SceneML: A Proposal for Annotating Scenes in Narrative Text

Rob Gaizauskas
University of Sheffield
r.gaizauskas@sheffield.ac.uk

Tarfah Alrashid
University of Sheffield
ttalrashid1@sheffield.ac.uk

Abstract

We propose a scheme to annotate *scenes* in narrative texts. Adopting the widely held definition that a scene is a structural unit in narrative discourse where time, location and characters do not change, we propose: (1) a conceptual model for semantic annotation of scenes and relevant related semantic and textual elements and relations in narratives and (2) a concrete XML-based syntax for scene annotation. We illustrate our ideas via the annotation of an extended example and discuss various issues that have been and remain to be addressed.

1 Introduction

In this paper we put forward an initial proposal for a scheme to annotate *scenes* in narrative texts. Once the scheme is stable we aim to annotate sufficient text to enable a high accuracy classification model to be trained to carry out scene segmentation automatically.

There are several good reasons for doing this. First, there is considerable interest within the field of narratology in studying the structure of stories. A program that could automatically identify scene boundaries in stories would be an invaluable tool, allowing large-scale analysis and quantification of the number, type and length of scenes in different authors' writing.

Secondly, there has been a considerable amount of research integrating vision and language for different tasks, including:

- automatic text illustration or story-picturing;
- aligning books with movies; and,
- automatic generation of image descriptions.

Text illustration, or *story-picturing*, is the task of associating a relevant image with a given piece of text. This problem has been addressed by a number of researchers (Joshi et al., 2004; Feng and Lapata, 2010; Agrawal et al., 2011), but their focus has been on finding suitable images for a given piece of text. However, in illustrating a whole story, the story first needs to be segmented into scenes, as these are the appropriate units to have an associated illustration. Thus a system that could automatically segment a story into scenes, rather then relying on manual segmentation or purely structural information (paragraph or chapter boundaries), would be highly beneficial in such an application setting.

Aligning books with movies is the task of matching segments of a book with shots or sequences of shots in movie release of the book. Zhu et al. (2015) describe a system for tackling this task, which they conceive as a sentence to shot or or sentence sequence to shot sequence task. While this is one way to tackle the broader problem, there may also be merit in aligning scenes in books with scenes in movies, rather than starting at the lower level of sentence to shot alignment. Doing this however, presupposes a notion of scene and the ability to automatically segment both books and movies into scenes.

Image description is the task generating appropriate linguistic descriptions of image content. Considerable work has been done in this area (Kulkarni et al., 2011; Yang et al., 2011; Dai et al., 2017). While object detection capabilities have improved immensely in the past 5 years, image description involves more than simply listing object types in the image (Wang and Gaizauskas, 2016): choosing which objects to mention is important and this is at least in part dependent on the scene type. Models of scene types and of what object types are frequently mentioned in what scene type descriptions are likely to contribute significantly to improving image description. A likely source of such models are narrative accounts which contain descriptions of scenes, including the objects and actions found within them. However, mining them will require the capability to automatically segment narratives into scenes.

This paper is structured as follows. Section 2 gives a brief overview of some definitions of a scene from different areas of research and discusses work related to scene segmentation. Section 3 presents our framework for the annotation of scenes in narrative text Section 4 provides some examples of our annotation scheme in practice and Section 5 discusses some of the choices we made in specifying the annotation scheme as well as some outstanding issues and challenges. Finally, section 6 briefly conclude the results and provide future work suggestions.

2 Related Work

2.1 Scenes

The notion of what makes a scene varies, depending on the context in which it has been mentioned and the area in which it has been used. Here, we present an overview of scene definitions, followed by our own. Callaway and Lester (2002) stated that a segment of a narrative text is considered a scene if it is contiguous in time, character and place. Changes in any of those three elements (i.e. when characters change their location or time; when the scene switches to other characters) signal a scene change. (Kozima and Furugori, 1994), in their paper Segmenting Narrative Text into Coherent Scenes, defined a scene as a piece of text that has common characteristics with movie scenes in that they both have characters and objects in a certain situation that includes a specific time, place and background. However, a scene in a text does not necessarily begin with cue phrases. From the perspective of drama, (Polking, 1990) defined a scene as “a division within an act of a play, indicated by a change of locale, an abrupt shift in time or the entrance or exit of a major character”.

