Bazinga! A Dataset for Multi-Party Dialogues Structuring

Paul Lerner, Juliette Bergoënd, Camille Guinaudeau, Hervé Bredin, Benjamin Maurice, Sharleyne Lefevre, Martin Bouteiller, Aman Berhe, Léo Galmant, Ruiqing Yin, et al.

Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 3434–3441
Marseille, 20-25 June 2022

Bazinga! A Dataset for Multi-Party Dialogues Structuring

Paul Lerner[1], Juliette Bergoënd[*], Camille Guinaudeau[1], Hervé Bredin[2]
Benjamin Maurice[*], Sharleyne Lefevre[*], Martin Bouteiller[*]
Aman Berhe[*], Léo Galmant[*], Ruiqing Yin[*], Claude Barras[3]
[1]Université Paris-Saclay, CNRS, LISN, [2]IRIT, Université de Toulouse, CNRS, [3]Vocapia Research
[1,3]91400, Orsay, France, [2]Toulouse, France
first.last@lisn.upsaclay.fr, herve.bredin@irit.fr
[*]Work done while at Université Paris-Saclay, CNRS, LISN

Abstract

We introduce a dataset built around a large collection of TV (and movie) series. Those are filled with challenging multi-party dialogues. Moreover, TV series come with a very active fan base that allows the collection of metadata and accelerates annotation. With 16 TV and movie series, *Bazinga!* amounts to 400+ hours of speech and 8M+ tokens, including 500K+ tokens annotated with the speaker, addressee, and entity linking information. Along with the dataset, we also provide a baseline for speaker diarization, punctuation restoration, and person entity recognition. The results demonstrate the difficulty of the tasks and of transfer learning from models trained on mono-speaker audio or written text, which is more widely available. This work is a step towards better multi-party dialogue structuring and understanding. *Bazinga!* is available at `hf.co/bazinga`. Because (a large) part of *Bazinga!* is only partially annotated, we also expect this dataset to foster research towards self- or weakly- supervised learning methods.

Keywords: Multi-Party Dialogues, Speaker Diarization, Entity Linking

1. Introduction

Multi-party dialogues, while ubiquitous in everyday life, are understudied and lack of large-scale, open-access datasets (Mahajan and Shaikh, 2021). To fill this gap, we propose *Bazinga!*, a dataset of 16 TV and movie series, filled with multi-party dialogues, covering a wide range of genres (*e.g.* drama or comedy). Although dialogues in series are mostly scripted, they emulate everyday-life situations: interruptions, speech overlap, disfluencies, *etc.* Each of these characteristics rises a scientific challenge (Park et al., 2022; Szymański et al., 2020).

The discrepancy between training datasets (either mono-speaker or written text) and human interactions (on which systems are actually deployed) has been noticed by Szymański et al. (2020) in the field of Automatic Speech Recognition (ASR). They note that modern ASR systems achieve 2-3% Word Error Rate (WER) on standard datasets but peak up to the 46%–73% range on dinner party conversations. We observe a similar trend for other tasks such as named entity recognition and speaker diarization.

While *Bazinga!* includes a gold-standard subset for evaluation, the majority of the corpus is silver-standard, *i.e.* annotated semi-automatically. We expect this *silver* subset to foster research towards self- or weakly-supervised learning methods.

Introduced in Section 2, the dataset provides gold-standard annotations for several tasks, ranging from speech processing to natural language processing, such as automatic speech recognition, punctuation restoration, named entity recognition, entity linking, addressee detection, and speaker identification. The *silver* subset includes timestamps for every word and speaker, which allows the training and evaluation of speaker diarization systems.

Bazinga! can be used for research on a wide range of tasks, listed in Section 3. In particular, we provide baseline results for speaker diarization, punctuation restoration, and person entity recognition. The results demonstrate the difficulty of the task and of transfer learning from models trained on mono-speaker data, which is more widely available.

The dataset is available at `hf.co/bazinga` after a quick registration step. It can then be loaded and processed in a few lines of Python code thanks to the `datasets` library (Lhoest et al., 2021). More details are provided in Section 4.

2. The *Bazinga!* Dataset

Bazinga! is a dataset built around the English audio tracks of 13 TV series (*24*, *Battlestar Galactica*, *Breaking Bad*, *Buffy the Vampire Slayer*, *ER*, *Friends*, *Game of Thrones*, *Homeland*, *Lost*, *Six Feet Under*, *The Big Bang Theory*, *The Office*, and *The Walking Dead*) and 3 movie series (*Harry Potter*, *Star Wars*, and *The Lord of the Rings*). Those series were selected to cover a wide range of different genres (*e.g.* drama, thriller, fantasy, comedy, or science fiction) and because we could gather high quality manual transcripts from websites maintained by their very active fan base.

