni, Gul Varol, Samuel Albanie, Triantafyllos Afouras, and Andrew Zisserman, "Watch, read and lookup: learning to spot signs from multiple supervisors," in *Asian Conference on Computer Vision*, 2020.

[12] Gul Varol, Liliane Momeni, Samuel Albanie, Triantafyllos Afouras, and Andrew Zisserman, "Read and attend: Temporal localisation in sign language videos," in *CVPR*, 2021.

[13] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv:1810.04805*, 2018.

[14] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean, "Distributed representations of words and phrases and their compositionality," *Advances in Neural Information Processing Systems*, 2013.

[15] Peter F Brown, Vincent J Della Pietra, Peter V Desouza, Jennifer C Lai, and Robert L Mercer, "Class-based n-gram models of natural language," *Computational linguistics*, vol. 18, 1992.

[16] Franz Josef Och and Hermann Ney, "Improved statistical alignment models," in *Computational Linguistics*, 2000.

[17] Qin Gao and Stephan Vogel, "Parallel implementations of word alignment tool," in *Software engineering, testing, and quality assurance for natural language processing*, 2008.

[18] Stephan Vogel, Hermann Ney, and Christoph Tillmann, "Hmm-based word alignment in statistical translation," in *COLING 1996 Volume 2: Computational Linguistics*, 1996.

[19] Elias Stengel-Eskin, Tzu-ray Su, Matt Post, and Benjamin Van Durme, "A discriminative neural model for cross-lingual word alignment," *arXiv preprint arXiv:1909.00444*, 2019.

[20] Akihiro Tamura, Taro Watanabe, and Eiichiro Sumita, "Recurrent neural networks for word alignment model," in *Computational Linguistics (Volume 1)*, 2014, pp. 1470–1480.

[21] Thomas Zenkel, Joern Wuebker, and John DeNero, "Adding interpretable attention to neural translation models improves word alignment," *arXiv preprint arXiv:1901.11359*, 2019.

[22] Hee-Deok Yang, Stan Sclaroff, and Seong-Whan Lee, "Sign language spotting with a threshold model based on conditional random fields," *IEEE transactions on pattern analysis and machine intelligence*, vol. 31, no. 7, pp. 1264–1277, 2008.

[23] Pinar Santemiz, Oya Aran, Murat Saraclar, and Lale Akarun, "Automatic sign segmentation from continuous signing via multiple sequence alignment," in *ICCV Workshops*, 2009.

[24] Iva Farag and Heike Brock, "Learning motion disfluencies for automatic sign language segmentation," in *ICASSP*, 2019.

[25] Hannah Bull, Michèle Gouiffès, and Annelies Braffort, "Automatic segmentation of sign language into subtitle-units," in *Computer Vision–ECCV*. Springer, 2020, pp. 186–198.

[26] Hannah Bull, Triantafyllos Afouras, Gül Varol, Samuel Albanie, Liliane Momeni, and Andrew Zisserman, "Aligning subtitles in sign language videos," in *International Conference on Computer Vision*, 2021, pp. 11552–11561.

[27] Don Tuggener, *Incremental coreference resolution for German*, Ph.D. thesis, University of Zurich, 2016.

[28] Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov, "Learning word vectors for 157 languages," in *Language Resources and Evaluation*, 2018.

[29] Branden Chan, Stefan Schweter, and Timo Möller, "German's next language model," in *Proceedings of the 28th International Conference on Computational Linguistics*, Barcelona, Spain (Online), Dec. 2020, pp. 6788–6796, International Committee on Computational Linguistics.

[30] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," *CoRR*, vol. abs/1810.04805, 2018.

[31] Ben Saunders, Necati Cihan Camgoz, and Richard Bowden, "Signing at Scale: Learning to Co-Articulate Signs for Large-Scale Photo-Realistic Sign Language Production," in *Computer Vision and Pattern Recognition*, 2022.

[32] Liliane Momeni, Hannah Bull, KR Prajwal, Samuel Albanie, Gül Varol, and Andrew Zisserman, "Automatic dense annotation of large-vocabulary sign language videos," in *Computer Vision–ECCV 2022*. Springer, 2022, pp. 671–690.