In addition, (Cutting, 2014) stated that a scene is a chunk of narrative that is sometimes synonymous with the concept of an event and often found in narrative arts. Dunne (2017) stated that the physical life of a setting presents the truth about the characters in the story or in the scene. The setting is defined by the physical life, the time and the physical objects of the setting; it is about knowing where and when the scene is taking place. For example, actions that occur at night differ from those that occur during the day, and outdoor settings are different from indoor settings. Physical life helps the reader imagine the image; therefore, it brings visual power to the scene.

By contrast, (Xiao et al., 2010) defined a scene as any place or location in which a person can take action or navigate. Scenes are often either related to or associated with specific actions (e.g. sleeping in a bedroom or reading in a library), and they are related to the space’s visual features. The environment is defined by its size and shape (e.g. a narrow corridor is for walking), by its material (e.g. snow, grass, wood) and by its objects (e.g. table and chairs). The scenes in their project are categorised into three levels: indoor, outdoor, and outdoor man made.

Our definition is different than those of others in that it only focuses on one feature of the scene (the place, or the setting, as it is called in some literature), and it considers time. More details of our definition are given in section 3.

2.2 Scene Segmentation

To our knowledge, the task of scene segmentation in text stories has not been investigated enough by researchers. Kozima and Furugori introduced a method of segmenting text into coherent scenes (Kozima and Furugori, 1994), suggesting that it could be used to resolve anaphora and ellipsis inside a scene. They defined a scene as a set of continuous sentences that have coherence between them. A scene in a text is like a movie scene in that it describes objects, such as characters or properties, in a certain time, place and background. Scene segmentation is done using a Lexical Cohesion Profile (LCP), which was first proposed by Kozima and Hideki (Kozima, 1993), to indicate boundaries of scenes in narrative text. It relies on the idea that coherent text is lexically cohesive (Morris and Hirst, 1991; Halliday and Hasan, 1976) so that the local cohesiveness leads to local coherence. An LCP keeps a record of lexical cohesion between words inside a window, by moving a window of a certain size (i.e. 50 words) word by word and measuring the lexical cohesion between the words each time a word is moved. There is a relationship between LCP and the alternation of a scene; when a window is inside a scene, the LCP value is typically high and words tend to be lexically cohesive; however, when a window crosses a scene boundary, the LCP value decreases and the words vary lexically. Thus, LCP identifies the alternation of a scene by detecting the valleys that are the minimum points.

In order to compute LCP, the lexical cohesiveness between words is computed first by spreading activation across the semantic network paradigm, which is an interval of [0,1]. Paradigm is a semantic network constructed systematically from a subset of an English dictionary called Glossem, that is used to analyse word meaning (Kozima and Furugori, 1993). The paradigm consists of 2,851 nodes and 295,914 links between them, nodes are constructed from the Glossem dictionary inputs. “Each node consists of a head-word, a word-class, an activity-value, and two sets of links: a referent and a refere”. Thus, LCP is computed by taking the record of cohesiveness $c(S_i)$ of a local text S_i at every position in the text. The best window size to use for LCP is 51 words.

The authors stated that LCP can be an indicator of scenes changes when comparing it to human judgements. However, cohesiveness and cohesion of text do not always work well. Sometimes, text cohesiveness *c(S)* does not work well on incoherent text that is lexically cohesive. In addition, in some cases some texts are coherent but are considered incohesive.

