TVD: a reproducible and multiply aligned TV series dataset

Anindya Roy[1], Camille Guinaudeau[1,2], Hervé Bredin[1], Claude Barras[1,2]
[1]Spoken Language Processing Group, CNRS-LIMSI, B.P. 133, Orsay, France.
[2]Université Paris Sud, Orsay, France.
{roy, guinaudeau, bredin, barras}@limsi.fr

Abstract

We introduce a new dataset built around two TV series from different genres, *The Big Bang Theory*, a situation comedy and *Game of Thrones*, a fantasy drama. The dataset has multiple tracks extracted from diverse sources, including dialogue (manual and automatic transcripts, multilingual subtitles), crowd-sourced textual descriptions (brief episode summaries, longer episode outlines) and various metadata (speakers, shots, scenes). The paper describes the dataset and provide tools to reproduce it for research purposes provided one has legally acquired the DVD set of the series. Tools are also provided to temporally align a major subset of dialogue and description tracks, in order to combine complementary information present in these tracks for enhanced accessibility. For alignment, we consider tracks as comparable corpora and first apply an existing algorithm for aligning such corpora based on dynamic time warping and TFIDF-based similarity scores. We improve this baseline algorithm using contextual information, WordNet-based word similarity and scene location information. We report the performance of these algorithms on a manually aligned subset of the data. To highlight the interest of the database, we report a use case involving rich speech retrieval and propose other uses.

Keywords: TV Series, transcripts, episode outlines.

1. Introduction

Rich in dialogue content and enjoying a wide fan base, TV series such as *The Big Bang Theory* and *Game of Thrones* are a potential source of data for both NLP applications (*e.g.* summarization) and information retrieval tasks (*e.g.* fans may like to retrieve all scenes where *Leonard invites Penny to dinner* or *Sheldon apologizes to Amy*).

However, given the obvious copyright restrictions, there are relatively few language resources directly built around them. Instead, researchers working on TV series typically provide experimental results and at most open-source versions of their algorithms with pre-extracted features (Laptev et al., 2008; Everingham et al., 2009). However, without the original dataset and associated annotations, it is difficult to reproduce or compare published results.

Data associated with TV series is multimodal. In addition to speech and visual data, there is an increasing amount of crowd-sourced textual content related to TV series on the web including manual transcriptions (The Big Bang Theory Transcripts, 2013; Game of Thrones Transcripts, 2013), detailed episode outlines (The Big Bang Theory Wiki, 2013; Game of Thrones Wiki, 2013) along with comments and discussions on Internet forums (The Big Bang Theory Forums, 2013). It would be interesting to combine these heterogeneous sources of data and present them in an integrated, self-connected framework staying within the limits imposed by copyright laws. This work is an effort in this direction.

We focus on three contributions. First, we acquire raw data related to two TV series from various sources (both crowd-sourced and commercial). We parse and structure the data into distinct tracks. Since copyright restrictions apply on part of the data, we make it reproducible by providing computer scripts to reproduce the tracks *locally* for research purposes, provided one has legally acquired the original source of the data. Second, we align corresponding units from different tracks (*e.g.* an event in an episode outline with all dialogue lines corresponding to that event). For this, we extend existing text alignment algorithms to better suit the data by using context-dependent TFIDF features and WordNet-based word similarity (Miller, 1995). This leads to a richer, multiply connected, more accessible representation of the data. Third, we propose an interesting use case of the aligned dataset: searching for high-level concepts such as "X *tries to convince* Y" which would be difficult using the dialogue tracks alone.

2. Related work

Datasets Existing TV series or movie-related datasets usually consist of *visual* metadata in terms of person identity (Everingham et al., 2009; Cour et al., 2009; Sivic et al., 2009; Tapaswi et al., 2012), pose (Ferrari et al., 2009) and physical actions (Laptev et al., 2008; Patron-Perez et al., 2010). On the contrary, easily reproducible/freely available *linguistically*-oriented datasets related to TV series/movies are harder to find. User review datasets for sentiment analysis (Maas et al., 2011) could be an example. But the dataset closest to the one in this work would be (Rich Speech Retrieval Dataset, 2011).[1] However, that dataset consists of videos from the `blip.tv` website. In contrast, this work involves two TV series.

Alignment of monolingual written texts has been studied in the context of comparable corpora *e.g.* Britannica corpus (Barzilay and Elhadad, 2003; Nelken and Shieber, 2006) and news articles (Quirk et al., 2004; Klerke and Søgaard, 2012). Related work specifically on TV series or movies may be grouped into the following:

1. Alignment of *text* to *video* (Laptev et al., 2008; Ronfard and Thuong, 2003; Cour et al., 2008; Sankar et al., 2009)

[1]`www.opensem.eu/rich-speech-retrieval-218.html`

2. Alignment of *text* to *text* (manual transcripts to subtitles (Everingham et al., 2006) or bilingual subtitles (Tsiartas et al., 2013)).

We also focus on text to text alignment but our main contribution is the alignment of dialogue to *crowd-sourced external texts* such as *episode outlines* which add valuable linguistic information.

