Cross-Document Coreference for Cross-Media Film Indexing

Eleftheria Tomadaki and Andrew Salway

Department of Computing, University of Surrey
Guildford, GU2 7XH, United Kingdom
[e.tomadaki, a.salway]@surrey.ac.uk

Abstract

Potentially, rich representations of film content could be extracted and merged from various texts, such as screenplays, audio description and plot summaries, in order to improve video indexing. As a first step, this requires solving the cross-document coreference (CDCR) task. The CDCR task is difficult in this new scenario because the texts each select and present information about film events very differently; furthermore, the set of possible events is relatively unconstrained. In order to propose new solutions for CDCR we first analysed how two different text types select and present information about the same film events. We present a corpus based analysis of the language used in plot summaries and in audio description, which suggests that while both use similar words to refer to entities, they use very different words to refer to events; there is little systematic relation between the words each use to refer to events. Based on our results, we propose and evaluate four heuristics for the CDCR task that match nouns, functional roles, some verbs, and take into account the number of expected matches according to event aspect. At best we achieved Precision of 49% and Recall of 32% based on 375 CDCR instances between plot summaries and audio descriptions. These figures are low compared to many information retrieval and extraction tasks but we believe that: (i) they may be close to the best possible given the differences between the text types and that they refer to an unconstrained set of events; (ii) they are high enough to start leveraging the information in the texts for video indexing purposes.

1. Introduction

Amongst the mass of multimedia information available today are many artifacts that in some way tell the same story, be it news or fiction. This fact can be exploited by multimedia indexing systems, for example, to generate rich descriptions of semantic video content from texts related to the video. An interesting and challenging scenario is film and the wide range of texts describing its content, such as novels, screenplays, audio description, plot summaries and reviews. These texts complement each other in providing different kinds of information about the story told by a film. It would be difficult, if not impossible, to extract much of this information directly from video data alone. However, before information extracted from each text type can be combined into a rich representation of semantic video content, it is necessary to solve the *Cross-Document Coreference* (CDCR) task. This is the task of deciding whether two linguistic descriptions refer to the same entity or event (Bagga, 1999).

Stories can be defined as comprising a sequence of causally connected events, where the causal connections typically relate to the goals and beliefs of characters (Bordwell and Thomson 1997). This paper focuses on CDCR between two interestingly different types of text that tell the stories of films – plot summaries and audio description. Plot summaries are written by film viewers to summarise the major events of the story. Audio description is mostly produced by trained experts to narrate what is happening on screen for visually impaired and blind film viewers. For more about audio description and its use for indexing narrative aspects of film video data, see (Salway and Graham 2003). Each text, along with the film itself, can be considered as a telling of the same story (Chatman 1978): each text selects and presents information about the events of a story differently. These differences mean that taken together the texts could provide richer indices for film video data, however the differences make the task of merging information from the two text types difficult.

2. Cross-Document Coreference for Video Indexing

Video can be indexed by applying a variety of natural language processing techniques on a range of texts relating to its content. Systems such as WEBSeek (Smith and Chang, 1997), PopEye (Netter, 1998), Informedia (Hauptmann, 2005), REVEAL THIS (Piperidis and Papageorgiou, 2005) and Google Video (2005) between them process HTML tags, closed captions and speech transcriptions. Other systems extract information from multiple texts related to restricted sets of events in a specialist domain. The KAB system processes dance descriptions and interpretations produced by experts (Salway, 1999). In order to index television coverage of soccer matches, the MUMIS system (Kuper et al, 2003) merges information about 31 football events, such as *goal*, *free kick* and *substitution,* from football transcriptions, tickers and news reports. The automatic identification of cross-document coreference is a crucial step in this merging. To date, CDCR algorithms have typically been used as part of cross-document summarisation systems, matching information in news articles. N-gram algorithms match two or more consecutive words, scored according to statistical methods, such as the document term frequency (TF) and the inverse document frequency (IDF), and are used in systems such as News Story Gisting (Doran et al, 2004). The ‘event centric’ algorithm is part of the MSR – NLP system (Vanderwende et al, 2004) and suggests a verb centred method, by matching the verb plus another word in a functional role – subject or object. The Boosting algorithm (Zhang et al, 2003) matches open class words and their synonyms, according to WordNet.

3. Corpus-based Analyses of Audio Description and Plot Summaries

Our initial observations about why plot summaries and audio descriptions are written suggest that they will select and present information about film events very differently. The corpus analysis reported here investigated: (i) the frequent open-class words in corpora of plot summaries and audio description (Section 3.1); (ii) whether there are regularities between the words used to refer to events in plot summaries and audio description (Section 3.2). We gathered a corpus of audio description scripts for 45 films, totalling 356,394 words spread across 9 categories and a corpus of 111 plot summaries spread across the same categories, totalling 13,761 words. For more information about the corpora and how they were gathered, see (Tomadaki 2006).