On the other hand, (Cutting, 2014) handled the problem of event segmentation. First, the study investigated the problem of narrative shifts in location, character and and time frame. These shifts may align with the viewer’s segmentation of the event (scenes) or may not agree with them. “Scene” is defined here as “a medium-size chunk found in all the narrative arts and often synonymous with the concept of an event”. The paper focused on the signals of new events by indicating the discontinuity across boundaries of visual narrative shots. A scene has three parameters: location, character and time. Narrative shifts are divided into seven types, and depend on the presence or absence of changes in any of the scene parameters (location, character and time). The presence of a change is denoted by (1), and the absence of a change is denoted by (0); the seven types then become 111, 110, 101, 100, 011, 010 and 001. These are applied to movies; then, this approach is compared to other approaches in text that include investigating the effects of shifts in location, characters, time and other factors on readers. It was clear that the shifts in the three indices were uncorrelated.

The study of Kauchak and Chen (Koshorek et al., 2018) investigated the topic of the segmentation of narrative documents. However, it did not tackle the problem of scene segmentation, but the type of text that is handled is narrative. Two narrative books were used here: “Biohazard” by Ken Alibek and ”The Demon in the Freezer” by Richard Preston. The authors of these books segmented them into sections, and these sections were used as the true segment boundary locations. “Biohazard” was split into three parts; two for experimenting and the third for testing, while ”The Demon in the Freezer” was dedicated to testing only. “Biohazard” had 213 true and 5,858 possible boundaries, while “The Demon in the Freezer” had 119 true and 4,466 possible boundaries. The segmentation here was handled as a classification problem in which the segmentation points could have been inferred from the boundaries between the sentences. Thus, data is preprocessed first by tokenisation, stop word removal and stemming. Features used here for the classification include: word group, entity group, entity chains, full name, pronoun, numbers, conversation, and paragraph. Using classification as an approach to the segmentation problem leads to the data losing its sequential nature. It might indicate segments with odd lengths (e.g. only sentences). Thus, it should include features that address the sequential nature of data.

3 The Annotation Framework

We distinguish the conceptual model that underlies our approach to annotation from the concrete syntax we employ in the annotation scheme (cf., e.g, Pustejovsky et al. (2011)).

3.1 The Conceptual Model

Following the general tendency we observe in the literature (Section 2.1), we shall treat a *scene* as a unit of a story in which the time, location and major characters are coherent, i.e. stay essentially the same. A change in any one these constitutes a change of scene.

Note that a scene is an abstract discourse element and does not exist apart from the narrative of which it forms a part. It comprises a location or setting, a time and one or more characters who participate in actions or events that unfold in the scene. By contrast, each of these things (location, time, character and events) *do* indeed exist in the real or fictive world (or *storyworld* as per narrative theory) in which the narrative is set. But the scene itself is an abstraction away from the potentially infinitely complex detail of that real or fictive world, a cognitive construct that makes finite, focussed narrative possible.

A scene is realised in the textual narrative as one or more *scene description segments (SDS)*. An SDS is a contiguous span of text that, possibly together with other SDSs, expresses a scene. Generally, a scene will consist of a single SDS unless that SDS contains embedded SDSs for other scenes, typically for temporally discontinous scenes (e.g. memories of past scenes or imagined future scenes) or spatially distinct locations that are topologically contained within or connected to the embedding SDS, or if the author is employing the narrative device of rotating between multiple concurrent scenes each of which is advancing a distinct story line (a technique very commonly used in action movies).

Since scenes change when *characters*, *time* or *location* change, it seems a good discipline in annotation to identify what these are for each scene. Furthermore, as we have seen in Section 2.2, others such as Cutting (2014) have found it useful to study types of scene changes. Identifying which of character, time and location changes between scenes will contribute to such studies. We see no need to re-invent the wheel and so are hopeful we can simply adopt the definitions, and annotation standards, for times, locations and spatial entities from Iso-TimeML and Iso-Space[1]. For characters, we propose to adopt the definition and annotation standards for named entities of type person, as developed for the ACE program and recently used in, e.g. the TAC 2018 entity discovery and linking task[2].