As depicted in Figure 1, each word of the manual transcripts have been annotated with the following metadata: timestamp (*when was the word pronounced?*), speaker (*who pronounced it?*), addressee (*who was it addressed to?*), entity linking (*is it referring to a per-*

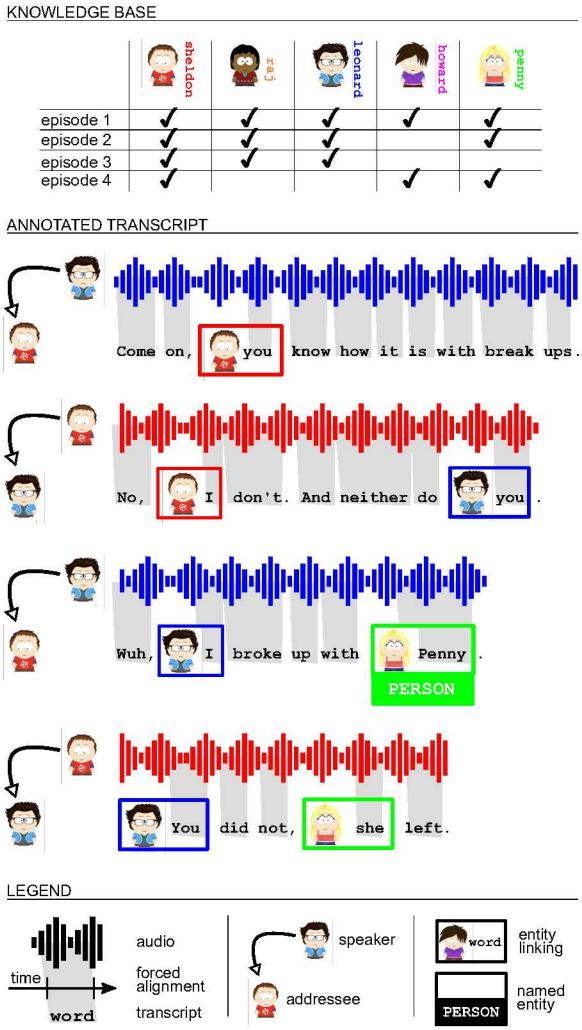

Figure 1: *Bazinga!* illustrated.

son? and who is that?), named entity (*is it the name of a person or just a pronoun?*).

2.1. Knowledge Base

We scrapped the `IMDb.com` page of each TV and movie series and gathered the list of episodes and characters. Each episode is assigned a unique identifier, its title, and the link to the episode's `IMDb.com` page. Each character is assigned a unique identifier (used throughout the dataset), their full name in the series, a link to the character's `IMDb.com` page, and the name of the actor playing the character.

As depicted in the upper part of Figure 1, we scrapped episode pages to provide the actual list of characters appearing in each episode. Episode pages contain lots of additional valuable details such as story lines, summaries, or synopses for each episode, but we did not integrate those in the first release of *Bazinga!*.

2.2. Audio

English audio tracks were extracted from (Zone 2) DVDs with a combination of several open-source tools such as `lsdvd` (to browse the content of DVDs programmatically), `HandBrake` (to extract episodes in MKV format), and `ffmpeg` (to resample audio tracks). They are provided as 16kHz mono-channel *wav* files (one file per episode).

2.3. Transcripts

Manual transcripts were scraped from public websites. They are provided as-is for the silver *Bazinga!* subset (Table 3) but were double-checked manually for parts of *Bazinga!*, using the process described in paragraph 2.5.

2.4. Speaker

Each word is assigned the unique identifier of the speaker who pronounced it. It was either directly part of downloaded transcripts (in which case we manually mapped each speaker to its unique identifier from the knowledge base) or it was not provided by the transcripts (in which case we designed an interactive annotation tool using speaker embedding and active learning to speed up the annotation process, depicted in Figure 2). A word pronounced by multiple speakers simultaneously is marked as such.

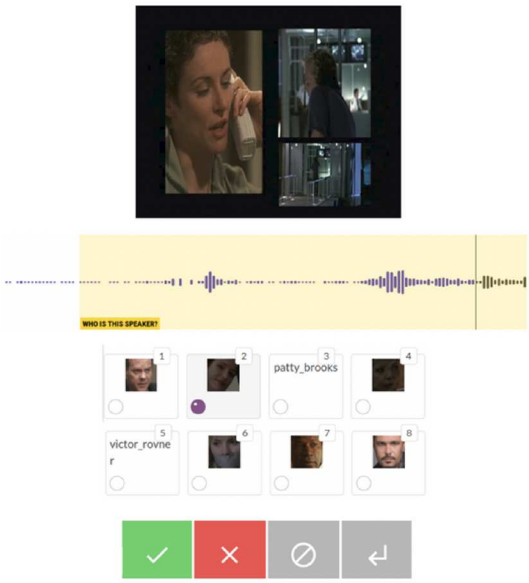

Figure 2: *Speaker* interactive annotation tool, based on Prodigy.

