The Aligned Multimodal Movie Treebank: An audio, video, dependency-parse treebank

Anonymous ACL submission

Abstract

Treebanks have traditionally included only text and were derived from written sources such as newspapers or the web. We introduce the Aligned Multimodal Movie Treebank, an English language treebank derived from naturalistic dialog in Hollywood movies which includes the source video and audio, transcriptions with word-level alignment to the audio stream, as well as part of speech tags and dependency parses in the Universal Dependencies formalism. AMMT consists of 31,264 sentences and 218,090 words, that will be the 3rd largest UD English treebank, and the only multimodal treebank in UD. To help with the web-based annotation effort, we also introduce the Efficient Audio Alignment Annotator (EAAA), a companion tool that enables annotators to speed-up significantly the annotation process.

Keywords: multimodal, video, audio, dependency parsing, treebank, Universal Dependencies

1 Introduction

Treebanks are fundamental resources in Natural Language Processing, and despite their central role most existing treebanks are derived from single-modality texts such as newspapers, blogs, and other online communities. The vocabulary, syntax, and statistics of spoken and written language can be quite different from one another (Caines et al., 2017). To complement these datasets, support experiments with naturalistic data, and aid the advent of conversational agents, we have created a new dataset, the Aligned Multimodal Movie Treebank, AMMT, the content of which is derived from language spoken in Hollywood movies. AAMT will be released publicly under an open source license and will be contributed to the Universal Dependencies (Nivre et al., 2016) treebanks.

The closest existing dataset to AMMT is Treebank-3 of the Penn Treebank (Marcus et al., 1993), which includes the Penn Treebank Switchboard corpus (Godfrey et al., 1992). This corpus contains nearly one million transcribed words from Switchboard annotated with part of speech tags, dysfluencies, and parse trees, and it also includes alignment between words and audio. However, several key differences between this dataset and our own exist. AMMT is being made open with this contribution and not restricted to LDC members, it is multi-modal, annotated with Universal Dependencies rather than Penn Treebank dependencies, and conversations are much shorter and more natural (Switchboard was designed to have long 10 minute conversations between strangers on the

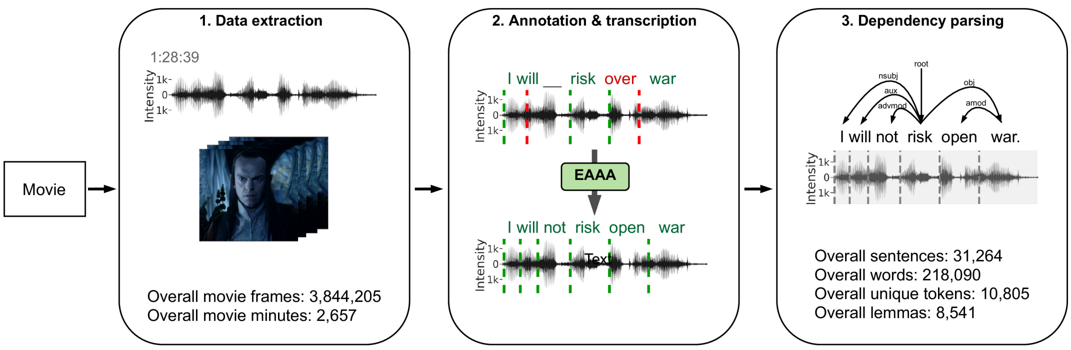

Figure 1: An overview our AAMT, our novel multimodal dataset, consisting of 21 transcribed and parsed movies. EAAA is a new transcription and alignment introduced below.