To the best of our knowledge, there is one other major work which deals with a similar task (Li et al., 2008). This uses *local* constraints for alignment instead of maximizing a *global* objective function as we do. Furthermore, their approach is limited to TFIDF features while we investigate the use of context-weighted TFIDF, WordNet-based word similarity, scene locations and their combination, showing how such additional information can improve performance. Also, they did not adequately specify the dataset they used making it difficult to reproduce results while we provide complete details about our dataset.

Use case (Aly et al., 2011) studied a rich speech retrieval task somewhat similar to the high-level concept retrieval task explored here but the data and query set used were different. (Rich Speech Retrieval Dataset, 2011)

3. Description of the dataset

The current version of the dataset focuses on two popular TV series, *The Big Bang Theory* (TBBT) (The Big Bang Theory Wiki, 2013) and *Game of Thrones* (GoT) (Game of Thrones Wiki, 2013).[2] A website has been created for the dataset: **`tvd.niderb.fr`** . Computer scripts required to reproduce the dataset locally are hosted at this website along with usage instructions. Table 1 provides an overview of the dataset while the following sections describe each track in detail.

3.1. Manual transcripts (MTR)

Manual transcripts contain exact transcription of dialogue lines and speaker identities. For TBBT, scene locations are also provided. Here is an extract of manual transcripts from an episode of TBBT:

```
Scene location: A Chinese restaurant.
Sheldon: I'm sorry, we cannot do this without Wolowitz.
Leonard: We can't order Chinese food without Wolowitz?
```

and from an episode of GoT:

```
Jorah Mormont : You need to drink, child. And eat.
Daenerys Targaryen : Isn't there anything else ?
```

We provide computer scripts to automatically download and parse manual transcripts from (The Big Bang Theory Transcripts, 2013; Game of Thrones Transcripts, 2013) for all 132 episodes of the first 6 seasons of TBBT and the first 5 episodes of GoT.

3.2. Subtitles (SUB)

While manual transcripts are mono-lingual and do not contain temporal information, DVD subtitles contain a close approximation of the dialogue (sentences often simplified to fit on-screen space contraints) in the original language (in this case, English) *and* some other local languages, and time stamps at the sentence level. Here is an extract of subtitles from an episode of TBBT:

```
00:13:21,520 -> 00:13:24,318
I'm sorry.
we cannot do this without Wolowitz.

00:13:24,480 -> 00:13:27,278
We can't order Chinese food
without Wolowitz?
```

Here is an extract of subtitles from an episode of GoT:

```
00:02:06,520 --> 00:02:07,953
You need to drink, child.

00:02:08,079 --> 00:02:09,990
(She sighs)

00:02:12,159 --> 00:02:14,070
And eat.
```

We provide computer scripts to automatically extract subtitles from DVD. For this, an optical character recognition step is applied as subtitles are actually stored as overlaid images on DVDs.

3.3. Automatic transcripts (ATR)

Similar to subtitles, DVDs usually contain audio tracks in multiple languages: one track in the original language (English in our case) and dubbed audio tracks in local languages. We ran an *off-the-shelf* automatic transcription system on all audio tracks in the DVD. As shown in the excerpt below, automatic transcripts provide word-level temporal alignment. It also groups words into speech turns associated with dummy speaker tags (such as FS8 and MS3 below) but not actual identities. Here is an extract of automatic transcripts from an episode of TBBT:

```
<SpeechSegment ch="1" spkid="FS8" lang="eng-usa">
<Word stime="250.43" dur="0.17" conf="0.794"> Why </Word>
<Word stime="250.66" dur="0.20" conf="0.834"> not </Word>
</SpeechSegment>
```

and from an episode of GoT:

```
<SpeechSegment ch="1" spkid="MS3" lang="eng-usa">
<Word stime="156.10" dur="0.07" conf="0.398"> The </Word>
<Word stime="156.60" dur="0.28" conf="0.393"> lands </Word>
<Word stime="156.95" dur="0.20" conf="0.515"> beyond </Word>
</SpeechSegment>
```

As transcripts may be considered derivative works of protected content, we do not actually own this data. However, as audio tracks are available in the TVD corpus, readers could use freely available text-to-speech tools (e.g. Kaldi (Povey et al., 2011)) to re-compute this data.

3.4. Episode outlines (OL) and summaries (SUM)

We provide computer scripts to automatically download and parse episode outlines and summaries from (The Big Bang Theory Wiki, 2013) for 69 episodes of TBBT. For GoT, we provide scripts to download episode outlines for 17 episodes and summaries for 30 episodes from (Game of Thrones Wiki, 2013). In each case, the corresponding content license is respected. Outlines provide a chronologically ordered list of main events taking place in each episode. For TBBT, we additionally have scene locations (as in manual transcripts) in the outline.

[2] Other series are planned to be covered in later versions of the dataset.

Track	# Episodes TBBT	# Episodes GoT	Type	Manual	Time-stamped	Multi-lingual	Identity	Location	Original source
Manual transcripts	132	5	dialogue	✔			✔	✔	WWW
Subtitles	17	10	dialogue	✔	✔	✔			DVD
Automatic transcripts	17	10	dialogue		✔	✔			speech-to-text
Episode outlines	69	17	description	✔				✔	WWW
Summaries	69	30	description	✔				✔	WWW
Speech turns	6	_	metadata	✔	✔		✔		(Tapaswi et al., 2012)
Face tracks	6	_	metadata	✔	✔		✔		(Tapaswi et al., 2012)
Shots	6	_	metadata	✔	✔				(Bäuml et al., 2013)

Table 1: Overview of each track in the TVD dataset. Number of episodes reflects the status of the dataset at the time of writing this paper.