3.1. Frequent Open-Class Words

An analysis of the 100 most frequent words in the corpora shows that both contain an unusually high number of frequent open-class words (nouns, verbs, adverbs and adjectives). The audio description corpus includes 41 open class words in the top 100 and the plot summary corpus includes 27; note, only 6 appear in the in the top 100 words of the general language, British National Corpus, see the list in (Kilgarrif 2002).

Kinds of words	**Audio description frequent OCW**	**Plot summary frequent OCW**
Words that refer to characters, human body	man, head, eyes, hand, face, hands, Tom, men, John, woman	man, family, wife, men, father, son, woman, Harry, Tom
Words that refer to events or states	looks, turns, takes, walks, sits, stands, open, pulls, smiles, stares, goes, look, puts, steps, watches, opens, runs, stops, go	get, help, love, finds, discovers, named
Words that refer to objects and space	room, car, side, table, window, bed, door, way	-
Colours	white, black, red	-
Words that refer to time	-	now, time, new, years, day, young
Miscellaneous	water	life, world, way, war, story, earth

Table 1: Open class words (OCW) in the top 100 words of the corpora of audio description and plot summaries (the lists were not lemmatised). Underlined words are common to the top 100 of both corpora.

The corpora are similar in the words referring to characters words - *man, woman, men,* Table 1. The rest of the frequent open class words are different, implying significant differences between the kinds of information given by both texts. In audio description frequent words refer to the human body (*head, eyes, hand, face*), objects (*door, window, table*) and colours. In plot summaries, frequent words refer to time (*day, now*). Most interesting to us, the words referring to events are completely different. This contrast is apparent in a classification of the most frequent verbs in each corpus, according to the types of process (Halliday 1994) that they refer to, (Tomadaki and Salway, 2005), Table 2.

Borrowing terms from a set of generic actions for human movement (Gavrila 1999), audio description includes words referring to *stand-alone actions* (look, stare, watch, turn, walk, go, step) and *interactions with objects*, (open, put). Plot summaries describe processes involving complex sequences of human movement and implying something about characters' goals and beliefs (helps, love, discovers, escape, murder). This contrast can be explained by the functions of each text type: audio description is intended to communicate what is being depicted on-screen whereas a plot summary must condense the essential elements of the story. It is further indicated by the differences in mental processes.

Process	**Verbs in audio description**	**Verbs in plot summaries**
Material	turn, take, walk, sit, stand, open, go, put, step, hold, close, wear, carry, run, fall, lift, throw, kiss, lead, get, give, cross	get, help, find, go, take, meet, become, kill, make, destroy, play, save, come, escape, move, lose, try, murder, die, leave
Relational	be	be, have
Mental	watch, see	love, discover, want, know, see, decide, seem
Verbal	-	tell
Behavioural	look, smile, stare, glance, nod	-

Table 2: The 30 most frequent verbs in the audio description and plot summaries in lemmatised corpora wordlists, categorised according to functional grammar processes (Halliday, 1994)

3.2. Cross-Document Coreference

The contrasts between the verbs used to refer to events in plot summaries and audio description seem to present challenges for the CDCR task; it is not possible to match two linguistic descriptions according to whether they contain the same verb. However, it might be reasonable to think that even if two descriptions of the same event do not use the same verb, they will at least use verbs that share some lexical relation, e.g. synonymy, troponymy, entailment (Fellbaum 1998). Thus we created a data set comprising 355 instances of CDCR between plot summaries and audio description. In summary, our analysis showed, perhaps surprisingly, that there is very little systematic relation between the words used to refer to events in plot summaries and audio description. However, we were able to note some degree of regularity in the number of audio description utterances co-referring with a plot summary clause according to the aspect of the verb included in the plot summary.

