SEAS: Sentence Extraction and Alignment from Subtitles

Josh Stephenson and Libby Barak
Montclair State University
New Jersey, USA
josh@brightmediums.com, barakl@montclair.edu

Abstract

Subtitles for movies and television offer a unique resource for extracting multilingual alignments and studying linguistic phenomena. However, existing tools for subtitle alignment are difficult to reproduce, lack standardized evaluation mechanisms, and frequently exhibit obvious errors. Moreover, these methods fail to leverage recent advances in sentence alignment, namely sentence embeddings. To address these shortcomings, we combine time-based methods with state-of-the-art sentence alignment techniques. Our method achieves substantial improvements in alignment performance over gold-standard data and a downstream Machine Translation task. Our submission includes a curated corpus of gold-standard alignments for English-Spanish and English-German subtitles, along with their corresponding subtitle files, a novel annotation tool, and the full code to reproduce our method.

1 Introduction

Subtitles are one of the most substantial sources of translation data created through manual annotation. Notably, they complement primary data sources such as books, web and legal data, which are commonly used to train language models, by providing rich cross-lingual spoken dialogue. Importantly, subtitle alignment presents multiple challenges due to the lack of one-to-one correspondence between utterances and subtitles across languages, substantial noise from non-dialogue elements, and variability in subtitle counts between source and target.

Previous subtitle extraction and alignment projects have produced large collections of sentence alignments from subtitle data (Tiedemann, 2007, 2012, 2016). However, past subtitle alignment methods have relied too heavily on timestamp data rather than the semantic properties of language and have employed less stringent data cleansing procedures. As a result, these datasets contain obvious errors. Unfortunately, subtitle evaluation has largely been limited to small samples of manually annotated sentence pairs, restricting the ability to quantify the impact of such errors on downstream tasks.

English: `- Stop.`
Spanish: `Agrándalo y vuélvelo a reproducir.`

English: `You're okay.`
Spanish: `Estás bien. Descansa.`

English:
German: `- Was ist los?\n- Treten Sie von ihr...`

Table 1: Sample errors from the English-Spanish and English-Germancorpus from Opus 2018.

Recent breakthroughs in sentence alignment using sentence embeddings have demonstrated significant progress (Schwenk and Douze, 2017). However, to the best of our knowledge, these advances have yet to be applied to subtitle data, either alone or in combination with time-based methods traditionally used for subtitle alignment. We propose a novel two-step method for subtitle alignment: (1) Preprocessing, which includes timecode drift correction, partitioning using dialogue gaps, and other data cleansing steps, and (2) Embedding-based alignment, which leverages Vecalign (Thompson and Koehn, 2019), a sentence embedding framework, to align extracted sentences based on normalized cosine similarity.

Our system achieves 93.12% F1 scores on a manually annotated sample of Spanish and German subtitles, consisting of roughly $2,950$ alignments for each language. Additionally, our results demonstrate significant improvements over previous subtitle datasets in a downstream Machine Translation (MT) task, using one million English-Spanish alignments generated by our method. Our submission includes the complete code for the model used to generate the data, the annotated and raw data, along with a novel annotation tool we developed to curate manual alignments.

Proceedings of the 21st Conference on Natural Language Processing (KONVENS), Volume 1: Long and Short Papers, pages 203–212, 2025

2 Related Work

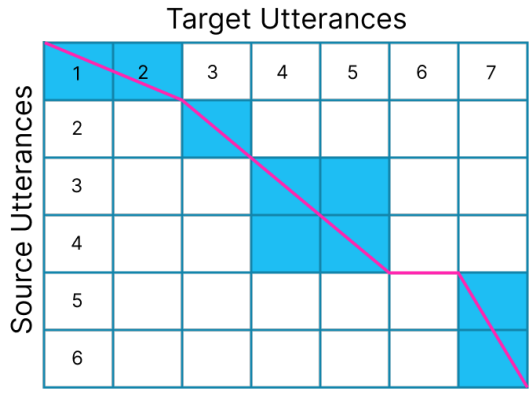

Figure 1: Visualization of a sentence alignment matrix.

