Matching Verb Attributes for Cross-Document Event Coreference

Eleftheria Tomadaki and Andrew Salway
Department of Computing, University of Surrey
Guildford, Surrey, UK
GU2 7XH
{e.tomadaki, a.salway}@surrey.ac.uk

Abstract

Collateral texts of different genre can describe the same filmed story, e.g. audio description and plot summaries. We deal with the challenge of cross-document coreference for events by matching verb attributes. Cross document coreference is the task of deciding whether two linguistic descriptions from different sources refer to the same event. This is important for reliable information integration, as well as generating richer machine-executable representations of multimedia material in retrieval and browsing systems. Corpora of audio description and plot summaries were analysed to investigate how they describe the same film events. This analysis shows that events are described by different verbs in the two corpora and has inspired the algorithms for cross-document event coreference, which match verb attributes, rather than verbs themselves. The preliminary evaluation was encouraging, showing a significantly better performance than the baseline algorithm.

1 Introduction

The present era can be characterised by a vast amount of information available in different forms of media; text documents, images, audio and video files etc. Many kinds of electronic information artefacts convey the same story; a fire event, for example, can be broadcast on television or radio, or narrated in a newspaper by the people that were affected; or a fictional story, for example Cinderella can be presented in films, theatre, books, pantomime etc. Information can be conveyed in the form of stories in history, science, current affairs, financial news, fiction etc. The process of narrating a story comprises a sequence of causally connected events organised in space and time. Matching events can be one way to acquire major information about a story.

This research is motivated by the fact that associating information in different texts representing the same story can on the one hand enhance the collection and verification of most available information about one story and more reliable information integration, and on the other hand provide richer machine-executable representations of multimedia material in retrieval and browsing systems, such as film databases.

Natural language textual descriptions can be collateral to a moving image and represent its content in words. Extracting information from collateral text (Srihari, 1995) can address higher levels of semantic video content than video processing alone, as language can express more information than colours, shapes, motion etc. and enhance video indexing, retrieval and browsing. Films entail stories and their content can be described by a range of collateral texts; a story told in a novel can be turned into a film. Novels can total 100,000's words and give detailed descriptions of charaters' cognitive states, which can be expressed by facial expressions in the moving images. Screenplays are the directors' scripts including dialogue, character and setting descriptions as well as instructions to the camera totaling 10,000's words. Audio descriptions are detailed descriptions of the characters' appearance and facial expressions, settings and what is happening on screen at the moment of speaking totaling 1,000's words. Audio description is scripted before it is recorded and includes time-codes to indicate when each utterance is to be spoken, enabling the alignment of the narration with the visual images. Plot summaries narrate the major events of the film in 100's words and include character's desires and goals. The challenge is to understand what is common in different collateral descriptions representing the same events. Consider for example, how the same event (*burned*) for the film English Patient is described in different collateral texts, Figure 1. Each source is heterogeneous, using different vocabulary, grammar structures, amount and kinds of information. These different collateral descriptions can be aligned to audio description fragments, which are temporally associated to the film data;

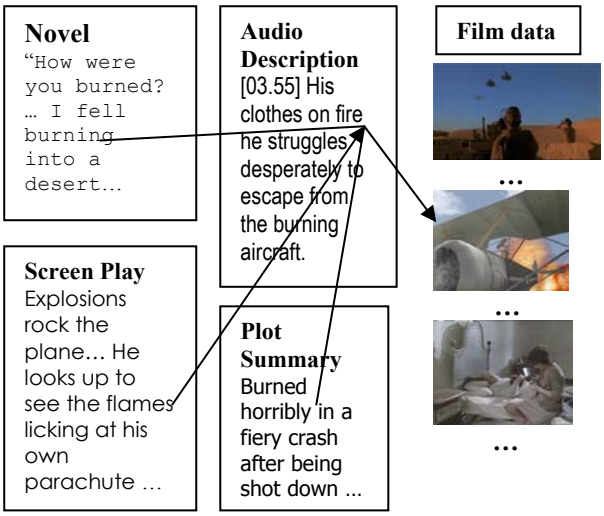

Figure 1: Different collateral descriptions for the same film event.