2.5. Forced Alignment

We use an in-house forced alignment toolkit to align manual transcripts with the audio tracks automatically. Therefore, each word in the manual transcript is given a start and end time in the audio track. Since this process is automatic, these timestamps should be considered silver-standard.

That being said, each word is also assigned a score indicating how confident the toolkit is about these auto-

matic timestamps. This proved helpful when double-checking the manual transcripts, as regions with the lowest forced-alignment confidence scores tend to indicate that the transcript might be incorrect (*e.g.* because of the inclusion of stage directions). Similarly, one could filter out low-scores regions when training or evaluating speaker diarization systems.

2.6. Addressee

We manually annotated *Bazinga!* in terms of addressee: *to whom the current speaker is speaking?* Each sentence is tagged with the unique identifier of the addressee (if the current speaker speaks to one specific character), a list of identifiers (if the current speaker addresses a group of characters), or marked as addressed to no one in particular.

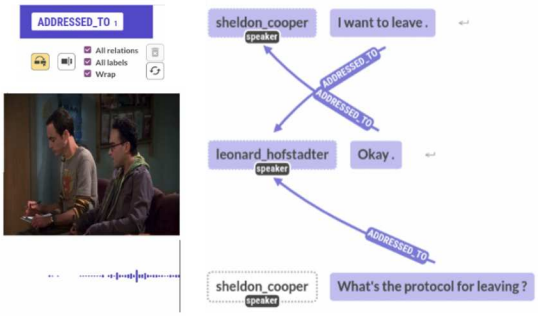

Figure 3: *Addressee* interactive annotation tool, based on Prodigy.

As depicted in Figure 3, we designed another interactive tool to perform this manual annotation that displays both the transcript and the video clip to ease the task of the human annotator.

Figure 4 depicts the interactions between the five main characters of *The Big Bang Theory*'s first season. These interactions, obtained from the addressee gold-standard annotations, represent the number of tokens addressed by one character to another. We can clearly see patterns emerging. First, 36.9% of tokens are exchanged between Sheldon Cooper and Leonard Hofstadter and more than 37% of the interactions are done by Sheldon Cooper. On the contrary, Raj Koothrappali's interactions represent only 5% of the interactions (among those 5% are addressed to himself).

2.7. Entities

Entity linking annotations are provided for each word in the transcript files. Those annotations gather named entities for persons, pronouns, nicknames, family relations, and denominations. We annotate person-named entities; entities such as organizations or countries are not annotated. If an entity is related to a character in the knowledge base, we annotate it with its corresponding identifier. If a name is composed of multiple words, we annotate each word with the corresponding identifier

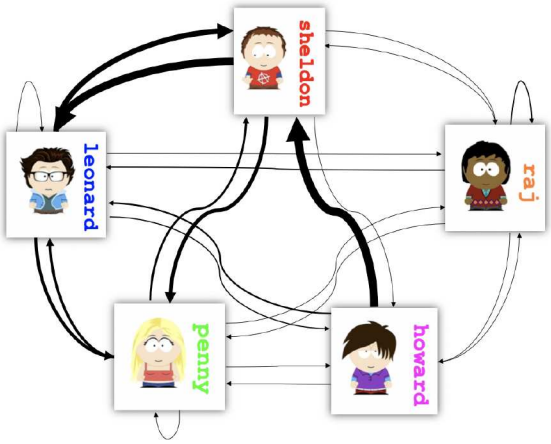

Figure 4: Interactions between the five most important characters in the first season of The Big Bang Theory. The thicker the line from X to Y, the larger the number of tokens addressed by X to Y.

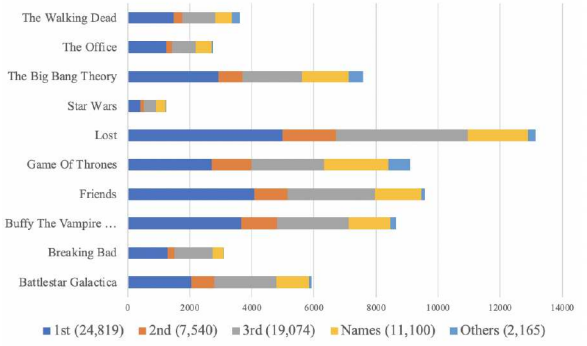

Figure 5: First (1st), second (2nd), third (3rd) persons pronouns, names and others category distribution for entity linking (number of elements for each category in brackets).

in the knowledge base. We annotate celebrity names or unknown characters not in the knowledge base with the "UNKNOWN" label. From these annotations, person-named entities are identified with a "PERSON" tag to allow person-named entity evaluation.