The core challenges in subtitle alignment are largely rooted in sentence alignment, which involves constructing a bipartite graph from parallel documents segmented at sentence boundaries. All methods have two main components: (1) a **scoring function**, which estimates the likelihood that two sentences are translations of one another, and (2) an **alignment algorithm**, which generates a hypothesized alignment between two documents using the scoring function.

Any document is essentially an ordered list of sentences, and aligning two documents starts by representing them as a large two-dimensional matrix, with the sentences of one document along the columns and those of the other along the rows (see Figure 1).

Previous sentence alignment methods have used various scoring functions, but most traditionally relied on sentence length in characters or tokens, while others incorporated lexical information (Brown et al., 1991; Gale and Church, 1993; Chen, 1993; Wu, 1994; Moore, 2002; Varga et al., 2007). Nearly all have employed some form of Dynamic Programming (DP) as the alignment algorithm (Bellman, 1954). Later methods introduced strategies such as clustering, pruning, and lexical anchoring to improve results (Braune and Fraser, 2010; Sennrich and Volk, 2010; Papineni et al., 2002).

The distinctive nature of subtitle translation introduces unique alignment challenges due to frequent insertions, deletions, and paraphrasing. Without extensive subtitle-specific preprocessing, common sentence alignment methods produce a high error rate, even when incorporating sentence-length or utterance-duration information (Tiedemann, 2007). Each subtitle has a corresponding start and end time which can be used for alignment instead of using DP. Notably, Tiedemann found that doing so produced superior results. His algorithm searches linearly through the target subtitles based only on overlapping timecodes to resolve insertions and deletions (see Table 2).

Sentence embeddings have addressed many of the key shortcomings of earlier alignment techniques. Thompson and Koehn (2019) applied multilingual sentence embeddings from the LASER embedding toolkit (Artetxe and Schwenk, 2018) to sentence alignment, using normalized cosine similarity as the scoring function and an alignment algorithm based on Fast Dynamic Time Warping (Salvador and Chan, 2007). This approach achieved state-of-the-art results in both alignment quality and efficiency. The method, Vecalign, first concatenates n adjacent sentences into a single unit (where n is a hyperparameter) before feeding them into the algorithm, enabling one-to-many and many-to-many alignments.

Despite these advances, subtitle alignment tools have not kept pace with recent developments in sentence alignment. The OPUS dataset contains over 8.5 billion sentence alignments across sixty-two languages (Tiedemann, 2007). In later work, the dataset has been extended and updated while maintaining the original method. Unfortunately, it lacks extensive evaluation procedures and was created before the introduction of sentence embeddings. Its evaluation consisted of only thirty alignments from each of five titles in two language pairs.

Our study overcomes these limitations by releasing an open-source suite of tools built on sentence embeddings and enabling comprehensive evaluation.

3 Methods

Unlike corpora such as books, websites, manuals and legal documents, subtitle documents contain more than just dialogue, which necessitates meticulous preprocessing prior to alignment.

Our tools include various preprocessing steps before alignment, which we have conceptually separated into three stages: (1) data cleansing, (2) sentence boundary detection, and (3) partitioning. For data cleansing, we first verify correct language labeling using n-grams (unigrams up to 5-grams)[1], detect and correct synchronization issues between

[1] mislabeled files are thrown out.

English (source)		Spanish (target)	
19	00:01:10,320 --> 00:01:15,617 Aren't you a professor of Physics?	12	00:01:10,446 --> 00:01:14,784 ¿No es usted Profesor de física?
95	00:07:06,342 --> 00:07:07,593 {\an8}[radio beeps]		
21	00:01:20,038 --> 00:01:21,915 You were my student.	14	00:01:20,498 --> 00:01:23,126 Fuiste mi estudiante.
22	00:01:22,123 --> 00:01:23,123 Behave yourself!		Compórtese.
25	00:01:25,835 --> 00:01:30,882 In your physics course, did you teach the theory of relativity?	16	00:01:25,628 --> 00:01:27,305 En su curso de física,
		17	00:01:27,505 --> 00:01:30,591 ¿enseñó la teoría de la relatividad?
543	00:36:40,531 --> 00:36:42,408 But I don't see it that way.	349	00:36:40,491 --> 00:36:41,834 No lo veo de esa forma.
		350	00:36:42,034 --> 00:36:45,329 Conozco tus logros académicos.