Here is an extract from an outline of an episode from TBBT:

```
Scene location: Hallway outside apartments
Event: Penny gives Leonard the key to her apartment.
Event: The four guys get into a discussion about
Superman's flight skills in front of Penny.
Scene location: Hallway
Event: Leonard invites Penny over.
```

and from an outline of an episode from GoT:

```
Event: Khal Drogo's khalasar is several days from
Pentos crossing the plains known as the Flatlands...
Event: The Dothraki make camp and Daenerys is helped
from her horse by Ser Jorah and her handmaidens...
```

Events are expressed using either words extracted from the dialogue (extractive summarization) such as 'Sheldon tells his "*good idea for a T-shirt*"' or descriptive phrases (abstractive summarization). Such phrases may describe a *rich* (high-level) *speech act*: "tries to convince", "apologizes to" and "tries to work out how", a *physical act*: "hugs", "make camp" and "helped from (her) horse", or a *state of mind*: "is shocked", "is not impressed by". The abstractive case is important because it brings new information not explicitly stated in the dialogue.

On the contrary, summaries are concise descriptions of each episode. Here is an extract from a summary of an episode from TBBT:

```
When Leonard and Sheldon meet Penny, Leonard is
immediately interested in her (saying "our babies
will be smart and beautiful"), but Sheldon feels
his friend is chasing a dream he'll never catch
(adding "not to mention imaginary")...
```

and from a summary of an episode from GoT:

```
Three rangers of the Night's Watch: Ser Waymar Royce,
Will, and Gared depart from the Wall to investigate
reports of wildlings in the Haunted Forest which
lies to the north...
```

3.5. Shots, speech turns and face tracks

Additionally, manual annotations of both audio stream (who speaks when) and visual stream (shot boundaries, face tracks, face identity) for first 6 episodes of TBBT are included (Tapaswi et al., 2012; Bäuml et al., 2013). Example:

```
53.2314 2.7039 leonard
57.7743 2.1954 other
61.5371 1.1235 sheldon
```

Henceforth, we denote transcripts and subtitles as **dialogue** tracks, summary and episode outline tracks as **description** tracks, and shots, speech turns and face tracks as **annotation** tracks.

Each track is composed of a sequence of chronologically ordered *units*. Depending on the type of track, a unit may be a dialogue line, a word or an event.

In the TVD dataset, we conveniently represent tracks as components of a linguistic annotation graph (Bird and Liberman, 2001). In this directed acyclic graph, each vertex represents a time point within a TV series. In the case of DVD subtitles and automatic transcripts, the position of a such a point is known precisely in time. In the case of manual transcripts and episode outlines, it is unknown. In these cases, however, a time point is usually constrained to lie *between* other points conforming to a chronological ordering.

Each unit is represented by a directed edge between two vertices corresponding to the starting and ending times of the unit. Figure 1 shows a compact visualization of such a graph corresponding to an episode from *The Big Bang Theory*. For clarity, only MTR, OL and SUM tracks are shown, represented by red, green and blue components (vertices and edges) respectively. Alignment between two tracks is indicated via vertices and edges *in common* between the sub-graphs corresponding to these tracks. These common components shown in black.

4. Alignment of tracks

Some tracks like manual transcripts and episode outlines do not have explicit timing information while others like subtitles and automatic transcripts do. To enhance the usability of the dataset, we provide scripts to automatically align the following pairs of tracks:[3]

1. **Manual transcripts** (MTR) ⇔ **subtitles** (SUB): merges time-stamps from SUB with exact dialogue and speaker identity from MTR.
2. **Subtitles** (SUB) ⇔ **automatic transcripts** (ATR): enhances time resolution from *sentence*-level (SUB) to *word*-level (ATR).
3. **Episode outlines** (OL) ⇔ **manual transcripts** (MTR): merges speaker and dialogue lines from MTR with event descriptions from OL. Figure 1 shows an example of such an alignment.

[3]hosted on the dataset website **tvd.niderb.fr** .

Episode n°	1	2	3	4	5	ALL
OL	35	31	27	26	26	145
MTR	315	281	347	395	424	1762

Table 2: N° of units in episode outlines (OL) and manual transcripts (MTR) of the first 5 episodes of GoT.

Episode n°	1	2	3	4	5	6	all
OL	17	15	15	26	19	13	105
MTR	326	244	250	229	228	255	1532

Table 3: N° of units in episode outlines (OL) and manual transcripts (MTR) of the first 6 episodes of TBBT.

Later versions of the dataset are planned to include scripts to automatically align *other* pairs of tracks *e.g.* summaries (SUM) with manual transcripts (MTR).