The 10 most frequent events were identified in the plot summary corpus after lemmatising the corpus wordlist: *help* (12 instances)*, meet* (11), *kill* (10)*, bring* (8)*, tell* (8)*, force* (8)*, find* (7)*, discover* (7)*, love* (6) and *murder* (6). Each plot summary instance, including a clause with one of these words (appearing mostly as verbs and on a very few occasions as nouns), was correlated (by one of the authors) to one or more audio description utterances, which could be consecutive or scattered in different parts of the script. Table 3 shows examples of CDCR relating to plot summary instances of *murder.*

Film	**Plot summary clause**	**Audio description utterance/s**
One Hot summer night	When the businessman is murdered the police naturally eye the woman as the top suspect.	[00.02.44] The driver pulls a gun. [00.02.57] The van driver shoots the middle aged man and he slumps back…
See No Evil, Hear No Evil	A man is murdered.	[00.10.19] The woman pulls a gun. She shoots him
Midnight in the Garden of Good and Evil	*The mysteries surrounding Billy's murder...*	[00.10.33] Billy is lying face down, blood on his back. ... [53.05.58] Jim shoots Billy in the back

Table 3: Example CDCR pairs for *murder*

We analysed the audio description utterances associated with each of the 10 frequent plot summary events to identify words that occurred unusually frequently in the co-referring audio description. Examples we found included, for the plot summary event *kill – body, gun, falls, shoots* and *police*, and for the plot summary event *love – kiss, kisses, gaze, gently, lips*. However, such examples were rare, and could only be detected for some of the very most commonly occurring plot summary events. Unless the corpora were much larger, or the CDCR task was reduced to a small set of common events, then it seems unlikely that a CDCR algorithm can rely on matching verbs either directly or via any kind of lexical relation; there is a large tail of infrequent events in the plot summary corpus – for more details see (Tomadaki 2006). Only in 3.2% of the CDCR instances did the audio description refer to the event expressed in the plot summary with the same word or with a synonym. Instead, sometimes, the audio description describes one or a series of related actions where there is, in common with the plot summary clause, at least one entity in the same functional role or a combination of entities.

One regularity we observed that might be helpful for CDCR solutions is a correspondence between the aspect of a verb – punctual or durative (Comrie 1976) – in the plot summary utterance, and the number of matching plot summary utterances. The events *discover*, *meet*, *bring*, *murder*, *kill*, *find* and *tell* are considered to be punctual and they co-refer with a mean average of 5 audio description utterances. By contrast, the durative *help*, *love* and *force* co-refer with a mean average of 29 audio description utterances. Punctual events usually occur in one part of the film and consequently in one part of the audio description, e.g. a murder usually happens quite quickly in one scene and is thus expressed in a few audio description utterances. 'Durative' events are expressed in multiple parts of the film and dispersed in the audio description, e.g. an event where the characters fall in love can be shown in different scenes throughout a film.

4. Proposed CDCR Solutions

We argue that the characteristics of CDCR in this scenario present new challenges for the task of automatically identifying instances of CDCR, and so previous approaches need to be adapted accordingly. In particular, it seems that we will have to rely on matching words referring to entities and their functional roles, because there is such little correspondence between the words used to refer to film events in audio description and plot summaries. Here we propose and evaluate four heuristics for the CDCR task, geared towards the film scenario.

4.1. Creating a Gold Standard Dataset

To create a gold standard dataset, five volunteers identified 375 CDCR instances between the plot summary and the audio description for the films 'Spiderman' and 'Chocolat'. The resulting data set will be made available on the web, see (Tomadaki 2006).

First they were asked to identify the events expressed by the plot summaries and then identify them in the audio descriptions. The identification of events in the plot summary was straightforward, as the pairwise agreement between the annotators was 90%; some volunteers annotated as events a few sentences including more than one verbs conveying different actions with different participants, while others annotated as events only clauses including one verb, which makes the task more focused. The annotators consolidated their answers, annotating plot summary clauses including one verb. The task is time-consuming and challenging when it comes to the event identification in the audio description, totalling four to six hours. The pairwise agreement between all annotators was quite low, totalling 62% in both films. The answers were quite different in events referred to in multiple utterances or because some utterances referred to more than one event, not all being annotated. All annotators noted that after the first couple of hours the task of annotating the audio description became laborious as they did not allow themselves to have multiple breaks. The annotators have finally concluded that the task was subjective due to the different inferences made by each person. They were then asked to reassess their annotations, considering annotations detected by the others and deciding whether to include them in their answers. After the data reconciliation the pairwise agreement increased to 95% for both films. For more details see (Tomadaki 2006).

4.2. Heuristics for Cross-Document Coreference

We propose four heuristics for the cross-document coreference task, and evaluate each on the plot summary/audio description data set. For each heuristic we first identify the events in the plot summary, following an algorithm which adds grammatical and functional roles (subject, object), deletes sentences having the verbs *be* and *have* as main verbs (as they normally denote states), resolves pronouns and finally separates clauses giving them a unique identification number. The words are matched in their base form. The parsing was realised using the Connexor tagger (www.connexor.com) and the pronoun resolution using ANNIE in GATE (www.gate.ac.uk). We will show how each heuristic is applied for the first clause referring to an event in the plot summary for the film 'Spiderman': *A rather odd thing happened to the life of nerdy high-school student Peter Parker: after being bitten by a radioactive spider…* Note that the Precision and Recall statistics relate to 375 instances of CDCR between 21 different plot summary events in the two films and their audio descriptions.