Table 2: English-Spanish alignment examples, (1) 1-to-1 noise free alignment; (2) English subtitle without dialogue; (3) 2 English sentences from 2 speakers aligned with 1 Spanish sentence; (4) 1 English sentence in 1 subtitle aligned with 1 Spanish sentence spanning 2 subtitles; (5) 1 English subtitle overlaps with 2 Spanish subtitles in 2 sentences.

source and target subtitle timecodes, and then use a multi-step process based on regular expressions to strip everything but dialogue. After cleansing, we apply regular expressions to detect sentence boundaries and extract utterances.

Finally, partitioning is a novel technique for splitting the subtitle files into smaller files based on gaps in the dialogue. In essence this leverages the time-based metadata of subtitles selectively, ensuring the semantic merging and alignment does not merge fragments across prosodic boundaries. Prior to this partitioning, our analysis of false positives revealed errors that involved merging one character's complete thought with another character's verbal filler. Because verbal fillers (e.g., "umm," "uh,") carry minimal semantic weight, Vecalign may fail to recognize them as separate utterances, leading to unintended merges. Often these merges occur across a clear prosodic boundary as evident by the timecodes. To mitigate these errors, our final stage of preprocessing uses a 3-second threshold for prosodic boundaries during partitioning, based on our analysis of gap length and psycholinguistic findings on prosodic boundary duration (Jun, 2006). See Appendix for full details on preprocessing.

Finally we use Vecalign for alignment of extracted utterances (Thompson and Koehn, 2019). Since some utterances are short phrases that need to be merged with adjacent ones, we overlap them in groups of six before inputting them into Vecalign.

4 Evaluation and Results

We evaluate our proposed subtitle alignment method, SEAS, against manually annotated alignments and on a downstream MT task. As a first baseline, we implement a timecode-only system to replicate the method used to generate the OPUS dataset (Tiedemann, 2007; Lison and Tiedemann, 2016)). To evaluate how well Vecalign can handle sentence alignment when subtitle data is not well preprocessed, we also produce baselines that include each of the preprocessing methods one-by-one to embedding-based alignment to assess the impact of each individual component.

	Preprocessing			Recall	Precision	F-1
	1	2	3			
0	-	-	-	66.67	58.26	62.18
1	✘	✘	✘	33.31	41.87	37.11
2	✓	✘	✘	45.07	52.40	48.46
3	✘	✓	✘	76.30	72.10	74.14
4	✓	✓	✘	93.01	92.08	92.54
5	✓	✓	✓	93.80	92.45	**93.12**

Table 3: System configuration details w/o preprocessing steps detailed in section 3. Recall, precision and F1 as percentage.

4.1 Alignment quality assessment

We report performance across five titles containing Spanish and English subtitles for each of the baselines described above. For each predicted alignment, we search linearly for an exact match in the gold standard alignments. Although some sentence alignment tasks allow for soft alignments based on edit distance (Buck and Koehn, 2016), we find that speech-based alignment errors regularly exceeded the 5% tolerance for soft recall, so we only look for exact matches.

Table 3 presents the results for Spanish. We observe overall similar results on German as reported in Appendix A. Our first baseline, timecode-only, yields an F1 score of 62.18% without the use of sentence embeddings or DP as part of the methods. In contrast, our embeddings-only systems (Table 7), achieves an F1 score of just 37.1%. Cleansing alone (i.e., preprocessing 1) improves the F1 score by approximately 11% over raw subtitles, while sentence boundary detection (preprocessing 2) alone yields a 37% improvement. When both cleansing and sentence boundary detection are applied together, performance rises to 92.54% (about 55% increase). Finally, we compare the results using additional preprocessing of time-based partitioning on dialogue gaps of 3 seconds. As can be seen, this step results in only a marginal improvement of a further increase to 93.12% (see Limitations for additional discussion).

4.2 Downstream Machine Translation Task

Dataset	chrF2	TER	BLEU
OPUS	39.8	70.3	14.6
SEAS	45.3	68.4	19.0

Table 4: BLEU scores for equivalent transformer models trained on OPUS versus SEAS.