4.1. Baseline algorithm

Our baseline algorithm for track alignment follows the standard approach in (Nelken and Shieber, 2006) based on dynamic time warping (DTW) for aligning comparable corpora. Let $U_{1:N} \equiv \{u_{1,1}, \cdots, u_{1,N}\}$ and $U_{2:M} \equiv \{u_{2,1}, \cdots, u_{2,M}\}$ be two tracks. Each u is a unit to be aligned.[4] Let $s(i,j)$ be the *local* similarity score between unit $u_{1,i}$ and $u_{2,j}$ for $1 \leq i \leq N,\ 1 \leq j \leq M$. Then the *global* alignment score $\mathscr{S}(i,j)$ is calculated as:

$$\mathscr{S}(i,j) = \max \begin{cases} \mathscr{S}(i-1,j-1) + s(i,j) \\ \mathscr{S}(i-1,j) + s(i,j) \\ \mathscr{S}(i,j-1) + s(i,j) \end{cases} \quad (1)$$

Once all scores $\{\mathscr{S}(i,j)\}$ have been computed, we find the best alignment path by backtracking from $\mathscr{S}(N,M)$ to $\mathscr{S}(1,1)$.

For the first and second alignment tasks (MTR $\Leftrightarrow$ SUB and SUB $\Leftrightarrow$ ATR respectively), the units u are words and we define the local similarity score $s(i,j)$ as 1 iff the words $u_{1,i}$ and $u_{2,j}$ are identical, and 0 otherwise:

$$s(i,j) = \mathbf{1}_{\{u_{1,i}=u_{2,j}\}} \quad (2)$$

This effectively transforms these tasks into the longest common subsequence problem (Myers, 1986). This relatively straightforward problem has been successfully addressed in the context of subtitle alignment. For a description, please consult (Everingham et al., 2006).

In this paper, we instead focus on the third (and most challenging) of the three alignment tasks: OL $\Leftrightarrow$ MTR. In this case, units $\{u_{1,i}\}$ are events in OL and $\{u_{2,j}\}$ are dialogue lines in MTR. Note that words in OL may have limited overlap with words in MTR (ref. abstractive summarization in Section 3.4.). Hence, just using the number of words in common as in Equation 1 might not be a good solution (Nelken and Shieber, 2006). Instead, we define the local similarity score $s(i,j)$ as the probability that unit $u_{1,i}$ should be aligned with unit $u_{2,j}$ estimated by fitting a logistic regression model to the cosine similarity between TFIDF vectors of $u_{1,i}$ and $u_{2,j}$ (Nelken and Shieber, 2006). We denote this algorithm as `TFIDF`. We compare `TFIDF` with a `NAIVE` alignment algorithm which groups subsequent dialogue lines in MTR into N equal-sized groups and aligns each group with an event in OL.

In this paper, these alignment algorithms are evaluated using the first 5 episodes of GoT and first 6 episodes of TBBT. The total number of OL and MTR units used in the evaluation are 145 and 1762 for GoT and 105 and 1532 for TBBT respectively. Episode-wise breakup of these numbers are shown in Tables 2 and 3 for the two TV series.

In Tables 5 and 7, performance is reported for GoT and TBBT in terms of alignment accuracy, *i.e.* ratio of number of dialogue line units in MTR aligned correctly with their corresponding event units in OL to total number of dialogue lines in MTR expressed as a percentage (Sankar et al., 2009).[5] `TFIDF` outperforms `NAIVE` on both GoT and TBBT, its accuracy being quite high on GoT. `TFIDF` accuracy on TBBT is lower than GoT probably because TBBT events are described by one sentence while GoT events are typically described by 5 to 10 sentences. In Sections 4.2.-4.4., we investigate 3 ways to increase accuracy on TBBT.

	GoT episode n°					
Algorithm	1	2	3	4	5	all
`NAIVE`	22.2	29.0	13.8	31.9	28.3	25.1
`TFIDF`	95.9	93.9	99.4	97.2	97.4	96.8

Table 4: Alignment accuracy on first 5 episodes of GoT using `NAIVE` and `TFIDF`.

	TBBT episode n°						
Algorithm	1	2	3	4	5	6	all
`NAIVE`	20.1	42.0	42.1	27.2	33.3	36.4	33.5
`TFIDF`	31.4	60.5	69.1	65.0	73.7	52.7	58.8

Table 5: Alignment accuracy on first 6 episodes of TBBT using `NAIVE` and `TFIDF`.

4.2. Contextual information

Instead of considering only terms inside the current unit u, we take a context of fixed length around current unit u and weight term frequencies inside this context via a window function centred on u. This algorithm is denoted as Context-dependent `TFIDF` or `CTFIDF`. Table 6 compares the alignment accuracy of 5 best-performing window functions. The asymmetric window $[1\ 1.25\ 0]$ performs the best. Table 7, rows 2 and 3 shows `CTFIDF` outperforms `TFIDF` for 4 out of 6 episodes and by 3.7% overall.

4.3. WordNet

To exploit semantic similarity between non-identical words ignored by TFIDF, we add Wu-Palmer similarities (Wu

[4] Note that the tracks are assumed to satisfy the three conditions for DTW: (1) boundary constraint (the first and last units in both tracks much be aligned), (2) temporal monotonicity (units are strictly chronologically ordered) and (3) temporal continuity.

[5] For reference, manual alignments were produced for all episodes used in the evaluation.