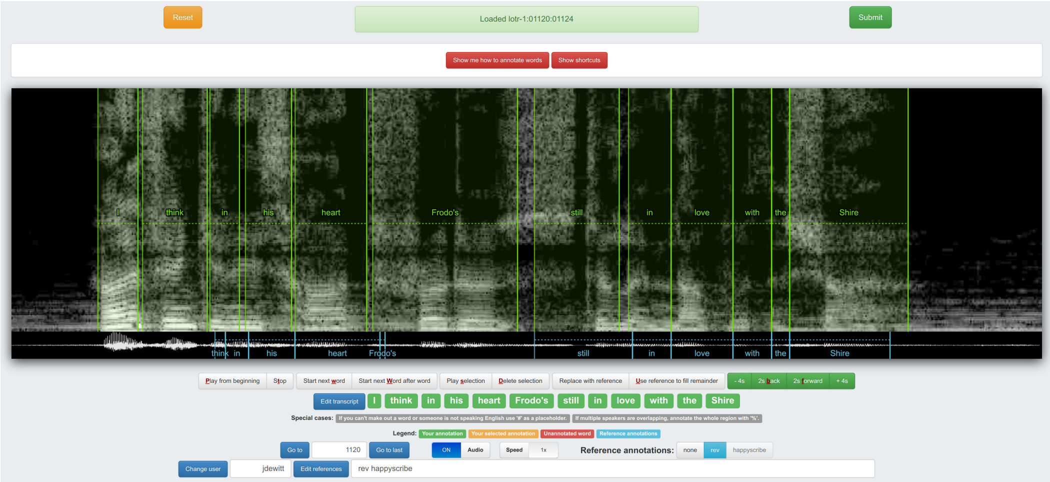

Figure 2: A screenshot of EAAA, the Efficient Audio Alignment Annotator. EAAA allows annotators to browse entire long movies, to play audio segments, play portions of the audio segments, edit the transcript, review multiple reference annotations, and annotate and change word boundaries. EAAA also includes an in-application walkthrough as well as extensive keyboard shortcuts. The main annotation area shows a spectrogram with annotated words. Words can be dragged with a mouse and similarly word boundaries can be adjusted with the mouse. The audio for individual words can be played by clicking them, while any audio segment can be played by clicking and dragging the portion that should be played. At the bottom, in blue, one or more reference annotations are shown which can be toggled on the fly. Annotators can start with a blank slate or initialize annotations from any reference annotation. Audio speed can be controlled as necessary.

phone discussing one of a preselected list of topics). AMMT also includes many more speakers and its audio quality allowed us to recover almost all spoken words. For practical experiments, AMMT is also significantly more entertaining for subjects.

Our contributions are: 1. AMMT is the first large-scale treebank to include audio and video. 2. AMMT includes fine-grained ms-level word boundaries. 3. AMMT is parsed in the Universal Dependencies framework and is the 3rd largest English UD treebank. 4. A new tool, Efficient Audio Alignment Annotator (EAAA), for rapid word boundaries annotation in large corpora.

2 Dataset

The AMMT dataset is an English language treebank based on 21 Hollywood movies that provides transcriptions with word-level alignment to the audio stream, part of speech tags and dependency parses in the Universal Dependencies formalism, as well as a tool chain to enable access to the source video and audio. The dataset consists of 31,264 sentences, 218,090 words, 8,541 lemmas and 10,805 unique tokens. The counts of POS tags and the most frequency dependencies are shown in appendix A. The 21 movies from which the dataset is derived were included in full and are listed in table 3 along with per-movie statistics.

Movies were chosen to be appropriate for many ages, with the highest rating being PG-13. Their release dates range from 1995 to present, and belong to a variety of movie genres (including action, adventure, animation, comedy, drama, fantasy, family, and sci-fi in the IMDB categorization). They were also selected to have verbose scripts, in the top 50% of randomly sampled movies. Movies which included extensive signing such as musicals were omitted. Copies of the movies were obtained and extracted in full including opening and closing credits. Special features and after-credits scenes were omitted.

2.1 Transcription pipeline

The audio track was originally transcribed using the Google Cloud Speech-to-Text API (Google, 2020). It was then corrected by annotators[1] and then further extensively corrected by 7 in-house annotators (3 male, 4 female; undergraduates and masters students).

Transcription was verbatim without any corrections for dysfluencies or mistakes. Instructions were provided to the annotators to standardize the transcripts and eliminate problematic audio segments. Foreshortened words (*’round* vs *around*)

[1]Annotators hired from *Rev.com* and *happyscribe.com* depending on the movie.

were transcribed as they were said including the foreshortening. Abbreviations were always expanded (*dr.* vs *doctor*). Cardinal and ordinal numbers were spelled out, while long numbers were written as spoken including conjunctions such as *and* (e.g., *five hundred and five*).

Manual transcription was carried out simultaneously with word boundary annotation using a purpose-built tool, EAAA (see section 3). EAAA presented annotators with a spectrogram for 4 second segments of a movie, along with the ability to replay and slow down any sub-segment and seek throughout the movie. In some cases, annotators could hear specific words but could not clearly identify in the spectrogram where those words occurred. Annotators were instructed to annotate what they heard regardless of the spectrogram, sometimes leading to such short words having zero-length intervals. In addition to the transcript, other common cases were noted, but not transcribed. Foreign sentences (e.g., Elvish in the movie *The Lord Of The Rings*) were marked but not included in the corpus, although one-off foreign words in English sentences were transcribed. All cases of singing, unintelligible speech, and multiple speakers overlapping were noted and eliminated from the dataset.