As a second evaluation step, we train a transformer translation model using fairseq[2] on 1 million English-Spanish alignments from OPUS and compare it against the same model trained on 1 million alignments from the data generated by SEAS[3]. We then evaluated the translation quality of these two machine translation models against the English-Spanish corpus from the Workshop on Statistical Machine Translation (WMT) 2013 using sacrebleu (Bojar et al., 2013)[4]. Comparing the BLEU scores, our model outperforms the OPUS-trained language model by a score of 4.4, an improvement of 23%. For full comparison, we also report two other common metrics for translation quality for both data sets: chrF2 and TER (see Table 4).

5 Discussion

In this paper, we present a novel method, **SEAS**, for subtitle alignment. Previous work has either focused on timecodes which are unique to subtitle data or do not leverage sentence embeddings. We present here the first application of sentence embeddings to subtitle content. Moreover, we evaluate our method on both manually annotated data and a downstream task showing significant improvements on both. As shown in our analysis, this data can offer significant improvements as training data for tasks that require multilingual aligned text, such as machine translation. We hypothesize that additional models can benefit from the unique properties of aligned text that carries informal language, diverse topics, and conversational data directed at a wide range of ages.

In this work, we addressed key challenges in subtitle alignment, focusing on improving the quality of extracted multilingual sentence alignments through tailored preprocessing and state-of-the-art sentence embedding techniques. With our procedures for cleansing, sentence boundary detection, and partitioning, our system demonstrates substantial improvements in alignment quality, achieving F1 of 93.12% on English-Spanish and 92.55% on English-German. These results underscore the importance of preprocessing steps tailored specifically for subtitle data, which address unique challenges such as fragmented sentences, language misclassification, noise, and timing discrepancies.

Our current evaluation has focused on English, Spanish, and German. In future work, we aim to extend our analysis to languages with less rigid sentence structure or different writing systems. We also contribute a novel annotation tool that facilitates the creation of high-quality gold-standard alignments, enabling further research on dialogue-based corpora. By releasing a small but meticulously curated dataset of English-Spanish and English-German alignments, we aim to bridge the gap in resources for evaluating alignment systems

[2]https://github.com/facebookresearch/fairseq

[3]Precisely, 1,026,231 alignments for both datasets.

[4]https://github.com/mjpost/sacrebleu/

and fostering advancements in this field.

Code and Data Availability

All code and data required to replicate this research and extended the data, including the gold-labeled subtitles and the annotation tool, are available on Github[5]. The authors intend to continue refining the annotation tool and welcome any open-source contributions.

Limitations

Although this study offers significant extensions to previous evaluation methods, our corpus of manual annotation is limited in size and language scope. For energy conservation, we limited our evaluation to allow for through inspection while minimizing the energy footprint of the full-scale of data.

In this work, we use Vecalign for the embedding-based method Thompson and Koehn (2019). However, Steingrímsson et al. (2023) present a similar approach, SentAlign, including several adjustments, such as incorporating Gale and Church's algorithm for optimizing documents that exceed a length threshold and using LaBSE for sentence embeddings rather than LASER (Feng et al., 2020). Although SentAlign demonstrated better performance than Vecalign on certain test datasets, we observed that it was less effective for subtitle alignment. Due to space limitations, we omit this analysis from our results. Future research could investigate adjusting hyperparameters and configurations to enhance the performance of SentAlign with subtitle data specifically.

We recognize the importance of subtitle-specific preprocessing to alignment quality. However, our sentence boundary detection method can be improved in several ways, which we hope to address in the future. The common use of insertions, deletions, and paraphrasing results in variability that complicates the use of time gaps between dialogue for determining sentence boundaries. Furthermore, reliance on punctuation is effective for certain languages, particularly Western languages, but proves inadequate for others. For sentence boundary detection, we utilized the regular expressions detailed in Appendix. While these expressions are effective, they are not comprehensive and may be insufficient for non-Western languages that do not employ traditional Western punctuation for sentence boundaries. Though, for English, Spanish, and German, this approach outperforms both NLTK and SpaCy alone on this dataset. Future improvements should focus on adapting these methods for additional languages.