Pronouns for the first, second, and third persons singular are annotated, as well as plural. First person singular pronouns are annotated thanks to the current speaker. Plural persons are annotated as "multiple_persons". We use addressee annotations to annotate second person singular and plural pronouns. Pronouns such as "they" or "them" are not annotated, as names that refer to a group, such as a whole family (e.g. "The Lannisters").

Denominations such as "Queen" or "King" are also annotated with the character related to the denomination, for example "King of the Seven Kingdoms." However, such titles are not considered entity-linking when they are used as a common noun, e.g. "She is a queen."

Family relationships are also part of the entity linking process : "Joey's mother" $\longrightarrow$ joey_tribbiani, gloria_tribbiani.
Figure 5 reports the tokens category distribution for the entity linking annotation. First and third person pronouns are the most frequent tokens used for entity denomination, followed by proper names for all TV series. The *Others* category (denominations, nicknames, or family relationships) is the least used in the dataset to refer to an entity.

2.8. Gold- vs. Silver-standard

As summarized in Table 1, approximately 7% of *Bazinga!* dataset comes with gold-standard annotations. Transcripts (gold- or silver-standard) are available for the whole dataset, while only half of *Bazinga!* utterances are assigned a speaker. Person entities and addressee are only available as gold-standard since the process is fully manual.

Annotations	Gold	Silver	Total
Knowledge base	100%	–	100%
Transcript	7%	93%	100%
Speaker	48%	–	48%
Entities	7%	–	7%
Addressee	7%	–	7%
Timestamps	–	100%	100%

Table 1: Percentage of *Bazinga!* with gold- or silver-standard annotations. Note that 100% of transcripts are written by humans but those were extracted automatically from websites. Only 7% were double-checked after download.

Table 2 provides a closer look at the content of the *gold* subset, *i.e.* episodes for which every annotation (transcript, speaker, entities, and addressee) is gold-standard. It covers 9 TV series (every episode in their respective first season) and 7 Star Wars movies. This amounts to a total of 120+ episodes, 500k+ words, more than one thousand speakers, and approximately 30 hours of speech (estimated using the timestamps obtained through forced alignment).

	Episodes	Tokens	Speakers	Speech
Battlestar Galactica	13	56k	119	3.0h
Breaking Bad	7	29k	58	1.6h
Buffy the Vampire...	12	68k	99	3.4h
Friends	24	82k	110	4.0h
Game of Thrones	10	54k	126	3.0h
Lost	25	101k	131	4.9h
The Big Bang Theory	17	58k	41	3.1h
The Office	6	23k	25	1.1h
The Walking Dead	6	25k	38	1.3h
Star Wars	7	72k	281	3.9h
Total	127	569k	1,028	29.4h

Table 2: Content of the *gold* subset.

Table 3 focus on the *silver* subset, *i.e.* episodes for which at least one annotation (among transcript, speaker, entities, and addressee) is either missing or available in silver-standard. It covers 13 TV series (including episodes of subsequent seasons of the 9 TV series of the gold subset) and 2 movie series. This amount to around 1,600 episodes, 7 million words, and approximately 400 hours of speech (again, estimated through forced alignment).

	Episodes	Tokens	Speakers*	Speech
24	168	886k	–	46.0h
Battlestar Galactica	58	262k	100+	14.3h
Breaking Bad	54	236k	20+	13.5h
Buffy the Vampire...	131	666k	200+	35.1h
ER	305	2,043k	–	100.4h
Friends	209	702k	190+	38.2h
Game of Thrones	50	247k	200+	14.0h
Homeland	58	278k	–	14.7h
Lost	79	300k	200+	14.3h
Six Feet Under	50	314k	30+	17.2h
The Big Bang Theory	190	581k	70+	33.2h
The Office	182	700k	140+	38.4h
The Walking Dead	93	260k	50+	13.7h
Harry Potter	8	81k	50+	4.6h
The Lord of the R...	3	28k	30+	1.7h
Total	1,638	7,584k	1,300+	399.4h

Table 3: Content of the *silver* subset. *Speakers that said at least 100 words.