After transcription and word boundary alignment the text was segmented into sentences. Annotators marked the end of each sentence manually and fixed capitalization (of both proper nouns and sentences as needed). Throughout this process some critical punctuation was introduced as annotators saw fix.

2.2 Dependency parsing pipeline

We parsed all transcriptions with Stanza (Qi et al., 2020) using the standard English model. One full time annotator, over the course of a year, annotated and corrected the parse of every sentence. Three team members knowledgeable about linguistics and universal dependencies were on hand to answer questions about edge cases. The fact that a single person annotated the dataset provides it with internal consistency that is not achievable when multiple annotators.

2.3 Validating annotator performance

After the annotation process we sampled 300 sentences to compute the accuracy of the annotations. Sentences for this experiment were chosen uniformly across movies and sentence length, focusing on sentences that were between 5 and 20 words long to avoid the effect of very short or very long sentences. See table 1. Overall, the accuracy of the annotations was very high, with 99.53% accuracy on POS tagging, 98.95% on correctly placing dependencies (UAS), and 98.31% on correctly identifying the type of a dependency relation. MLAS ties together POS and LAS into a single number, 96.72, which measures the accuracy of the annotations (Straka, 2018).

Metric	Precision	Recall	F1 Score	AligndAcc
Words	100.00	100.00	100.00	N/A
UPOS	99.53	99.53	99.53	99.53
UAS	98.95	98.95	98.95	98.95
LAS	98.31	98.31	98.31	98.31
CLAS	97.75	97.71	97.73	97.71
MLAS	96.74	96.70	96.72	96.70

Table 1: The accuracy of AMMT syntactic annotations. 300 sentences of length 5 through 20 uniformly sampled across movies were reannotated by an expert annotator who did not contribute to the dataset otherwise.

We also found word boundaries inter-coder agreement to be remarkably high, with less than 15ms on average for all words in a single movie (*Lord Of The Rings*) annotated by 5 annotators.

2.4 Performance of existing parsers

We compared our annotations against those produced by Stanza (Qi et al., 2020) in fig. 3. Stanza was the original parser used to initialize the treebank before extensive human correction. This likely biases the results toward Stanza in subtle ways which we do not investigate here (Berzak et al., 2016), beyond section 2.3, where we measure the accuracy of the corrected annotations.

Note that performance on short sentences, fewer than 3 words, and long sentences, with more than 20 words, is far worse than average-case performance. This trend is not observed in other corpora such as the English Web Treebank (EWT) (Silveira et al., 2014), where performance increases for short sentences (although these are very infrequent) while the performance drop for long sentences is half or less than that seen in AMMT. The performance drop for short sentences appears to be driven by part of speech tag errors, see the relative drop in POS accuracy between fig. 3(a,b,c) — perhaps such sentences require more context to be correctly interpreted. While the performance drop for long sentences appears to be driven by incorrectly identified relationships, see the relative drop in UAS accuracy between fig. 3(a,b,c).

Metric	Precision	Recall	F1 Score	AligndAcc
Words	99.51	99.75	99.63	N/A
UPOS	97.64	97.88	97.76	98.13
UAS	88.02	88.24	88.13	88.46
LAS	85.68	85.89	85.78	86.10
CLAS	83.40	83.01	83.20	83.29
MLAS	81.38	80.99	81.18	81.27

(a) All sentences

Metric	Precision	Recall	F1 Score	AligndAcc
Words	99.45	99.53	99.49	N/A
UPOS	91.49	91.56	91.53	92.00
UAS	91.31	91.38	91.35	91.82
LAS	88.76	88.83	88.80	89.25
CLAS	86.49	86.06	86.28	86.71
MLAS	75.87	75.50	75.68	76.06

(b) Short sentences, fewer than 3 words

Metric	Precision	Recall	F1 Score	AligndAcc
Words	99.52	99.78	99.65	N/A
UPOS	98.44	98.70	98.57	98.92
UAS	80.47	80.68	80.57	80.86
LAS	78.78	79.00	78.89	79.17
CLAS	76.32	76.06	76.19	76.28
MLAS	74.02	73.77	73.90	73.98

(c) Long sentences, more than 20 words

Figure 3: (a) The overall accuracy of Stanza on AMMT. Performance drops significantly for (b) short sentences which are common in speech as well as for (c) long sentences.