The use of time gaps for partitioning offers modest improvement as a net positive result from correcting for erroneous merges across prosodic boundaries. Resolving Vecalign errors without introducing new ones requires a more nuanced approach than simple partitioning. A promising direction could be to modify the scoring function to reduce alignment scores proportionally when overlaps occur across gaps, thereby discouraging mismatches due to timing discrepancies. We believe that a variation of this approach could significantly reduce or even eliminate nearly all false positives. Together, these results highlight the critical roles of both cleansing and sentence boundary detection in accurate sentence extraction and alignment from subtitles, while static partitioning offers only a marginal benefit as currently implemented.

Ethics Statement

This research tackles the limitations of existing subtitles datasets by introducing a novel approach that surpasses previous methods, while also ensuring full reproducibility and extensibility by making the complete code publicly available. We obtained a large collection of subtitle files from OpenSubtitles[6] a website that freely offers them for download. These subtitles were originally created by translators contributing to the public domain. Our acquisition was conducted lawfully, with permission from the website operators to obtain them in bulk.

To reduce the environmental impact of our work, we restricted our machine translation downstream task to 1 million alignments—a number insufficient for achieving fluency but adequate for comparative analysis. By releasing the code, we aim to enhance transparency in our methodology and encourage external contributions to the codebase, gold-labeled alignments, and resulting alignments. Future work should seek to optimize the loading of sentence embeddings between language pairs to optimize computational efficiency and speed.

The code used to generate the OPUS dataset is open-source[7]. However, it follows different preprocessing steps than those we employed and proved challenging to run given changes in the field since

[5]https://github.com/joshstephenson/SEAS

[6]https://opensubtitles.org

[7]https://github.com/Helsinki-NLP/subalign

its creation. To ensure a fair comparison that isolates alignment performance rather than preprocessing differences, we recreated their algorithm to the best of our ability.

Given that the subtitle data reflects naturalistic dialogues spanning a wide range of topics, we recognize that it may include strong language and other linguistic or social characteristics typical of this genre. It is important to note that we utilize this data in alignment with prior releases and established practices for research within the field.

Acknowledgments

We thank OpenSubtitles for generously providing a wealth of subtitles that made this research possible.

References

Mikel Artetxe and Holger Schwenk. 2018. Margin-based parallel corpus mining with multilingual sentence embeddings. *arXiv preprint arXiv:1811.01136*.

Richard Bellman. 1954. The theory of dynamic programming. *Bulletin of the American Mathematical Society*, 60(6):503–515.

Ondřej Bojar, Christian Buck, Chris Callison-Burch, Christian Federmann, Barry Haddow, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia. 2013. Findings of the 2013 Workshop on Statistical Machine Translation. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 1–44, Sofia, Bulgaria. Association for Computational Linguistics.

Fabienne Braune and Alexander Fraser. 2010. Improved unsupervised sentence alignment for symmetrical and asymmetrical parallel corpora. In *Coling 2010: Posters*, pages 81–89.

Peter F Brown, Jennifer C Lai, and Robert L Mercer. 1991. Aligning sentences in parallel corpora. In *29th Annual Meeting of the Association for Computational Linguistics*, pages 169–176.

Christian Buck and Philipp Koehn. 2016. Findings of the WMT 2016 bilingual document alignment shared task. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 554–563, Berlin, Germany. Association for Computational Linguistics.

Stanley F Chen. 1993. Aligning sentences in bilingual corpora using lexical information. In *31st Annual Meeting of the Association for Computational Linguistics*, pages 9–16.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2020. Language-agnostic bert sentence embedding. *arXiv preprint arXiv:2007.01852*.

William A. Gale and Kenneth W. Church. 1993. A program for aligning sentences in bilingual corpora. *Computational Linguistics*, 19(1):75–102.

Sun-Ah Jun. 2006. *Prosodic typology: The phonology of intonation and phrasing*, volume 1. Oxford University Press.

Pierre Lison and Jörg Tiedemann. 2016. OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 923–929, Portorož, Slovenia. European Language Resources Association (ELRA).

Robert C Moore. 2002. Fast and accurate sentence alignment of bilingual corpora. In *Conference of the Association for Machine Translation in the Americas*, pages 135–144. Springer.