These standards are appropriate for annotating *all* mentions of times, locations/spatial entities and persons, in texts. However, we are interested in specific times, locations and persons, namely those which represent the time of, location of and characters of the scene in which they occur. Thus, conceptually, we are interested not just in these entities, but in *relations* between these entities and the narrative construct which is the scene.

3.2 SceneML Elements

SceneML comprises two main element types:

1. *Entities*: scenes, scene description segments (SDSs), locations, times, characters
2. *Relations*: scene-scene narrative progression links (there are other relations, but for now we represent them via attributes in entities – see below in Section 5 for a discussion of alternatives and whether this a good way to go).

We give a brief description of each of these here. An extended example that illustrates the use of each is provided in the next section.

[1] `https://www.iso.org/standard/37331.html` and `https://www.iso.org/obp/ui/#iso:std:60779:en`.

[2] See `http://nlp.cs.rpi.edu/kbp/2018/` and `https://www.ldc.upenn.edu/sites/www.ldc.upenn.edu/files/english-entities-guidelines-v5.6.6.pdf`.

Scenes The primary element in SceneML. As indicated above, scenes are a core basic unit in any form of storytelling and involve a coherent setting (location and time) and set of characters, who participate in some form of action and/or dialogue that advances the story. They have as attributes a unique `id`, `time` and `location`. They include a set of `character` sub-elements, since there may be arbitrarily many characters per scene,

Scene description segments (SDSs) SDSs are text spans that are the textual components of scenes. Each SDS is a component of exactly one scene, though a scene may comprise many SDS's. An SDS has the following attributes: a unique `id` and a unique `scene_id`, which is the id of the scene which includes the SDS as a component.

Time Times are essentially timex3 elements as proposed in ISO-TimeML. They have an `id` attribute and a text span. We also allow a special time value of `base`, which is the time in the storyworld of the narrated events.

Location Locations are either location or spatial entity elements from ISO-Space. Like times they have an `id` attribute and a text span. Other attributes from the ISO-Space elements may be adopted as we conduct further corpus annotation.

Character Characters are entities of type `person` as specified in the ACE English Annotation Guidelines for Entities (see note 2 above) with the exception that we do allow characters to be animals or non-human, provided they play the role of a character in the narrative. They have an `id` attribute and a text span. The `type` attribute may be adopted from the ACE specification if it proves useful,

Narrative Progression Links Narrative progression links (`nplinks`) express the type of narrative progression between textually adjacent scenes. For now we identify four types of progression: `sequence`, when one scene follows on from another when, e.g. characters move from one location to another; `analepsis` (or flashback) when then we are taken to another, earlier time and possibly other details such as location and characters change as well; `prolepsis` (or flashforward) when then we are taken forward in time; `concurrent` when we are taken to another location with different characters, where another thread of the story is developing at the same time as the textually preceding scene.

4 Example SceneML Annotations

Figure 1 shows the annotation of several pages of a children's story called *Bunnies from the Future* by Joe Corcoran (Corcoran, 2016). Much of the text has been elided and some of the annotations have been simplified to improve readability. Specifically rather than explicitly annotating times, locations and persons in the text and then using their `id` attribute value in the `scene time` or `location` attribute or in the `character` element, we use the text string itself as the value of the attribute or data element.