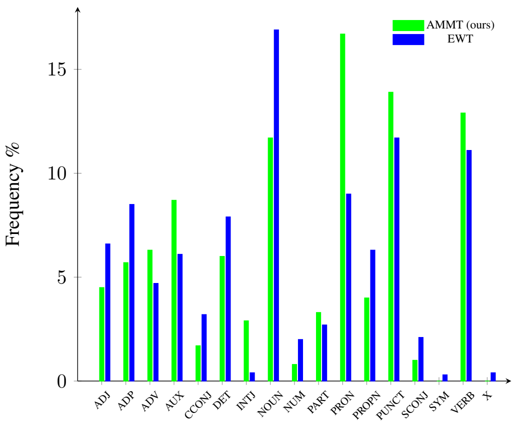

Figure 4: Comparing POS frquency in EWT, a treebank derived from text on the web, and AMMT, our new benchmark derived from spoken language. Among many differences, note that nouns are much less common and pronouns are far more common.

3 Tools

To efficiently annotate the alignment between word onsets and offsets and the audio stream we created a new tool[2], the Efficient Audio Alignment Annotator (EAAA). EAAA enables annotators to start with a rough transcript and approximate alignment between words and the audio track. Annotators can simultaneously correct the transcript while annotating new words. An overview of the EAAA interface is shown in fig. 2.

Tools such as Praat (Boersma, 2001) also allow for annotating audio corpora with word boundaries. Unlike Praat, EAAA is web-based making it easier for annotators to use. Data such as spectrograms and wave files seen by annotators is pre-processed on the server-side, making EAAA extremely fast. Since EAAA is a single-purpose tool meant for transcription and fine-grained alignment, it provides custom features which significantly speed up the annotation process like keyboard shortcuts, the ability to handle audio files of any length, and a streamlined interface. EAAA also handles multiple concurrent annotators, sharing and comparing multiple annotations directly.

EAAA pre-processes movie files into 4 second segments that overlap by 2 seconds and computes spectrograms for segment with Librosa (McFee et al., 2015). Storage is provided by a local Redis database which is not exposed to the internet. In addition, EAAA includes a telemetry server which collects comprehensive information during the annotation process including every transcript change, keyboard shortcut used, and mouse press.

4 Conclusion

AAMT and EAAA are open source and AAMT will be contributed to the UD treebanks [3]. To handle the source copyrighted material AAMT will provide multiple aligned audio samples and video clip samples from every movie allowing users to obtain their own copies of the movies and then realign to the dataset.

Most datasets for evaluating and training parsers are focused on written rather than spoken language. With the rise of conversational agents, AAMT can serve as a more predictive benchmark in this domain.

At present, no end-to-end video-and-audio to parse systems exist, despite the fact that humans can use visual information to disambiguate and contextualize auditory information. We hope that AMMT and its tooling will support further work on conversational agents, multimodal end-to-end parsing, as well as psychophysics and neuroscience with naturalistic stimuli.

[2] EAAA along with annotation guidelines is open source and available on GitHub, link redacted to protect anonymity.

[3] AAMT annotations were approved by an IRB.

References

Yevgeni Berzak, Yan Huang, Andrei Barbu, Anna Korhonen, and Boris Katz. 2016. Anchoring and agreement in syntactic annotations. *arXiv preprint arXiv:1605.04481*.

Paul Boersma. 2001. Praat, a system for doing phonetics by computer. *Glot. Int.*, 5(9):341–345.

Andrew Caines, Michael McCarthy, and Paula Buttery. 2017. Parsing transcripts of speech. *EMNLP 2017*, page 27.

John J Godfrey, Edward C Holliman, and Jane McDaniel. 1992. Switchboard: Telephone speech corpus for research and development. In *Acoustics, Speech, and Signal Processing, IEEE International Conference on*, volume 1, pages 517–520. IEEE Computer Society.

Google. 2020. Speech-to-text: Automatic speech recognition — google cloud.

Mitchell Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of english: The penn treebank.

Brian McFee, Colin Raffel, Dawen Liang, Daniel PW Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto. 2015. librosa: Audio and music signal analysis in python. In *Proceedings of the 14th python in science conference*, volume 8, pages 18–25. Citeseer.

Joakim Nivre, Marie-Catherine De Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, Christopher D Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, et al. 2016. Universal dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1659–1666.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108.