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.

Stan Salvador and Philip Chan. 2007. Toward accurate dynamic time warping in linear time and space. *Intelligent Data Analysis*, 11(5):561–580.

Holger Schwenk and Matthijs Douze. 2017. Learning joint multilingual sentence representations with neural machine translation. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*, pages 157–167, Vancouver, Canada. Association for Computational Linguistics.

Rico Sennrich and Martin Volk. 2010. Mt-based sentence alignment for ocr-generated parallel texts.

Steinþór Steingrímsson, Hrafn Loftsson, and Andy Way. 2023. Sentalign: Accurate and scalable sentence alignment. *arXiv preprint arXiv:2311.08982*.

Brian Thompson and Philipp Koehn. 2019. Vecalign: Improved sentence alignment in linear time and space. 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 1342–1348.

Jörg Tiedemann. 2007. Improved sentence alignment for building a parallel subtitle corpus: Building a multilingual parallel subtitle corpus. *LOT Occasional Series*, 7:147–162.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in opus. In *Lrec*, volume 2012, pages 2214–2218. Citeseer.

Jörg Tiedemann. 2016. Finding alternative translations in a large corpus of movie subtitle. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3518–3522.

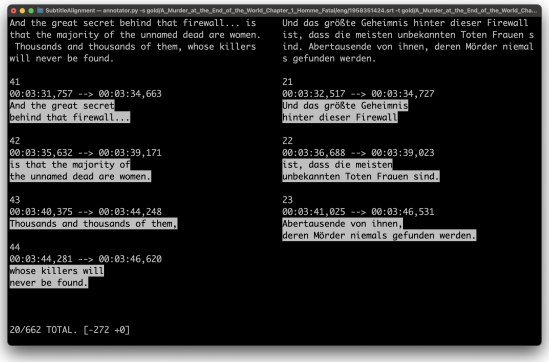

Figure 2: Annotator showing many-to-many alignment between 4 English subtitles and 3 German subtitles.

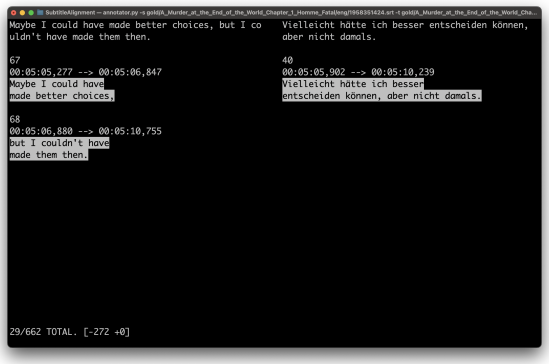

Figure 3: Annotator showing 2-to-1 alignment between English and German.

Dániel Varga, Péter Halácsy, András Kornai, Viktor Nagy, László Németh, and Viktor Trón. 2007. Parallel corpora for medium density languages. *Amsterdam Studies In The Theory And History Of Linguistic Science Series 4*, 292:247.

Dekai Wu. 1994. Aligning a parallel english-chinese corpus statistically with lexical criteria. *arXiv preprint cmp-lg/9406007*.

A Appendix

A.1 Annotation Tool

Annotating subtitle files is a labor-intensive task. To streamline this process, we developed a tool integrated with our extraction and alignment system. After extracting and generating hypothesis alignments, the tool presents these alignments alongside the corresponding subtitles in a visual editor with a Vim-like interface, implemented in Python with the curses library. Users have tools for editing alignments, primarily to remove elements that should have been filtered out during preprocessing or fragments that do not belong in the current alignment (though the tool does not explicitly restrict corrections to translations). Additional functions allow for deleting, merging, and splitting alignments.

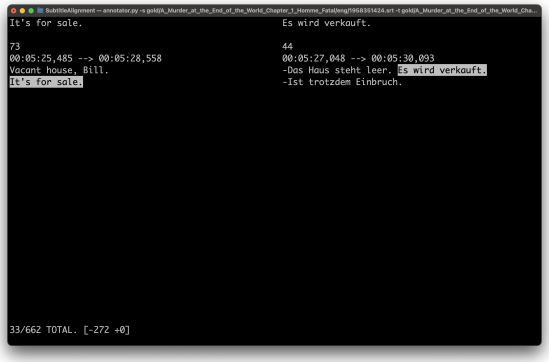

Figure 4: Annotator demonstrating how single subtitles can include more than one alignment.