5 Discussion

Representing Relations in XML One issue that came up repeatedly in developing the annotation scheme was whether to represent relational information between two entities via an explicit relation element or via an attribute or element associated with one of the entities. So, for example, every `sds` is associated with exactly one `scene`. We have chosen to record this via the `scene_id` attribute of the `sds` element. But it could have been represented in a separate `sds-scene` link entity with two attributes, one for `sds_id` and one for `scene_id`. These representations are informationally equivalent. There may turn out to be processing considerations that favour one over the other; but then the less efficient representation could always be converted to the more efficient by an automated process.

```
<scene id="s1" loc="pod" time="base">
    <character>I</character>
    <character>Skip</character>
    <character>Pockets</character>
</scene>
<sds id = "sds1" scene = "s1">
As we approached our destination, Skip started to issue instructions to the pod about approach vectors ... I
was about to say something ... when Skip opened the door and
</sds>
<scene id="s2" loc="bubble shaped large room" time="base">
    <character>I</character>
    <character>Skip</character>
    <character>Pockets</character>
    <character>Trouble</character>
    <character>Methusaleh</character>
</scene>
<sds id = "sds2" scene_id="s2">
I stumbled out into a blaze of light and noise ... I was in a large room ... It was bubble shaped ... All around
me in the bubble were bunnies ... ” ... Right now, we have to take him to Methuselah. ... I was just about to
ask if Skip was coming in with me, but the door had already opened and they were manoeuvring me through.
</sds>
<scene id="s3"  loc="cylindrical room"  time="base">
     <character>I</character>
    <character>Methusaleh</character>
</scene>
<sds  id = "sds3" scene_id="s3">
I found myself in a room that was cylindrical, like the pod only bigger ... There, in front of, or above, me
(zero gravity is so confusing) was the oldest rabbit I had ever seen ... Methuselah then told me about how the
Bunnies from the Future first came into being.
</sds>
<scene id="s4" loc="planet earth"  time="a long time ago">
    <character>humans</character>
    <character>bunnies</character>
</scene>
<sds  id = "sds4" scene_id="s4">
It was after the plants had turned nasty  a long time ago, even for this old bunny  and the story sounded more
like a legend than real history ... The war was lost, but the bunnies never stopped searching for a way to
achieve victory and to reclaim the planet.
</sds>
<sds id = "sds5" scene_id="s3">
What happened to people? Where are they now? I interrupted, rather rudely ...
<istlink id = "isl1" type = "sequence" scene1 = "s1" scene2 = "s2"/>
<istlink id = "isl2" type = "sequence" scene1 = "s2" scene2 = "s3"/>
<istlink id = "isl3" type = "analepsis" scene1 = "s3" scene2 = "s4"/>
```

Figure 1: Example SceneML Annotations for Chapter 2 of Corcoran (2016)

So probably visual economy and ease of annotation are the primary considerations. Other places where this problem arises are in character-scene relations. We have chosen to handle this by associated multiple character sub-elements with scenes. But these too could be represented as link relations between character and scene ids.

Scene transition signals Some sentences serve to signal a scene transition. Consider this slight variant of one of our example sentences above: *Skip opened the door and I stumbled from the pod out into a blaze of light and noise*. Two questions arise here: (1) should such sentences be included in the first scene, the second scene, in both scenes or in neither? or be split somehow in the middle? (2) should we annotate them, as e.g., `scene_transition_signals`, much the way that temporal and spatial signals (e.g. *before* or *in front of*) are annotated in ISO-TimeML and ISO-Space? We are learning towards annotating them as signals separate from each scene, which will have the advantage of assisting supervised learning algorithms to identify scene transition markers; but we have not reviewed sufficient data yet to make an evidenced recommendation.

6 Conclusion

In this paper we have proposed an annotation framework (conceptual model and XML syntax) for annotating scenes in narrative texts. The definition of scene is based on the relatively widely shared view in narrative studies that scenes change whenever time, location or principal characters shift. Following this view we propose an abstract scene entity, which is realised in text via one or more scene description segments, contiguous sequences of sentences describing the action and dialogue in a scene. Scenes have associated time, location and character information and we propose to adopt previously developed annotation standards for these things. We illustrated our proposal via an extended example and discussed various issues relating to it.

Future work will take the form of an iterative cycle of annotating texts (expanding the type of texts covered not just the quantity) and refining the annotation specification and guidelines. Of course some text will need double annotation and inter-annotator agreement will be assessed; we will also assess the feasibility of crowd sourcing annotation and of using texts with pre-existed scene annotations (e.g. plays, screen plays, etc.). As noted in the introduction our aim is to annotate a large enough corpus to be able train and evaluate and automatic scene segmenter. This will then help enable a range of applications, including narrative analysis tools, book to movie alignment, image description and narrative generation.