Episode n°	1	2	3	4	5	6	all
[0 2 1]	44.0	61.5	67.4	64.6	79.8	53.6	61.8
[1 2 0]	46.9	**63.4**	**68.7**	60.2	**80.8**	53.6	62.2
[1 1.25 0]	46.9	**63.4**	**68.7**	**61.2**	**80.8**	**54.0**	**62.5**
[1 2 1]	47.5	62.9	67.4	57.8	79.8	54.0	61.6
[1 2 4 2 1]	**51.6**	60.0	65.7	56.8	77.3	53.1	60.7

Table 6: Alignment accuracy using CTFIDF with 5 best-performing window functions on TBBT. Top accuracy for each episode in **bold**.

and Palmer, 1994) computed over WordNet senses (Miller, 1995; Bird et al., 2009) of *all* word pairs $\{(w_1, w_2) : w_1 \in u_{1,i} \wedge w_2 \in u_{2,j}\}$ to TFIDF cosine similarity.

Results are reported in Table 7. Note that that this algorithm TFIDF.W (row 4) outperforms TFIDF (row 2) overall. However, coupled with CTFIDF (CTFIDF.W, row 5), this algorithm cannot improve on CTFIDF (row 3).

4.4. Scene locations

TBBT OL and MTR tracks contain scene locations which could be useful for alignment *e.g. A Chinese restaurant*, *Hallway outside appartments* (ref. Sections 3.1., 3.4.). One way to incorporate scene locations would be to first align OL and MTR tracks at the scene level using scene locations, and then, align units inside each (aligned) scene.[6]

However, since OL and MTR are crowd-sourced and not created by a single person or group, their terminology is seldom consistent: *e.g.* scene location *living room* in OL is equivalent to *Sheldon and Leonard's apartment* in MTR. Also, scene locations may be arbitrarily dropped leading to unequal number of scene locations in OL and MTR. So, trying to first align tracks at the level of scene locations was found to be impractical.

Instead, we simply add scene locations to the set of units to be aligned, in the same way as dialogue lines in MTR and events in OL. However, a distinction between scene location units and *other* units is made by weighting local similarity scores between them differently, as follows:

$$\tilde{s}(i,j) \leftarrow \quad 1.2 \times s(i,j) \text{ if } u_{1,i} \in \mathbf{U}_L \wedge u_{2,j} \in \mathbf{U}_L \quad (3a)$$
$$\tilde{s}(i,j) \leftarrow \quad 0.1 \times s(i,j) \text{ if } u_{1,i} \in \mathbf{U}_L \wedge u_{2,j} \notin \mathbf{U}_L \quad (3b)$$
$$\tilde{s}(i,j) \leftarrow \quad 0.1 \times s(i,j) \text{ if } u_{1,i} \notin \mathbf{U}_L \wedge u_{2,j} \in \mathbf{U}_L \quad (3c)$$
$$\tilde{s}(i,j) \leftarrow \quad 1.0 \times s(i,j) \text{ if } u_{1,i} \notin \mathbf{U}_L \wedge u_{2,j} \notin \mathbf{U}_L \quad (3d)$$

where $\mathbf{U}_L$ is the set of scene location units. Note that the similarity score between a scene location and a non-scene location unit is given a non-zero (but low) weight to account for the fact that scene locations are sometimes mentioned in the dialogue. All four weights have been tuned manually to optimize performance.

Results are reported in Table 7. Note that scene locations coupled with CTFIDF (CTFIDF.S) outperforms CTFIDF by 3.3% overall. Scene locations enhanced with WordNet (CTFIDF.SW) further improves accuracy by 3.5% leading to the highest alignment accuracy among all algorithms. WordNet seems to work better on the restricted vocabulary of scene locations than on the entire text (Section 4.3.).

[6]Note that this approach is similar to the two-step approach for aligning comparable corpora involving paragraph matching followed by sentence matching in (Barzilay and Elhadad, 2003).

5. Use cases

Exploitation of speech transcripts for information retrieval is not new (de Jong et al., 2007). A novel feature of our work is to automatically align *external* crowd-sourced textual descriptions to speech transcripts from TV series, thus boosting its accessability. Next, we describe a simple use case to illustrate this point. Episode outlines provide compact textual descriptions at the scene or sub-scene level, *e.g*,

```
Leonard apologizes to Penny who accepts it.
```

The key terms *apologize* and *accept (the apology)* do not occur in the corresponding extract from the manual transcript (abridged):

```
Penny: What's going on?
Leonard: ... I regret my participation in what was, at
the very least, an error in judgement. The hallmark of
the great human experiment is the willingness to reco-
gnise one's mistakes. Some mistakes, ...
Penny: Leonard.
Leonard: Yeah.
Penny (hugs him): We're okay.
```

This observation shows that episode outlines may be more useful for scene retrieval when searching using high-level concepts such as rich speech acts ("X *apologizes* to Y") or states of mind ("X *isn't impressed* by Y") as listed in column 1 of Table 8, in contrast to transcriptions. In fact, our dataset is well-suited for this because we automatically align outlines with manual transcripts, and manual transcripts with DVD subtitles/automatic transcripts, thus transferring time-stamps from the latter to the former. This allows us to match each outline with its corresponding scene in the video.