The interface displays the hypothesis alignment at the top, with the source on the left, and target on the right, while the raw subtitles are shown below with extracted portions highlighted (Figure 2). This design enables quick and intuitive validation of alignments, including many-to-many (Figure 2), two-to-one (Figure 3), and even compound subtitles (Figure 4). The annotation tool along with the gold standards for extracted and aligned sentences is a part of our contribution.

A.2 Preprocessing

A.2.1 Language Verification

The first step is to verify that the language is labeled correctly. Many subtitles have been reported as one language when their contents or a significant portion of it are actually another language. To do this, we use n-grams (unigrams up to 5-grams) to detect languages against an known set of supported languages [8].

A.2.2 Data Cleansing

We next use a multi-step process based on regular expressions to clean the content by stripping everything but dialogue.

1. Strip brackets and their contents. This goes for curly brackets which are usually used to change the position of subtitles on the screen, and square brackets which are often used for captioning or speaker tagging.

2. Strip HTML tags and their contents unless they are `<b>`, `<i>`, or `<font>` tags in which

[8] https://github.com/pemistahl/lingua-py

case we only strip the tags, preserving their contents.

3. Exclude entire subtitles if they contain URLs.

4. Exclude entire subtitles if they are tagged for lyrics to soundtracks if they contain musical note unicode characters, or start with a pound sign.

A.2.3 Correcting Timecodes

Lastly, many subtitles can be out of sync with respect to their relative timecodes even when representing the same utterances in different languages and have similar durations. Steps later in the alignment process that rely on timecodes to first partition subtitle files before semantic alignment will not be successful if this is not addressed. To account for that, we run a first pass at sentence embedding alignment storing a pointer between extracted sentences and related subtitles. If the timecodes of subtitles that correspond to the aligned sentence pairs have more than a 2 second average drift between them, we adjust all timecodes in the target by that duration.

A.3 Sentence Extraction

Subtitles do not adhere to sentence boundaries. While a complete sentence may occasionally fit neatly within a single subtitle, sentences more often span multiple subtitles—sometimes as many as six or seven during extended monologues. Conversely, short, distinct utterances from multiple speakers are frequently compressed into a single subtitle. Ellipses introduce additional ambiguity: a single ellipsis may indicate a trailing-off utterance or a non-terminal pause within dialogue. This variability complicates the use of time gaps between subtitles as a cue for sentence boundaries. Additionally, while punctuation-based approaches are effective for certain languages, particularly Western ones, they often prove inadequate for others.

For sentence boundary detection, we utilized the regular expressions listed in Table 5. While these expressions are effective, they are not comprehensive and may be insufficient for non-Western languages that do not employ traditional Western punctuation for sentence boundaries. However, for English, Spanish, and German, this approach outperforms both NLTK[9] and SpaCy[10] alone on this dataset. Future improvements should focus on adapting these methods for additional languages.

Pattern that begins an utterance:
`\A[A-Z¿¡-]`
Start of string followed immediately by either an inverted question mark, inverted exclamation point or hyphen.
Pattern that ends an utterance:
`(?<!\.)[.!?]\s*\Z`
An exclamation point, question mark or single period (would not match ellipses), followed by optional whitespace and end of string.

Table 5: Regular expressions used for sentence boundary detection.

A.4 SEAS Comparative Results

Due to space restriction, we focus on English-Spanish subtitles in the main paper. Our results on each subset of the annotated data in Spanish is consistent with the results reported in the paper. We extend this evaluation by annotating the alignment of same English source subtitles over English-German data. The results on English-German are similar to the results on Spanish. We provide the German results on the same data below.