References

Agrawal, R., S. Gollapudi, A. Kannan, and K. Kenthapadi (2011). Enriching textbooks with images. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*, CIKM '11, New York, NY, USA, pp. 1847–1856. ACM.

Callaway, C. B. and J. C. Lester (2002, August). Narrative prose generation. *Artif. Intell. 139*(2), 213–252.

Corcoran, J. (2016). *Bunnies from the Future*. www.freekidsbooks.org.

Cutting, J. E. (2014, jun). Event segmentation and seven types of narrative discontinuity in popular movies. *Acta Psychologica 149*, 69–77.

Dai, B., S. Fidler, R. Urtasun, and D. Lin (2017). Towards diverse and natural image descriptions via a conditional gan. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2970–2979.

Dunne, W. (2017). *The dramatic writer's companion: tools to develop characters, cause scenes, and build stories*. University of Chicago Press.

Feng, Y. and M. Lapata (2010). Topic models for image annotation and text illustration. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 831–839. Association for Computational Linguistics.

Halliday, M. A. K. and R. Hasan (1976). *Cohesion in English.* Longman.

Joshi, D., J. Z. Wang, and J. Li (2004). The story picturing engine: finding elite images to illustrate a story using mutual reinforcement. In *Proceedings of the 6th ACM SIGMM international workshop on Multimedia information retrieval*, pp. 119–126. ACM.

Koshorek, O., A. Cohen, N. Mor, M. Rotman, and J. Berant (2018, June). Text segmentation as a supervised learning task. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, New Orleans, Louisiana, pp. 469–473. Association for Computational Linguistics.

Kozima, H. (1993). Text segmentation based on similarity between words. In *Proceedings of the 31st annual meeting on Association for Computational Linguistics*, pp. 286–288. Association for Computational Linguistics.

Kozima, H. and T. Furugori (1993). Similarity between words computed by spreading activation on an English dictionary. In *Proceedings of the sixth conference on European chapter of the Association for Computational Linguistics* -, Morristown, NJ, USA, pp. 232. Association for Computational Linguistics.

Kozima, H. and T. Furugori (1994). Segmenting narrative text into coherent scenes. *Literary and Linguistic Computing 9*(1), 13–19.

Kulkarni, G., V. Premraj, S. Dhar, S. Li, Y. Choi, A. C. Berg, and T. L. Berg (2011). Baby talk: Understanding and generating image descriptions. In *Proceedings of the 24th CVPR*. Citeseer.

Morris, J. and G. Hirst (1991). Lexical cohesion computed by thesaural relations as an indicator of the structure of text. *Computational linguistics 17*(1), 21–48.

Polking, K. (1990). . *Writing A to Z: The terms, procedures, and facts of the writing business defined, explained, and put within reach. Cincinnati, OH.* Writer's Digest Books. ISBN 0-89879-435-8.

Pustejovsky, J., J. Moszkowicz, and M. Verhagen (2011). ISO-Space: The annotation of spatial information in language. In *Proceedings of the Sixth Joint ACL-ISO Workshop on Interoperable Semantic Annotation*, pp. 1–9.

Wang, J. and R. Gaizauskas (2016). Don't mention the shoe! a learning to rank approach to content selection for image description generation. In *Proceedings of the 9th International Natural Language Generation Conference (INLG16).*

Xiao, J., J. Hays, K. A. Ehinger, A. Oliva, and A. Torralba (2010, jun). SUN database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 3485–3492. IEEE.

Yang, Y., C. L. Teo, H. Daumé III, and Y. Aloimonos (2011). Corpus-guided sentence generation of natural images. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 444–454. Association for Computational Linguistics.

Zhu, Y., R. Kiros, R. Zemel, R. Salakhutdinov, R. Urtasun, A. Torralba, and S. Fidler (2015). Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*, ICCV '15, Washington, DC, USA, pp. 19–27. IEEE Computer Society.