3. The *Bazinga!* Tasks

This section describes a range of tasks for which the *Bazinga!* dataset could be useful for training and/or evaluating machine learning approaches. We also provide baseline results for some of them. Due to the bimodal nature of the *Bazinga!* dataset (it contains audio files and their transcript), those tasks can be approximately categorized into two groups – natural language processing tasks and speech processing tasks – though this boundary might be slightly fuzzy for some.

3.1. Recommended Experimental Protocol

We recommend using season 1 episodes (and the first movie of each movie series) as test set, seasons 2 and 3 (as well as movies 2 and 3) as development or validation set, and the rest of *Bazinga!* as training set.

3.2. Speaker Diarization and Identification

Speaker diarization is the process of partitioning a multi-party conversation into homogeneous temporal segments according to the identity of the speaker. Table 4 reports the performance obtained by a speaker diarization pipeline pretrained on DIHARD (Ryant et al., 2020) dataset and available in `pyannote.audio` (Bredin et al., 2020). Voice activity detection and agglomerative clustering thresholds were tuned separately for each series, using silver-standard annotations of their respective seasons 2 and 3 (or movies 2 and 3 for Star Wars).
Because speaker timestamps are inferred from silver-standard timestamps (obtained by automatic forced

alignment), one should take those numbers with a grain of salt, especially as far as false alarm rate is concerned [1].

	DER%	FA%	Miss.%	Conf.%
Battlestar Galactica	63.6	28.1	15.3	20.2
Buffy the Vampire...	49.0	16.1	17.6	15.2
Friends	59.7	21.9	17.1	20.6
Game of Thrones	55.3	22.3	13.0	20.0
Lost	69.6	19.4	30.5	19.8
Star Wars	63.6	26.0	15.4	22.2
The Big Bang Theory	42.5	13.7	10.4	18.3
The Office	45.3	19.7	12.2	13.4
The Walking Dead	72.7	15.4	28.5	28.7
Overall gold-*Bazinga!*	57.9	20.3	17.8	19.8

Table 4: Diarization error rate (DER) on the *Bazinga! gold* subset using `pyannote.audio` speaker diarization pipeline (FA = false alarm, Miss. = missed detection, Conf. = speaker confusion).

Bazinga! can also be used for speaker identification, either in a supervised manner (using the *silver* subset for training speaker models), weakly supervised manner (using the episode-level list of speakers available in the knowledge base) or in an unsupervised manner using a combination of speaker diarization, named entity detection, and addressee detection (Bredin et al., 2014).

3.3. Automatic Speech Recognition

Although much larger datasets exist for training automatic speech recognition (ASR), the *Bazinga!* dataset contains multi-party conversations that might prove useful to evaluate ASR systems in close to real-life conditions. However, we leave it for future work.

3.4. Punctuation Restoration

As most automatic speech recognition systems provide output without punctuation, the task of punctuation restoration can improve the readability of transcripts and increases the effectiveness of subsequent processing. Punctuation restoration was performed on the manual transcript of the *Bazinga!* dataset thanks to bidirectional Recurrent Neural Network with attention mechanism proposed by Tilk and Alumäe (2016). Similar to what was done in the paper, the model was trained using the monolingual version of the Europarl corpus (v9) (Koehn, 2005) using hidden layers of 256 units and a learning rate of 0.02. In order to fit the TV series dialogues peculiarities (short sentences, interjection, etc.), a second model was trained on the whole *Bazinga!* silver subset using the same parameters.
The two models are applied on the *Bazinga!* gold subset and are evaluated in terms of overall precision and recall. The results, presented in Table 5, are significantly worse than the results obtained by Tilk and Alumäe (2016) on their English reference transcripts test set (Prec: 70.0, Rec: 59.7), when trained on the *Bazinga!* silver subset, for most TV series. Moreover, we can see a huge gap between the model trained on EU and the one trained on *Bazinga!*. This highlights the peculiarities of TV series listed above.

Tested on / Trained on	EU		*Bazinga!*	
	Prec	Recall	Prec	Recall
Battlestar Galactica	50.65	23.88	60.9	49.62
Breaking Bad	40.41	18.16	60.68	53.42
Buffy the Vampire...	41.32	17.11	60.62	48.42
Friends	39.13	16.23	58.89	46.05
Game of Thrones	48.39	24.84	62.65	52.17
Lost	44.73	19.47	64.33	55.16
Star Wars	52.80	26.10	59.20	46.80
The Big Bang Theory	43.25	21.02	59.49	54.75
The Office	43.67	17.73	64.25	55.53
The Walking Dead	28.53	16.25	41.13	47.17
Overall gold-*Bazinga!*	43.29	20.08	59.21	50.91

Table 5: Overall recall and precision for punctuation restoration on the *Bazinga!* gold subset using models trained on the Europarl corpus (EU) and the *Bazinga!* silver subset.