[9] https://www.nltk.org/
[10] https://spacy.io/

Title	TP	FN	FP	Recall	Precision	F-1
3 Body Problem S1E1	351	202	139	63.47	71.63	67.31
A Murder at the End of the World S1E1	348	308	230	53.05	60.21	56.40
Better Call Saul S5E2	270	379	246	41.60	52.33	46.35
Outer Range S2E5	277	170	117	61.97	70.30	65.87
Yellowstone S5E8	422	136	102	75.63	80.53	78.00
Total	1668	1195	834	66.67	58.26	62.18

Table 6: System 0. Strictly based on time-code alignment after cleansing and sentence boundary detection.

Title	TP	FN	FP	Recall	Precision	F-1
3 Body Problem S1E1	173	380	475	26.70	31.28	28.81
A Murder at the End of the World S1E1	326	327	642	33.68	49.92	40.22
Better Call Saul S5E2	240	406	426	36.04	37.15	36.59
Outer Range S2E5	178	273	337	34.56	39.47	36.85
Yellowstone S5E8	281	277	518	35.17	50.36	41.41
Total	1198	1663	2398	33.31	41.87	37.11

Table 7: System 1 on English-Spanish. cleansing: No. Sentence boundary detection: No. Partitioning: No.

Title	TP	FN	FP	Recall	Precision	F-1
3 Body Problem S1E1	295	256	342	46.31	53.54	49.66
A Murder at the End of the World S1E1	379	272	373	50.40	58.22	54.03
Better Call Saul S5E2	265	382	389	40.52	40.96	40.74
Outer Range S2E5	237	210	262	47.49	53.02	50.11
Yellowstone S5E8	320	239	457	41.18	57.25	47.90
Total	1496	1359	1823	45.07	52.40	48.46

Table 8: System 2 on English-Spanish. cleansing: Yes. Sentence boundary detection: No. Partitioning: No.

Title	TP	FN	FP	Recall	Precision	F-1
3 Body Problem S1E1	354	199	202	64.01	63.67	63.84
A Murder at the End of the World S1E1	587	105	235	84.83	71.41	77.54
Better Call Saul S5E2	498	164	186	75.23	72.81	74.00
Outer Range S2E5	324	125	133	72.16	70.90	71.52
Yellowstone S5E8	465	99	106	82.45	81.44	81.94
Total	2228	692	862	76.30	72.10	74.14

Table 9: System 3 on English-Spanish. cleansing: No. Sentence boundary detection: Yes. Partitioning: No.

Title	TP	FN	FP	Recall	Precision	F-1
3 Body Problem S1E1	518	42	49	92.50	91.36	91.93
A Murder at the End of the World S1E1	669	30	32	95.71	95.44	95.57
Better Call Saul S5E2	608	57	75	91.43	89.02	90.21
Outer Range S2E5	425	35	41	92.39	91.20	91.79
Yellowstone S5E8	523	42	39	92.57	93.06	92.81
Total	2743	206	236	93.01	92.08	92.54

Table 10: System 4 on English-Spanish. cleansing: Yes. Sentence boundary detection: Yes. Partitioning: No.

Title	TP	FN	FP	Recall	Precision	F-1
3 Body Problem S1E1	519	41	47	92.68	91.70	92.18
A Murder at the End of the World S1E1	678	24	30	96.58	95.76	96.17
Better Call Saul S5E2	614	51	69	92.33	89.90	91.10
Outer Range S2E5	427	32	39	93.03	91.63	92.32
Yellowstone S5E8	530	35	41	93.81	92.82	93.31
Total	2768	183	226	93.80	92.45	93.12

Table 11: System 5 on English-Spanish. cleansing: Yes. Sentence boundary detection: Yes. Partitioning: 3 sec gaps.

Title	TP	FN	FP	Recall	Precision	F-1
3 Body Problem S1E1	535	22	22	96.05	96.05	96.05
A Murder at the End of the World S1E1	606	51	62	92.24	90.72	91.47
Better Call Saul S5E2	553	49	86	91.86	86.54	89.12
Outer Range S2E5	432	29	33	93.71	92.90	93.30
Yellowstone S5E8	506	35	35	93.82	93.19	93.50
Total	2632	186	238	93.40	91.7	92.55

Table 12: System 5 on English-German. Better Call Saul had a lot of different formatting that most subtitle files which yielded a lot of false positives. Improvements to preprocessing would decrease false negatives and false positives.