Natalia Silveira, Timothy Dozat, Marie-Catherine De Marneffe, Samuel R Bowman, Miriam Connor, John Bauer, and Christopher D Manning. 2014. A gold standard dependency corpus for english. In *LREC*, pages 2897–2904. Citeseer.

Milan Straka. 2018. Udpipe 2.0 prototype at conll 2018 ud shared task. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 197–207.

A Appendix

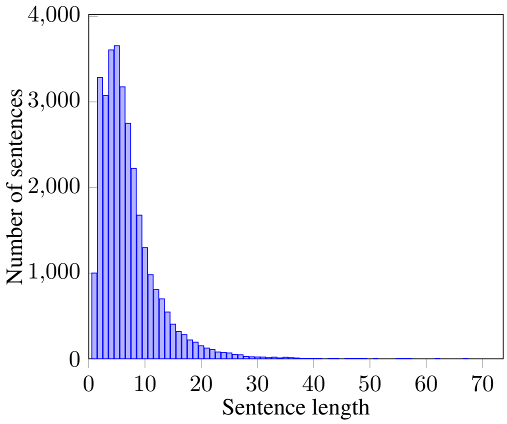

Figure 5: The distribution of sentence lengths. Most sentences are quite short. The mean sentence length is 6.97 words long. Compare to standard corpora derived from written sources like the English Web Treebank (15.33 words/sentence) long and the Penn Treebank (23.73 words/sentence in the test set).

Aligned Multimodal Movie Treebank	
sentences	31,264
tokens	218,090
lemmas	8,541
types	10,805
num. movies	21

Table 2: Basic statistics of the dataset

Movie	Year	Time (s)	Sentences	Tokens	Types	Rating	FPS	Frames
Ant-Man	2015	7027	1412	9846	1956	PG-13	23.98	168507
Aquaman	2018	8601	1003	7218	1563	PG-13	23.98	206251
Avengers: Infinity War	2018	8961	1372	8479	1780	PG-13	23.98	214884
Black Panther	2018	8073	1139	7571	1628	PG-13	23.98	193590
Cars 2	2011	6377	1801	11404	2060	G	23.98	152920
Coraline	2009	6036	933	5428	1251	PG	23.98	144743
Fantastic Mr. Fox	2009	5205	1162	8457	1892	PG	23.98	124815
Guardians of the Galaxy 1	2014	7251	1104	8241	1799	PG-13	23.98	173878
Guardians of the Galaxy 2	2017	8146	1180	9332	1839	PG-13	23.98	195341
The Incredibles	2003	6926	1408	9369	1966	PG	23.98	166085
Lord of the Rings 1	2001	13699	1424	10538	2011	PG-13	23.98	328502
Lord of the Rings 2	2002	14131	1620	11017	2085	PG-13	23.98	338861
Megamind	2010	5735	1351	8833	1748	PG	23.98	137525
Sesame Street Ep. 3990	2016	3440	718	4218	804	TV-Y	29.97	103096
Shrek the Third	2007	5568	999	7192	1586	PG	23.98	133520
Spiderman: Far From Home	2019	7764	1705	12004	1988	PG-13	23.98	186180
Spiderman: Homecoming	2017	8008	1993	12258	2107	PG-13	23.98	192031
The Martian	2015	9081	1421	11360	2210	PG-13	23.98	217762
Thor: Ragnarok	2017	7831	1471	9651	1806	PG-13	23.98	187787
Toy Story 1	1995	4863	1240	7194	1545	G	23.98	116614
Venom	2018	6727	1301	7859	1527	PG-13	23.98	161313

Table 3: Statistics of the 21 movies from which AMMT is derived from. Movies were selected to be appropriate for most ages enabling a wide range of experiments. Movies are not randomly sampled, they were selected for their verbose scripts and subjects entertainment during experiments.

POS	Count
ADJ	9829
ADP	12464
ADV	13688
AUX	18965
CCONJ	3746
DET	12984
INTJ	6275
NOUN	25457
NUM	1835
PART	7202
PRON	36370
PROPN	8679
PUNCT	30301
SCONJ	2140
SYM	10
VERB	28139
X	6

Dependencies	Count
nsubj	25050
advmod	14003
obj	12825
det	12325
case	11274
aux	9286
cop	7830
obl	6653
mark	5693
amod	4958
xcomp	4306
nmod:poss	3996
discourse	3912
cc	3682
compound	3335
conj	3322
vocative	3134

Table 4: The distribution of POS tags (left), and the most common dependencies (right). There is a long tail of dependencies.