3.5. Person Entity Recognition

Named entity recognition for the entity type PERSON was also applied on the *Bazinga!* gold subset. Identifying the mentioned persons in the transcripts can be useful for several tasks in dialogue structuring (addressee detection, speaker diarization, etc.). Note that we use the gold-standard transcripts, which are case-sensitive, not the output of ASR. Two different named entity recognition models were used: Flair (Akbik et al., 2018) and spaCy[2]. Only the entity of type PERSON was considered for both of them, and the Recall and Precision metrics were used for evaluation. Table 6 shows first that spaCy performs better for this kind of data and, second, that the performance (especially recall) is lower for series that do not take place in the real world (Battlestar Galactica, Game of Thrones, and Star Wars). In comparison, the F1-score on the CONLL03 NER task for English is 93.1 for Flair and 91.6 for spaCy for all types of named entities. The results show the difficulty of transfer learning from models trained on mono-speaker data to our multi-party dialogues where the performance drops by 23.4 (resp. 18.5) points in F1 with Flair (resp. spaCy).

3.6. Entity Linking

The task of entity linking consists in assigning a unique identifier to words of interest in a text that can be named entities (locations, persons, organization, etc.), mentions, or surface forms. In the Character identification

[1] We compared *Bazinga!* silver-standard annotations with Serial Speakers (Bost et al., 2020) gold-standard annotations on season 1 of *Game of Thrones* and found that the former contains 20% of missed detection, 5% of false alarm, and very little speaker confusion ($< 1\%$).

[2] https://spaCy.io/

Tested on / Model	Flair		spaCy	
	Rec	Prec	Rec	Prec
Battlestar Galactica	0.48	0.78	0.53	0.83
Breaking Bad	0.71	0.66	0.75	0.76
Buffy the Vampire...	0.63	0.89	0.64	0.87
Friends	0.77	0.66	0.84	0.80
Game of Thrones	0.45	0.82	0.39	0.74
Lost	0.83	0.85	0.80	0.87
Star Wars	0.39	0.58	0.51	0.62
The Big Bang Theory	0.62	0.84	0.68	0.91
The Office	0.78	0.89	0.88	0.92
The Walking Dead	0.63	0.81	0.62	0.91
Overall gold-*Bazinga!*	0.63	0.78	0.66	0.82

Table 6: Recall and precision for person-named entity recognition with spaCy and Flair NER tools.

shared task, proposed by Choi and Chen (2018), the authors released the two first seasons of Friends annotated with 15,709 mentions and 401 entities. The *Bazinga!* gold set multiplies by four the number of annotated mentions (64,698) for entity linking and is more diverse. We do not provide a baseline for this task and leave it for future work, given that our setting is very different from traditional entity linking benchmarks and demands for specific research.

3.7. Addressee Detection

Addressee detection aims to answer the question *to whom is a speaker talking?* For natural language understanding, in the context of multi-party dialogues (TV series or movies dialogues, meeting recordings, dialogue system), it is crucial to understand to whom a sentence is addressed. To help answer this task, 73k+ sentences are manually annotated with addressee in the *Bazinga! gold* subset. We do not provide a baseline result for this task.

3.8. Continual Learning

Because TV series are built around a sequence of episodes happening in chronological order, most of the aforementioned tasks could also be addressed in a continual or incremental learning manner. For instance, as characters appear in (or disappear from) the storyline, the list of output classes of a speaker recognition or entity linking systems may evolve continuously.

4. Using *Bazinga!*

Both annotated transcripts and audio files can be downloaded using the open-source Python library `datasets`:

```
# install Huggingface "datasets" library
$ pip install datasets
```

Note that *Bazinga!* is hosted on `huggingface.co` as a private dataset. Please visit `hf.co/bazinga` to become a member of the *Bazinga!* organization. Once your membership request is approved, you can log into your account:

```
# authenticate with your Huggingface account
$ huggingface-cli login
```

and access the dataset from Python using the `datasets` library:

```
from datasets import load_dataset

# load "The Big Bang Theory" subset
dataset = load_dataset(
    "bazinga/bazinga",
    "TheBigBangTheory",
    use_auth_token=True)

# loop on the gold subset, episode by episode
for episode in dataset["gold"]:
    pass

# loop on the annotated transcript, word by word
for word in episode["transcript"]:

    # which word is pronounced?
    word["token"]       # "you"

    # when is "you" pronounced (in seconds)?
    word["start_time"]  # 151.880
    word["end_time"]    # 151.950

    # how confident is forced alignment
    word["confidence"]  # 0.99

    # who pronounces the word?
    word["speaker"]     # sheldon_cooper

    # who is the "sheldon_cooper" speaking to?
    word["addressee"]   # leonard_hofstadter

    # who is "you" referring to?
    word["entity"]      # leonard_hofstadter
```