In an initial toy experiment, we first transform each concept into a small set of query terms combined with AND ('.') or OR ('+') if needed, as shown in column 2 of Table 8. We search for these terms inside episode outlines and manual transcripts and retrieve all units containing these terms. In this experiment, we assume that alignment with video is perfect. In Table 8, we report results using episode outlines (columns 3-5) and manual transcripts (columns 6-8). The retrieved scenes were checked manually for relevance.[7] We observe that the number of relevant scenes retrieved as well as the precision is consistently higher with episode outlines than with transcripts. We are in the process of building a complete scene retrieval system starting from this initial idea.[8]

In addition to scene retrieval, this dataset may be used to study a variety of NLP and machine learning tasks such as automatic summarization (extractive and abstractive) (Liu and Liu, 2013), generation of natural language video descriptions (Krishnamoorthy et al., 2013), training linguistic and visual concept detectors (Laptev et al., 2008; de Jong et al., 2007), person identification (Bredin, 2012; Bredin et

[7]We consider X and Y to represent any person.

[8]Here, we have not addressed queries based on *physical* actions ("X *hugs* Y") as these have been studied elsewhere (Laptev et al., 2008; Patron-Perez et al., 2010). However, the same approach may be used for such queries too.

	TBBT episode n°						
Algorithm	1	2	3	4	5	6	all
NAIVE	20.1	42.0	42.1	27.2	33.3	36.4	33.5
TFIDF	31.4	60.5	69.1	65.0	73.7	52.7	58.8
CTFIDF	**46.9**	63.4	68.7	61.2	**80.8**	54.0	62.5
TFIDF.W	43.1	60.0	**72.1**	64.1	69.7	51.9	60.1
CTFIDF.W	**46.9**	63.4	68.7	64.6	76.8	54.0	62.4
CTFIDF.S	42.5	**66.8**	70.8	**66.0**	**80.3**	**68.6**	**65.8**
CTFIDF.SW	**43.4**	**72.7**	**78.1**	**72.8**	**80.3**	**68.6**	**69.3**

Table 7: Alignment accuracy on first 6 episodes of TBBT using various algorithms. Top 2 accuracies for each episode in **bold**.

		Outlines			Transcripts		
Concept	Query term(s)	# retrieved	# relevant	P (%)	# retrieved	# relevant	P (%)
X invites Y	invite + invitation	25	20	80.0	28	6	21.4
X apologises (to Y)	apologise + apology	28	19	67.9	60	19	31.7
X isn't impressed (by Y)	(not.impress) + unimpress	5	5	100.0	1	0	0.0
X tries to convince Y	try.convince	3	3	100.0	0	0	0.0
X is shocked	shock	8	8	100.0	7	3	42.9
X is surprised	surprise	8	5	62.5	30	8	26.7
X discusses (with Y)	discuss	38	38	100.0	22	5	22.7

Table 8: Scene retrieval results in terms of number of scenes retrieved, number of relevant scenes retrieved and precision (P) on the first 69 episodes of TBBT using high-level concepts as queries. X and Y may be any person, '.' → AND, '+' → OR, verbs were lemmatized, partial matches were allowed where suitable ('unimpress' → 'unimpressed', 'unimpressive'). Please see Section 5. for details.

al., 2012; Roy et al., 2014) and plot de-interlacing of TV series (Ercolessi et al., 2012). It is noteworthy that *Game of Thrones* closely follows a collection of fantasy novels from which it is adapted and could therefore lead to novel cross-media analysis techniques.

6. Conclusions

We present a new multi-track TV series dataset and provide computer scripts to locally regenerate it from source, conforming to copyright and license restrictions. We align tracks in the dataset and show that context-dependent TFIDF, scene locations and WordNet improve alignment accuracy.

The alignment of crowd-sourced external content (*e.g.* episode outlines) with time-stamped speech transcripts opens up new possibilities for scene retrieval using high-level concepts including rich speech acts (Rich Speech Retrieval Dataset, 2011) ("X *invites* Y", "X *tries to convince* Y"). Such concepts are often explicitly mentioned in episode outlines but not in speech transcripts. Initial experiments based on this idea yielded promising results.

The dataset may also be used for other tasks: summarization, training concept detectors and person identification.

7. Acknowledgements

This work was done in the context of the QCOMPERE project (funded by ANR) and the CHIST-ERA CAMOMILE project (funded by ANR, FNR and Tübitak).

8. References

Aly, R. B. N., Verschoor, T., and Ordelman, R. J. F. (2011). Utwente does rich speech retrieval at mediaeval 2011. In *Working Notes Proceedings of the MediaEval 2011 Workshop, Pisa, Italy*, volume 807, Aachen, Germany.

Barzilay, R. and Elhadad, N. (2003). Sentence alignment for monolingual comparable corpora. In *Proc. of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Bäuml, M., Tapaswi, M., and Stiefelhagen, R. (2013). Semi-supervised Learning with Constraints for Person Identification in Multimedia Data. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Bird, S. and Liberman, M. (2001). A formal framework for linguistic annotation. *Speech Communications*, 33(1-2):23–60.

Bird, S., Klein, E., and Loper, E. (2009). *Natural Language Processing with Python*. O'Reilly Media Inc.

Bredin, H., Poignant, J., Tapaswi, M., Fortier, G., Le, V., Napoleon, T., Gao, H., Barras, C., Rosset, S., Besacier, L., Verbeek, J., Quénot, G., Jurie, F., and Ekenel, H. (2012). Fusion of Speech, Faces and Text for Person Identification in TV Broadcast. In *Proc. of Workshop on Information Fusion in Computer Vision for Concept Recognition (ECCV 2012)*, October.