1.1 Towards Information Integration

A number of terms can describe the process by which information is extracted from different texts relating to the same theme and then associated and combined. The method followed in this work as a first step to integrate event-related information is called C*ross-Document Coreference*; this is the process of deciding whether two linguistic descriptions from different sources refer to the same entity or event and has been applied in specific sets of events, such as election and terrorist events (Bagga and Baldwin, 1999). Recent systems associate entities, extracting nouns and pronouns from different news texts and matching them (Radev and McKowen, 1998). Cross-document coreference appears to be a sub-task of cross-document summarisation by selecting and matching of the crucial information in multiple texts before summarising multiple documents. The task of selecting candidate phrases is expressed in the Document Understanding Conferences (DUC) and is based on the principle of relevance: syntactic patterns are significant, as they describe either a precise and well-defined entity or concise events or situations. Cross Document Structure Theory (CDST) describes several relations included in pairs of matched fragments tested on news articles (Zhang et al, 2003). CDST is tested on relations in homogeneous texts. Related research includes the term information merging describing the process of integrating information about a set of football events, e.g. goal, free kick etc.; the technique applied includes extraction and matching of a set of specific entities, such as football players' names etc from different texts, e.g. tickers, radio transcriptions etc. (Kuper et al, 2003).

Although the two kinds of texts presented in this paper, audio description and plot summaries, describe the same story, they are very different in the vocabulary used, the content and amount of event-related information included; cross-document event coreference in films is perhaps more challenging because it is harder to identify a set of common events.

The goal of the current work is to develop a computational account of how events are expressed in different narrative discourses of the same story in multimedia systems. We focus on the question of how information about an event can be related in different discourses. Our approach is inspired by the corpora analysis, which shows the challenge of matching events in heterogeneous texts, such as plot summary and audio description, as they include different verbs. However, several verb attributes, for instance nouns and proper nouns, are common in both kinds of texts. This analysis has led to the proposal of a method including algorithms that apply event cross-document coreference by matching combinations of verb attributes, rather than matching verbs themselves.

2 Collateral Texts for Films: audio description and plot summaries

Audio description (AD) narrates what is happening on screen for visually impaired people and is available for a range of television programmes, such as series, documentaries, films, children's programmes etc. It is produced by trained experts who follow guidelines while describing, for instance the use of present tense showing that the actions take place at the moment of speaking and the use of proper nouns when there are a lot of participants in a scene to avoid the confusion of the audience. The description is first prepared in electronic format, time-coded and then spoken. The audio description for films is a detailed, long description which involves a story, unfolded in a series of temporally and causally connected events, including characters and plot significant objects, location of the scene, who is speaking, what the characters are doing and wearing, facial expressions and body language, text shown on screen and colours. The following examples are from the audio description for the film English Patient from 3m 40s to 3m 55s:

[03:40] Bullets tear holes in the fuselage.
[03:47] The plane catches fire.
[03:55] His clothes on fire he struggles to escape

In contrast, plot summaries (PS) are short descriptions mentioning the major points of a filmed story, the protagonists and their intentions, locations, time and duration of main events and cause of certain actions. The film is described

according to the subjectivity of any author that decides to publish a film summary electronically, without following any guidelines. The following excerpt is from the plot summay for the film English Patient:

> Burned horribly in a fiery crash after being shot down while crossing the Sahara Desert ...

2.1 Corpora Analysis

Two corpora were created to represent and analyse the language used in audio descriptions and plot summaries. The corpora include nine different film categories selected by audio description experts based on the choice of vocabulary, grammar structures and kinds of information conveyed: children's live action and animation, action, comedy, period drama, thriller, dark, romantic and other. The present audio description corpus includes audio description scripts for 56 films, approximately 376,000 words (6,000-8,500 words per script). The current plot summaries corpus includes summaries for the same films (Internet Movie Database), totaling 9,500 words approximately (around 200-400 words per summary). The 100 most frequent words include 41 open class words in the audio description corpus, and 27 open class words in the plot summary corpus. This suggests that audio description and plot summaries are special languages, while comparing them with common language (2 open class words in the first 100 words of the BNC corpus) and other corpora of special languages (e.g. 39 open class words in the linguistics corpus). The most frequent words in both corpora are proper nouns and nouns referring to characters, plot significant objects and time, as well as verbs. However, only a few nouns and proper nouns are the same. In language, an event is typically realised in the form of a verb or noun. We analyse verbs having selected a verb classification based on the semantic properties of the verbs, used to structure and represent event-related information. In functional grammar, verbs can be categorised in six kinds of processes: material, mental, behavioral, existential, verbal and relational (Halliday, 1994).

According to the frequency results, around 70% of the verbs in both corpora represent material processes, Figures 2a and 2b. However, the verbs included in the material processes category differ in the two corpora. Audio description includes verbs describing motion such as *walk, come, open, fall* etc., which, if separated by the context, do not give explicit information about major events, whereas plot summaries include verbs such as *murder, escape, die, find, help, follow* etc. that refer to the story plot; for example, a murder event may be described in audio description as *he picks up the gun and points at the man…he pulls the trigger*. In plot summaries there are more verbs expressing mental processes (20%) than in audio description (7%). Interestingly, the quality of the mental processes is also different. Mental processes of seeing are mostly depicted in audio description, by verbs such as *watch* and *see*, whereas plot summaries include mental processes related to cognition or affection, what the characters believe and feel, i.e. verbs such as *love, want, know, plan, decide* etc. which are not encountered in audio description.