The first heuristic concentrates on matching entities, which tend to be characters or occasionally objects and locations:

Heuristic 1: If at least two head nouns in the plot summary clause appear in the audio description utterance (in any form), then MATCH = TRUE, else MATCH = FALSE

In the first plot summary event, the words to be matched according to heuristic 1 are: *Peter Parker / Peter/ Parker + thing +/ life +/ student +/ spider.* Eleven audio description utterances, including at least two head nouns or proper nouns were matched. Only three out of ten matches of the combination *Peter* and *spider* were correct, e.g. *[09.56.00] the spider inches its way down towards Peter*, as the word *spider* referred to other spiders as well as to the radioactive spider which bit *Peter*, e.g. *[03.02.57] Peter, with a spider and web emblazoned on his sweatshirt…*A problem arises when two nouns referring to the same entity are matched in the two texts, e.g. *Peter* and *student*, reducing Precision. Overall, Heuristic 1 achieved Recall of 23.4% and Precision of 30.4%, identifying 83/375 CDCR instances.

Heuristic 2 adds verbs and all the nouns to the keyword list to be matched:

Heuristic 2: If at least two nouns or one noun and one verb in the plot summary clause appear in the audio description utterance, then MATCH = TRUE, else MATCH = FALSE

In our example, the words to be matched are: *thing – occur - life - school - student - Peter Parker / Peter/ Parker – bite – spider*. Heuristic 2 retrieved fifteen utterances in total, including eleven spurious, while the gold standard includes five. Four utterances were matched including keywords such as *Peter, spider* and *bite* and two of them were correct, whereas the rest refer to another event, e.g. *[24.07.10] He zooms in on the Daily Bugle front page: Big Apple dreads Spider bite.*

Both Recall (26.2%) and Precision (32.4%) improve slightly on Heuristic 1.

Heuristic 3 has stricter matching criteria in order to increase Precision, by requiring that a noun appears in the same functional role in both plot summary clause and audio description.

Heuristic 3: If at least one noun in the plot summary clause appears in the audio description utterance in the same functional role AND at least one other noun or a verb in the plot summary clause appears in the audio description utterance then MATCH = TRUE, else = FALSE.

We match words with other words in the same functional role, logical subject/agent or object, following the terms in the Connexor tagger: *[Thing: subj] +/ occur +/ Peter Parker / Peter/ Parker + [spider: agt] +/ bite*. Heuristic 3 detected three matches in event 1 of the film

'Spiderman'. Two out of five utterances have been retrieved, including *spider* in the role of subject and *Peter,* or *Peter* in the role of object and *spider,* e.g. *[10.05.00] the spider bites Peter*. All matches are correct, whereas another two utterances have not been detected, as they either include the words *Peter* and *spider* in different functional roles, or because the pronoun resolution failed in the previous step. Heuristic 3 achieved the lowest Recall (21.5%) but the highest Precision (49.4%).

Heuristic 4 was designed to balance Precision and Recall by combining heuristic 2, and heuristic 3. It checks the event aspect (punctual or durative), according to an index of all plot summary events that we have created, and if punctual it retrieves the 5 highest ranked matches according to the following match score algorithm, if durative it retrieves all candidate matches. Heuristic 4 achieved Recall of 32.9% and Precision of 47.8%. One limiting factor is that not all references to the same entity can be resolved automatically, e.g. *Peter Parker* and *Spiderman,* and *chocolate shop*, *patisserie* and *chocolaterie*. When Heuristic 4 was evaluated after manual entity resolution Recall increased to 46.2% and Precision to 50.1%.

1st: Match according to heuristic 3 with two keywords in the same roles

2nd: Match according to heuristic 1 and 2 with three keywords

3rd: Match according to heuristic 3 with one keyword in the same role and another keyword

4th: Match according to heuristic 1, 2 and 3 with two keywords in utterances appearing within or close to the estimated temporal interval

5th: Match according to heuristic 1 and 2 with two keywords

6th: Match utterances appearing within or close to the estimated temporal interval including one keyword

7th: the rest

5. Conclusions

Though these figures are low compared to many information retrieval and extraction tasks we believe that: (i) they may be close to the best possible given the differences between the text types and that they refer to an uconstrained set of events; (ii) they are high enough to start leveraging the information in the texts for video indexing purposes. Until now, researchers have focussed on CDCR between texts of the same type, or texts referring to a restricted set of events. We showed that with perfect entity resolution we could get both Precision and Recall to around 50%. Further improvements, for a small set of common plot summary events, should be possible by identifying correspondences between individual plot summary words (love, murder, etc) and words commonly used to describe the events in audio description.