Bredin, H. (2012). Segmentation of TV Shows into Scenes using Speaker Diarization and Speech Recognition. In *Proc. of ICASSP 2012, IEEE International Conference on Acoustics, Speech, and Signal Processing*, Kyoto, Japan, March.

Cour, T., Jordan, C., Miltsakaki, E., and Taskar, B. (2008). Movie/Script: Alignment and Parsing of Video and Text Transcription. In *Proc. of European Conference on Computer Vision (ECCV)*.

Cour, T., Sapp, B., Jordan, C., and Taskar, B. (2009). Learning from Ambiguously Labeled Images. In *Proc. of the IEEE Conference on Computer Vision and Pattern*

Recognition (CVPR).

de Jong, F., Westerveld, T., and de Vries, A. (2007). Multimedia Search Without Visual Analysis: The Value of Linguistic and Ccontextual Iinformation. *IEEE Transactions on Circuits and Systems for Video Technology*, 17(3):365–371.

Ercolessi, P., Bredin, H., and Sénac, C. (2012). StoViz: Story Visualization of TV Series. In *Proc. of 20th ACM International Conference on Multimedia (ACM MM)*.

Everingham, M., Sivic, J., and Zisserman, A. (2006). "Hello! My name is... Buffy" - automatic naming of characters in TV video. In *Proc. of British Machine Vision Conference*.

Everingham, M., Sivic, J., and Zisserman, A. (2009). Taking the bite out of automatic naming of characters in TV video. *Image and Vision Computing*, 27(5).

Ferrari, V., Marin-Jimenez, M., and Zisserman, A. (2009). Pose search: Retrieving people using their pose. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Game of Thrones Transcripts. (2013). Game of Thrones Transcripts. `foreverdreaming.org/viewforum.php?f=67`. Accessed: 2013-10-23.

Game of Thrones Wiki. (2013). Game of Thrones Wiki. `gameofthrones.wikia.com/`. Accessed: 2013-10-23.

Klerke, S. and Sogaard, A. (2012). DSim, a Danish Parallel Corpus for Text Simplification. In *Proc. of the International Conference on Language Resources and Evaluation (LREC)*.

Krishnamoorthy, N., Malkarnenkar, G., Mooney, R., Saenko, K., and Guadarrama, S. (2013). Generating natural-language video descriptions using text-mined knowledge. In *Proc. of the 27th AAAI Conference on Artificial Intelligence (AAAI-2013)*, pages 541–547, July.

Laptev, I., Marszalek, M., Schmid, C., and Rozenfeld, B. (2008). Learning Realistic Human Actions from Movies. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Li, M., Liu, Y., Zhang, Y.-D., and Lin, S.-X. (2008). Synopsis alignment: Importing external text information for multi-model movie analysis. In *Advances in Multimedia Information Processing - PCM 2008*, volume 5353 of *Lecture Notes in Computer Science*, pages 287–295. Springer Berlin Heidelberg.

Liu, F. and Liu, Y. (2013). Towards Abstractive Speech Summarization: Exploring Unsupervised and Supervised Approaches for Spoken Utterance Compression. *IEEE Transactions on Audio, Speech and Language Processing*, 21(7).

Maas, A., Daly, R., Pham, P., D., H., Ng, A., and Potts, C. (2011). Learning Word Vectors for Sentiment Analysis. In *Proc. of the 49th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Miller, G. (1995). WordNet: A Lexical Database for English. *Communications of the ACM*, 38(11):39–41.

Myers, E. (1986). An O(ND) Difference Algorithm and its Variations. *Algorithmica*, 1(2):251–266.

Nelken, R. and Shieber, S. (2006). Towards robust context-sensitive sentence alignment for monolingual corpora. In *Proc. of the 11th Conference of the European Chapter of the ACL (EACL)*.

Patron-Perez, A., Marszalek, M., Zisserman, A., and Reid, I. (2010). High Five: Recognising human interactions in TV shows. In *Proc. of British Machine Vision Conference*.

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., Silovsky, J., Stemmer, G., and Vesely, K. (2011). The kaldi speech recognition toolkit. In *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*, December.

Quirk, C., Brockett, C., and Dolan, W. (2004). Monolingual machine translation for paraphrase generation. In *In Proc. of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 142–149.

Rich Speech Retrieval Dataset. (2011). Rich Speech Retrieval Dataset. `http://www.opensem.eu/rich-speech-retrieval-218.html`. Accessed: 2013-10-20.

Ronfard, R. and Thuong, T. (2003). A framework for aligning and indexing movies with their script. In *Proc. of International Conference on Multimedia and Expo (ICME)*.

Roy, A., Bredin, H., Hartmann, W., Le, V., Barras, C., and Gauvain, J.-L. (2014). Lexical Speaker Identification in TV Shows. *Accepted for publication in Springer Multimedia Tools and Applications*.

Sankar, K., Jawahar, C., and Zisserman, A. (2009). Subtitle-free Movie to Script Alignment. In *Proc. of British Machine Vision Conference*.