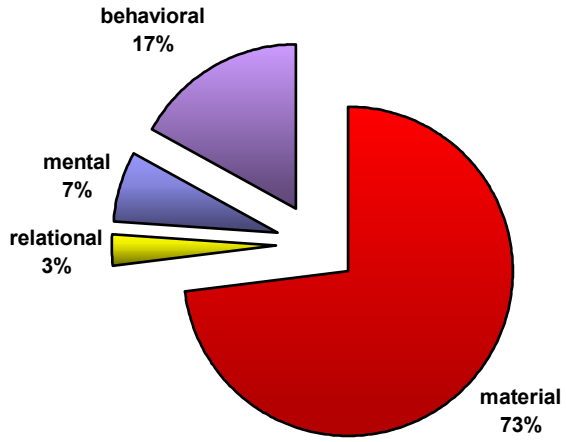

Figure 2a: 4 types of processes in a 376,000-word corpus of audio description based on the 30 most frequent verbs

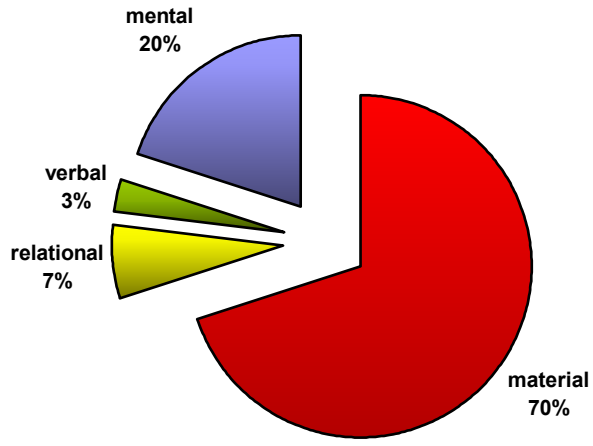

Figure 2b: 4 types of processes in a 9,500-word corpus of plot summaries based on the 30 most frequent verbs

The other verb categories encountered in audio description and plot summaries are different. In audio description, behavioral processes constitute

the 17% including verbs such as *smile, stare, look* and *glance*, as the narrators describe what can be seen on screen relatively to the characters physiological and psychological behaviour. These processes may be proved to be important as they can sometimes describe emotions, for example a *laughing* process can express a positive feeling related to the character and concerning the event that has just preceded in the story. On the contrary, the 30 most frequent verbs in plot summaries do not include the behavioral category, as the authors do not describe the character's behavior. Plot summaries also contain verbal processes (3%), such as *tell*, that are not mentioned in audio description due to the dialogue's presence that actually represents the verbal processes.

The frequency results suggest that the same events are described by different verbs in the two corpora. Material verbs may compose the biggest category in both corpora, but the verbs differ completely as shown in Tables 1 and 2.

Process	Verbs in audio description
Material	open, walk, run, step, hold, close, go, wear, fall, lift, stand, throw, carry, kiss, sit, lead, get, give, cross, join, make, jump
Relational	be
Mental	watch, see
Behavioral	smile, stare, look, glance, nod

Table 1: The 30 most frequent verbs describing 4 types of processes in audio description

Process	Verbs in plot summaries
Material	get, love, find, take, kill, help, go, become, plan, die, give, come, escape, make, murder, try, turn, change, follow, lose, need, run
Relational	be, have
Verbal	tell
Mental	want, know, decide, seem

Table 2: The 30 most frequent verbs describing 4 types of processes in plot summaries

In the following example, the *tending* event included in the plot summary is expressed by the verb *tend*, a series of moving images in the film and a series of audio description utterances including the verbal groups *make comfortable* and *wash*, Figure 3. These verbs cannot be matched as they are not synonyms to the verb *tend*.

Audio description
[23:54] Hana ***makes*** her patient ***comfortable***

Plot summary
a young, shell-shocked war nurse Hana remains behind to ***tend*** her doomed patient

Audio description
[45:09] Gently Hana ***washes*** the tender skin on the patient's chest.

Figure 3: Audio description utterances for the same plot summary event

The wordlists of the plot summary and the audio description for the film English Patient do not include any verbs mentioned in both texts. However, they share other open class words; interestingly, the most frequent ones are proper nouns and nouns expressing the characters of the story, locations etc, Table 3.