5. Related Work

TV and movie series have been used in various datasets, for tasks ranging over many fields of academic research: from natural language processing to computer vision and speech processing. TV series emulate the continuity of life with episodes spanning over several years. Thus, continual learning strategies could be applied in this context. However, it has not been explored yet. Mahajan and Shaikh (2021) recently reviewed the multi-party dialogue literature, we focus here on datasets built upon TV and movie series.

Serial Speakers (Bost et al., 2020) is a dataset focused on *serials*, i.e. TV series with highly continuous plots between the episodes, such as *Game of Thrones*. Their dataset was also annotated semi-automatically using forced alignment following subtitles' Optical Character Recognition (OCR). While two out of three series of their dataset is included in *Bazinga!*, the authors focus on narrative structure and speaker diarization, leaving named entities and coreference resolution aside.

EmotionLines (Hsu et al., 2018) is a dataset for emotion detection in multi-party dialogues, built upon the scripts of *Friends* and *Facebook messenger* messages.

(Poria et al., 2019) extend this dataset by aligning it with the audio-visual tracks of *Friends*.

With *OpenSubtitles2016*, (Lison and Tiedemann, 2016) introduce a large translation dataset consisting of 1689 bitexts spanning 2.6 billion sentences across 60 languages. The translation bitexts were inferred from subtitles of hundred of thousands of movies or TV episodes based on the timestamp of the subtitle.

(Kepuska and Rojanasthien, 2011) introduce a toolkit able to generate Automatic Speech Recognition (ASR) corpora from DVDs. Their approach is similar to ours, although they extract subtitles from the DVDs using OCR instead of manual transcriptions. Then, they refine subtitles timestamps using forced alignment, aiming for tasks such as prosodic analysis.

In the context of SemEval 2018 (Choi and Chen, 2018), person-named entities of the first two seasons of Friends were manually annotated similarly to our work (see Section 2.7). The task focuses on the textual modality, discarding any audio-visual information. Moreover, it is limited to a single serie, thus genre.

With *MovieQA*, (Tapaswi et al., 2016) introduce a multimodal question-answering dataset that aims at evaluating automatic story comprehension from both video and text of 408 movies. Their dataset consists of thousands of multiple-choice questions, with five candidate answers, requiring different types of reasoning.

(Everingham et al., 2006) approach is quite similar to ours (see Section 2.5): they align the text of timestamped subtitles and manual transcripts with speaker names to get the '*who speaks when*' annotation, although this is far less precise than forced alignment. However they aim for *face recognition* rather than for *speaker identification*.

6. Conclusion and Future Work

We propose a new dataset, *Bazinga!*, for multi-party dialogues structuring. We expect its large scale to foster research towards unsupervised or weakly-supervised learning. Continual learning is also an interesting lead, given that TV and movie series often span over several years.

We provide baseline results for three tasks: speaker diarization, punctuation restoration, and person entity recognition. Those will ease the track of progress in multi-party dialogues structuring. We leave other tasks for future work.

Because of domain mismatch between the data used to train our baselines and the *Bazinga!* dataset, speaker diarization and person entity recognition show poor performance, highlighting the need for further research on unsupervised adaptation and transfer learning, along with more datasets of multi-party dialogues. We believe that the *Bazinga!* dataset is a step towards this goal.

Depending on the interest of the community, we hope to keep annotating the dataset so that the *gold* subset tends towards a complete coverage. Feel free to contact the corresponding author (herve.bredin@irit.fr) if you would like to help, either by contributing annotations, hosting crowd-sourcing annotation interfaces, or proposing new tasks (such as dialogue acts classification and topic segmentation).

7. Acknowledgements

Manual transcripts were scrapped from the following websites and are shared for research purposes only: `fandom.com`, `foreverdreaming.org`, `springfieldspringfield.co.uk`, `ageofthering.com`, `hypnoweb.net`. The use of audio tracks is allowed under the Fair Use Clause of the Copyright Law and for research purposes only. All the other annotations were created by the authors of the paper and are shared as Creative Commons.

This work was partly funded by the Digiteo Foundation through the StoryArcs project (2016-1267D) and by the French National Research Agency (ANR) through the PLUMCOT (ANR-16-CE92-0025) and the MEERQAT (ANR-19-CE23-0028) projects.