Sivic, J., Everingham, M., and Zisserman, A. (2009). "Who are you?" – learning person specific classifiers from video. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Tapaswi, M., Bäuml, M., and Stiefelhagen, R. (2012). "Knock! Knock! Who is it?" Probabilistic Person Identification in TV Series. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

The Big Bang Theory Forums. (2013). The Big Bang Theory Forums. `forum.the-big-bang-theory.com/`. Accessed: 2013-10-20.

The Big Bang Theory Transcripts. (2013). The Big Bang Theory Transcripts. `bigbangtrans.wordpress.com`. Accessed: 2013-10-11.

The Big Bang Theory Wiki. (2013). The Big Bang Theory Wiki. `wiki.the-big-bang-theory.com/`. Accessed: 2013-10-11.

Tsiartas, A., Ghosh, P., Georgiou, P., and Narayanan, S. (2013). High-quality bilingual subtitle document alignments with application to spontaneous speech translation. *Computer Speech and Language*, 27(2):572 – 591.

Wu, Z. and Palmer, M. (1994). Verb Semantics and Lexical Selection. In *Proc. of the 32nd Annual Meeting of the Association for Computational Linguistics*, pages 133–139.

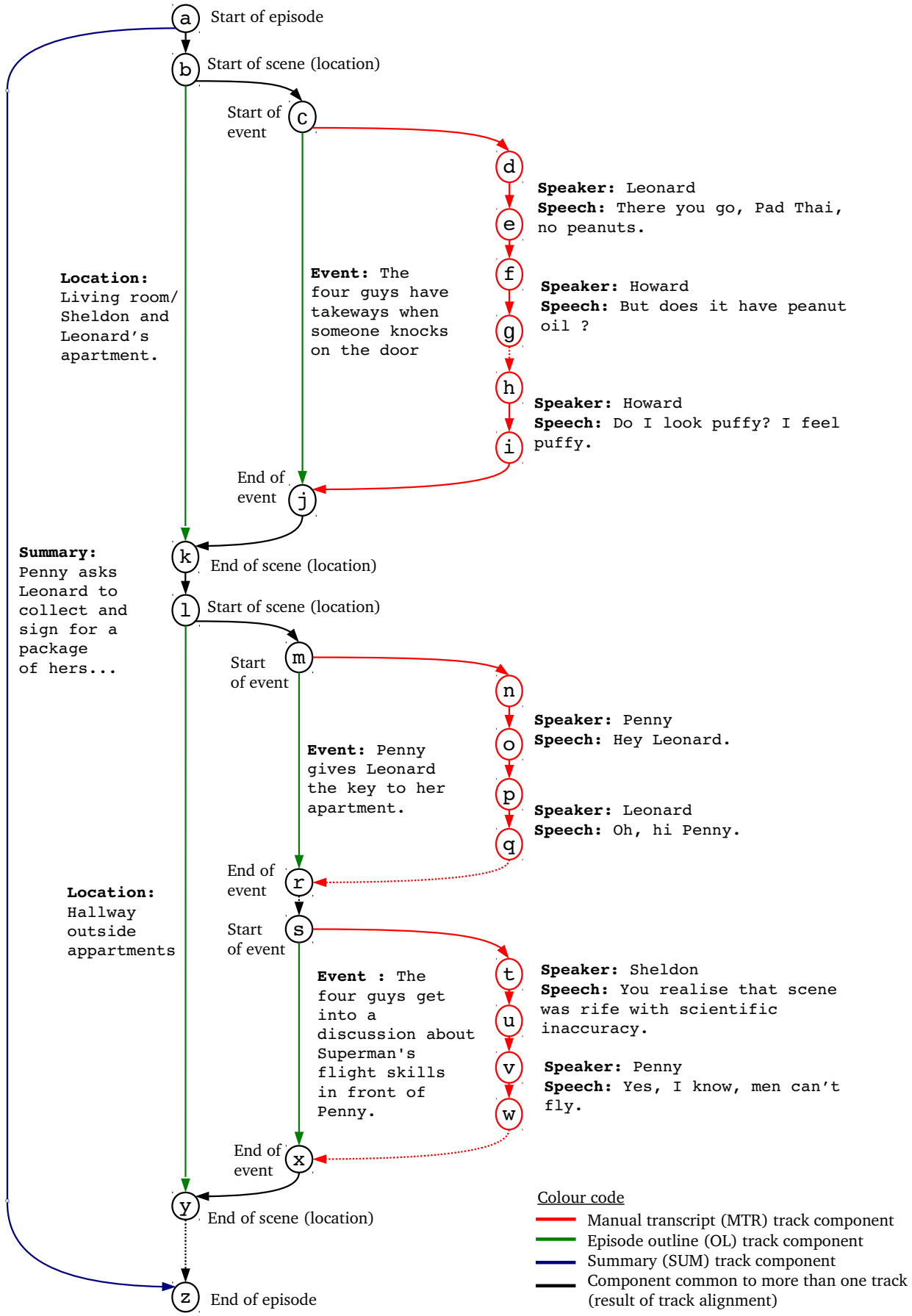

Figure 1: Linguistic annotation graph showing aligned (merged) manual transcript, episode outline and summary tracks from TVD database. The graph corresponds to Season 1 Episode 2 of *The Big Bang Theory* (ref. Sections 3. and 4.).