Common open class words	OCW PS	OCW AD	Cumulative OCW
Hana	1	73	74
Patient	1	33	34
Kip	1	31	32
Caravaggio	1	22	23
Desert	1	17	18
Nurse	1	6	7
Pilot	1	2	3
Burned	1	2	3

Table 3: Common open class words and their occurrence (OCW) in the PS and AD wordlists for the film English Patient

A major event described by one verb in the plot summary, such as *tend* in the example used, may not be explicitly expressed in the audio description, but implied through a series of other events and actions, e.g. *wash* and *make comfortable*. Common event attributes are only the participants *Hana* and *patient*. It is therefore possible to match their combination instead of matching the verb *tend*.

2.2 Creating Test Data

We focus on a method to identify and relate event related information in plot summaries and audio description. The human task involves reading plot summaries and watching the corresponding films, associating the events read to the events visualised. The annotators detect and number the events read in the plot summary. While watching the film, they are told the number of the scene each time a scene commences and they associate the number of the event visualised on screen to the number of the scene, e.g. in the film English Patient, the plot summary event 2 *burned*

horribly in a crash can be visualised in scene 2 of the film. The human task of matching the events can be characterised as cross-modal event coreference, as humans match events they have read to events they visualise on screen. This had caused disagreements on whether events not explicitely expressed by the visual images but inferred by the sound effects or the dialogue should be annotated or not. The annotation of all events, either explicit or inferred was taken into consideration for the preliminary evaluation of this work due to the multimedia nature of the data included.

2.3 Proposed Algorithms

To compute the human task of event association, we propose a method for cross-document event coreference by identifying and matching verb attributes. The task of event detection in plot summaries has not been automated and main events are already numbered by the human annotators that have read the plot summary. Having identified the main events in the plot summary, we have used the Connexor tagger to represent the plot summary sentences in terms of grammar and functional roles. The algorithms designed generate a list of combinations of event constituents, i.e. verbs and their attributes, according to the tags assigned and match them to the corresponding combinations in the audio description fragments, which are associated with the film data by time-codes and divided into scenes. The scene division was available as part of some scripts by the audio describers who authored the scripts, whereas we have separated the rest of the films according to the scene division in the visual data, i.e. when the location or time changes.

As shown from the verb frequency analysis, in 2.1, it is hard to match verbs from different collateral descriptions expressing the same event. However, characters, plot significant objects and usually locations can be matched. The suggested approach is to match the combination of all or most of the event ingredients, i.e. participants and their roles and circumstances. In the first algorithm, called *Keyword Combination List Generation and Matching (KC)*, the identified plot summary events are grammatically tagged by the Connexor part-of-speech tagger. We then apply rules combining the event constituents, Figure 4; the participants are usually expressed by nouns or proper nouns (as nominal heads), and the circumstances, e.g. location, time, expressed by nouns or adverbs etc. An obligation is to retrieve the combination of the event participants, or one participant and another keyword.

Find Proper Noun / Noun + other keyword:
a. Proper Noun / Noun + Proper Noun-s / Noun-s (+ Noun-s +/ Verb +/ Adverb +/ Adj.)
If no other Proper Noun / Noun is found then find
b. Proper Noun / Noun + Verb +/ Adverb +/ Adj.

Figure 4: A Keyword combination rule

In the sentence *A young, shell-shocked war nurse Hana remains behind to tend her doomed patient*, the algorithm looks for the following combinations: Hana + nurse / patient (+remains +/ tend +/ behind +/ young +/ shell-shocked +/ doomed), as *Hana* is a proper noun and *nurse* and *patient* nouns, and then for the verbs *remains* and *tend*, the adverb *behind* and the adjectives *young*, *shell-shocked* and *doomed*. The next step is to match the generated list of keywords to the audio description utterances including all possible combinations of these keywords without tagging the audio description.

The second algorithm, called *Keyword and Keyword Role Combination List Generation and Matching (KKRC)*, is based on the combination of the keywords and their functional roles in the sentence. Here we have used the machinese syntax function of the Connexor tagger, which assigns words with the roles of subject, agent, object etc. This time, the algorithm looks for the combination of the keywords in the specific roles assigned by the tagger, which means we have to tag the audio description script as well as the plot summary. An example of keyword role combination list rules is shown in Figure 5:

Find [keyword+subject/agent–role] + [other keyword+functional role]:
a.Find [keyword+ subject/agent-role] + [keyword + object-role]
If no [keyword +object-role] is found then
b.Find [keyword+subject/agent-role]+ [keyword + prepositional complement]…

Figure 5: A keyword-role combination rule

In our example, the algorithm generates and matches the combination of *patient* plus the role of object plus another participant, *Hana* plus the role of subject (plus the verb *tend*); Hana[subject] + patient[object] (+tend [verb] etc.).