8. Bibliographical References

Akbik, A., Blythe, D., and Vollgraf, R. (2018). Contextual String Embeddings For Sequence Labeling. In *Proceedings of the International Conference on Computational Linguistics*, pages 1638–1649.

Bredin, H., Laurent, A., Sarkar, A., Le, V.-B., Rosset, S., and Barras, C. (2014). Person Instance Graphs for Named Speaker Identification in TV Broadcast. In *Odyssey 2014, The Speaker and Language Recognition Workshop*, Joensuu, Finland, June.

Bredin, H., Yin, R., Coria, J. M., Gelly, G., Korshunov, P., Lavechin, M., Fustes, D., Titeux, H., Bouaziz, W., and Gill, M.-P. (2020). pyannote.audio: neural building blocks for speaker diarization. In *Proc. ICASSP 2020*.

Choi, J. D. and Chen, H. Y. (2018). SemEval 2018 task 4: Character identification on multiparty dialogues. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 57–64.

Mahajan, K. and Shaikh, S. (2021). On the Need for Thoughtful Data Collection for Multi-Party Dialogue: A Survey of Available Corpora and Collection Methods. In *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 338–352, Singapore and Online, July. Association for Computational Linguistics.

Park, T. J., Kanda, N., Dimitriadis, D., Han, K. J., Watanabe, S., and Narayanan, S. (2022). A review of speaker diarization: Recent advances with deep learning. *Computer Speech & Language*, 72:101317, March.

Ryant, N., Singh, P., Krishnamohan, V., Varma, R., Church, K., Cieri, C., Du, J., Ganapathy, S., and

Liberman, M. (2020). The Third DIHARD Diarization Challenge. *arXiv preprint arXiv:2012.01477*.

Szymański, P., Żelasko, P., Morzy, M., Szymczak, A., Ży\la Hoppe, M., Banaszczak, J., Augustyniak, L., Mizgajski, J., and Carmiel, Y. (2020). WER we are and WER we think we are. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3290–3295, Online, November. Association for Computational Linguistics.

Tilk, O. and Alumäe, T. (2016). Bidirectional recurrent neural network with attention mechanism for punctuation restoration. In *Proceedings of the 17th Annual Conference of the International Speech Communication Association*.

9. Language Resource References

Bost, X., Labatut, V., and Linares, G. (2020). Serial speakers: a dataset of TV series. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4256–4264, Marseille, France, May. European Language Resources Association.

Choi, J. D. and Chen, H. Y. (2018). Semeval 2018 task 4: Character identification on multiparty dialogues. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 57–64.

Everingham, M., Sivic, J., and Zisserman, A. (2006). Hello! my name is... buffy"–automatic naming of characters in tv video. In *BMVC*, volume 2, page 6.

Hsu, C.-C., Chen, S.-Y., Kuo, C.-C., Huang, T.-H., and Ku, L.-W. (2018). EmotionLines: An Emotion Corpus of Multi-Party Conversations. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Kepuska, V. Z. and Rojanasthien, P. (2011). Speech corpus generation from dvds of movies and tv series. *Journal of International Technology and Information Management*, 20(1):4.

Koehn, P. (2005). Europarl: A parallel corpus for statistical machine translation. In *Proceedings of the 10th Machine Translation Summit Conference*, pages 79–86. International Association for Machine Translation.

Lhoest, Q., Villanova del Moral, A., Jernite, Y., Thakur, A., von Platen, P., Patil, S., Chaumond, J., Drame, M., Plu, J., Tunstall, L., Davison, J., Šaško, M., Chhablani, G., Malik, B., Brandeis, S., Le Scao, T., Sanh, V., Xu, C., Patry, N., McMillan-Major, A., Schmid, P., Gugger, S., Delangue, C., Matussière, T., Debut, L., Bekman, S., Cistac, P., Goehringer, T., Mustar, V., Lagunas, F., Rush, A., and Wolf, T. (2021). Datasets: A Community Library for Natural Language Processing. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 175–184, Online and Punta Cana, Dominican Republic, November. Association for Computational Linguistics.

Lison, P. and Tiedemann, J. (2016). Opensubtitles2016: Extracting large parallel corpora from movie and tv subtitles. In *Proceedings of the 10th Language Resources and Evaluation Conference*, pages 923–929.

Poria, S., Hazarika, D., Majumder, N., Naik, G., Cambria, E., and Mihalcea, R. (2019). MELD: A Multimodal Multi-Party Dataset for Emotion Recognition in Conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 527–536.

Tapaswi, M., Zhu, Y., Stiefelhagen, R., Torralba, A., Urtasun, R., and Fidler, S. (2016). Movieqa: Understanding stories in movies through question-answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4631–4640.