6. Acknowledgments

This research was carried out as part of the Television in Words project (TIWO) supported by an Engineering and Physical Sciences Research Council (EPSRC) grant, GR/R67194/01: http://www.computing.surrey.ac.uk/personal/pg/A.Salway/tiwo/TIWO.htm.

The authors thank the members of the TIWO Round Table for sharing their knowledge of audio description and providing samples for our corpus.

7. References

Bagga A. and Baldwin B. (1999). Cross-Document Event Coreference: Annotations, Experiments, and Observations. In: *Proceeding of the ACL99 Workshop on Coreference and its Applications* (pp. 1-8).

Bordwell D. and Thompson M.K. (1997). *Film Art: An Introduction*. New York: McGraw-Hill.

Chatman S. (1978). *Story and Discourse: Narrative Structure in Fiction and Film*. NY: Cornell University Press.

Comrie B. (1976). *Aspect: An Intoduction to the Study of Verbal Aspect and Related Problems*, Cambridge: Cambridge University Press.

Doran W., Stokes N., Newman E., Dunnion J., Carthy J., and Toolan F. (2004). News Story Gisting at University College Dublin. In

Proceedings of Document Understanding Conference 2004, Boston, USA.

Fellbaum C. (1998). *An Electronic Lexical Database*. Cambridge: The MIT Press.

Gavrila D. (1999). The Visual Analysis of Human Movement. *Computer Vision and Image Understanding* 73 (1), (pp. 82-98).

Google (2005). *About Google Video,* http://video.google.com/video_about.html

Halliday M.A.K. (1994). *Introduction to Functional Grammar.* 2nd Edition, London: Edward Arnold.

Hauptmann A. (2005). Lessons for the Future from a Decade of Informedia Video Analysis Research. International Conference on Image and Video Retrieval. *Lecture Notes in Computer Science*, Volume 3568, August 2005, pp. 1-10.

Kilgariff A. (2002). *BNC wordlist.* ftp://ftp.itri.bton.ac.uk/bnc/all.num.o5

Kuper J., Saggion H., Cunningham H., Declerck T., de Jong F., Reidsma D., Wilks C. and Wittenburg P. (2003). Intelligent Multimedia Indexing and Retrieval through Multi-source Information Extraction and Merging. *Proceedings of International Joint Conference of Artificial Intelligence-2003 Workshop on Information Integration on the Web (*IJCAI *03),* pp. 409-414.

Netter K. (1998). 'POP-EYE and OLIVE - Human Language as the Medium for Cross-lingual Multimedia Information Retrieval.' The ELRA Newsletter (European Language Resources Association) November 1998, pp. 5-6.

Pickering M.J. and Rüger S.M. (2003). ANSES: Summarisation of news video. *Lecture Notes in Computer Science* 2728 Springer-Verlag, pp. 425-434.

Piperidis S. and Papageorgiou X. (2005). REVEAL-THIS: Retrieval of Multimedia and Multilingual Content for the Home User in an Information Society. In *Proceedings of the 2nd European Workshop on the Integration of Knowledge,Semantic and Digital Media Technologies*, London, U.K.

Salway A.J. (1998). Video Annotation: The Role of Specialist Text. PhD thesis, University of Surrey.

Salway A.J. and Graham M. (2003). Extracting Information about Emotions in Films. In *Procs. 11th ACM Conference on Multimedia 2003*, 4th-6th Nov. 2003, pp. 299-302.

Smith J.R. and Chang S.F. (1997). Visually Searching the Web for Content. *IEEE Multimedia* July-September, pp. 12-20.

Tomadaki E. (2006). Cross-Document Coreference between Different Types of Collateral Texts for Films. PhD thesis, University of Surrey.

Tomadaki E. and Salway A. (2005). Matching Verb Attributes for Cross-Document Event Coreference. In Erk, Melinger and Schulte im Walde (eds.) *Proceeidings of Interdisciplinary Workshop on the Identification and Representation of Verb Features and Verb Classes*, pp. 127-132.

Vanderwende L., Banko M., and Menezes A. (2004). Event-Centric Summary Generation. In *Proceedings of Document Understanding Conference 2004,* Boston, USA.

Zhang Z., Otterbacher J., and Radev D.R. (2004). Learning Crossdocument Structural Relationships using Boosting. In *Proceedings of Document Understanding Conference 2004,* Boston, USA.