3 Preliminary Evaluation

The preliminary evaluation of the algorithms has been realised for four films, based on the comparison with human annotations, in terms of precision and recall. We first compare the scenes' identification number of the *Computer-Retrieved Scenes (CRS)* with the scenes' identification number of the *Human Annotated Scenes (HAS)* to find the number of *Correct Computer-Retrieved*

Scenes (CCRS). To find the percentage of the algorithms' precision, we multiply CCRS by one hundred and then divide it to CRS: CCRS + 100/ CRS. To find the percentage of the algorithms' recall, we multiply CCRS by one hundred and divide it to HAS: CCRS * 100/HAS. We have assumed a linear relation between plot summary and film time for the baseline algorithm, which divides the number of the audio description scenes to the number of the plot summary sentences and allocates the first plot summary sentence to the first audio description scene etc. The baseline's low performance (Table 4) is mainly due to the fact that events are ordered differently in plot summaries and in audio description. Film content can be organised in shots and scenes, which relate to film time and the events that comprise the semantic video content, which relate to story time; audio description is temporally aligned with the video data in film time, whereas plot summary is not, relating only to the story time (Salway and Tomadaki, 2002).

Algorithm	Precision	Recall
Baseline	0.1875	0.0261
KC	0.5625	0.6806
KKRC	0.6497	0.4145

Table 4: The evaluation of the algorithms in terms of precision and recall

The evaluation of the KC algorithm presents a significantly better precision and recall than the baseline algorithm. Combining nouns and proper nouns can be useful to find characters although they may not always be plot significant, in which case the precision is low. The KKRC algorithm is more precise, as more retrieved scenes were accurate. Less scenes were retrieved, as assigning roles to characters can be strict sometimes.

4 Discussion

The corpora analysis suggests the heterogeneity of the audio description and plot summaries corpora and the challenge of relating pairs that describe the same events using different verbs, structures and amount of event-related information. This investigation guided the algorithms' approach to match verb attributes; characters and roles, objects, locations or other circumstances. This can show different relations in cross-document structures. The preliminary evaluation shows that precision is of more importance in our case and that semantic role matching is more precise than matching grammatical attributes. To increase the precision, an event classification for filmed stories may be proved useful; for example, the verbs *kill*, *love*, *escape*, *help*, *murder*, *plan* etc. are amongst the 30 most frequent verbs in the plot summary corpus. A preliminary evaluation of using systems such as CYC and WordNet to match events by query expansion has shown that the difference in the vocabulary choice used in the two corpora is not based on synonyms. Matching verb attributes in audio description and plot summaries may also automate the task of event decomposition into other events; for example a *tending* event may include *making comfortable*, *washing* etc. or a *fighting* event may include *kicking, punching, firing at* etc. The algorithms should also be tested on other kinds of data, such as news stories or witness accounts.

5 Acknowledgements

Our thanks to the EPSRC funded project TIWO (GR/R671940/1), the RNIB and ITFC for providing the audio description for this research

References

A. Bagga, A. and B. Baldwin. 1999. *Cross-Document Event Coreference: Annotations, Experiments, and Observations*. Workshop on Coreference and its Applications (ACL99)

A.J. Salway and E. Tomadaki, 2002. *Temporal Information in Collateral Texts for Indexing Video* Procs. LREC Workshop on Annotation Standards for Temporal Information in Natural Language

Connexor: http://www.connexor.com/demo/tagger/

D.R. Radev, and K.R. McKowen. 1998. Generating Natural Language Summaries from Multiple On-Line Sources. *Journal of Computational Linguistics* 24(3): 469-500

Internet Movie Database: http://www.imdb.com

J. Kuper, H. Saggion, H. Cunningham, T. Declerck, F. de Jong, D. Reidsma, Y. Wilks, and P. Wittenburg. 2003. *Intelligent multimedia indexing and retrieval through multi-source information extraction and merging*. 18th International Joint Conference of Artificial Intelligence, Acapulco

M.A.K Halliday. 1994. *Introduction to Functional Grammar*. 2nd Edition, London

R.K. Srihari. 1995. Computational Models for Integrating Linguistic and Visual Information: A Survey. *Artificial Intelligence Review* 8 (5-6), pp. 349-369

Z. Zhang, J. Otterbacher and D. Radev, 2003. *Learning Crossdocument Structural Relationships using Boosting* 12th International Conference on Knowledge Management