HELDA
HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

UNIVERSITY OF HELSINKI

https://helda.helsinki.fi

A coherence approach to data-driven inference in visual communication

Alikhani, Malihe

2019

Alikhani , M , Hiippala , T & Stone , M 2019 , ' A coherence approach to data-driven inference in visual communication ' , Language and Vision Workshop at 2019 Conference on Computer Vision and Pattern Recognition (CVPR) , Long Beach , United States , 16/06/2019 .

http://hdl.handle.net/10138/342541

publishedVersion

Downloaded from Helda, University of Helsinki institutional repository.

This is an electronic reprint of the original article.

This reprint may differ from the original in pagination and typographic detail.

Please cite the original version.

A Coherence Approach to Data-Driven Inference in Visual Communication

Malihe Alikhani
Computer Science
Rutgers University
ma1195@cs.rutgers.edu

Tuomo Hiippala
Department of Languages
University of Helsinki
tuomo.hiippala@helsinki.fi

Matthew Stone
Computer Science
Rutgers University
mdstone@cs.rutgers.edu

Abstract

When people communicate with visual imagery, they intend the audience to recover specific structures and inferences. We propose that these structures and inferences can be modeled using representations and algorithms modeled on approaches to natural language (NL) discourse, particularly coherence relations. We support our argument by describing two successful case studies where we use NL methods to annotate the integrated interpretation of diagrams and pictures in context and to infer the interpretation of these presentations.

1. Overview

We propose organizing image–text presentations and diagrams in terms of coherence relations, a fundamental construct from the theory of natural language discourse that is often invoked to explain the integrated interpretation of the diverse communicative actions in face-to-face conversation [6, 10]. To exemplify, Figure 1 presents steps from two recipes in which images are paired with instruction text. The juxtaposition of the text and the image in each step suggests specific but diverse inferential connections. Figure 1a depicts the action that is described in the text, *lower peaches*, while suggesting that the action has to be completed using a large spoon. The image illustrates the action in progress, and thus shows a moment in the middle of the process that is described in text. The text, on the other hand, provides specific information that is not depicted in the image, *30 to 60 seconds*. In contrast to Figure 1a, Figure 1b shows the result of the action that is described in the text.

Our work has explored the hypothesis that the inferential connections across modalities are fundamentally analogous to those between successive sentences in discourse. We therefore argue that discourse theory provides an important starting point for representing and learning communicative inferences involving visual content.

At the same time, there are also crucial differences between text and imagery. In this paper, we also showcase ways that fine-grained representation for visual communication must (1) acknowledge the inherent differences between text and images and (2) preserve information about how these modalities are combined.

Apart from visual scenes in photographs such as those in Figure 1, many of the entities collectively understood as 'images'—such as diagrams and infographics—integrate illustrations, line art, charts and other forms of graphic expression with natural language. Although both photographs and diagrams share features such as *compositionality* (that is, they consist of distinct elements organized into hierarchies), and *spatiality* (that is, the elements are meaningfully organized in 2D layout space), the kinds of semantic relations that hold between elements are fundamentally different [5]. Whereas photographs feature objects, attributes and interactions, limiting computational reasoning to a given visual context [14], diagrams can involve more diverse structures and reasoning [2].

For this reason, distinct visual modalities such as photographs and diagrams require different approaches to computational processing. Beyond individual modalities, an additional level of complexity emerges at the level of multimodal discourse—such as entire documents—which draw semantic relations between modalities for communicative purposes [11]. In what follows, we describe our previous studies of multimodal discourse to motivate and outline a trajectory for future research.

2. Coherence in Visual Instructions

Our work of image–text presentations [3] emphasizes that text and images are linked together using a constrained set of coherence relations, which can summarize the structural, logical and purposeful relationships between the contributions of text and the contributions of pictures.

To investigate computational methods for studying such inferential links, we have introduced a novel crowd-sourced resource [1]. Authors intend images to communicate specific messages while supplementing the interpretation in text. We make these messages precise by asking a series of questions from subjects for about 2400 image-text pairs.

TEXT: Lower peaches into boiling water and simmer until skins loosen, 30 to 60 seconds.

TEXT: Transfer to airtight container and freeze until firm.

Figure 1: Two steps in recipes illustrating diverse inferential relationships between text and accompanying imagery in instructions. The left image depicts action in progress while elaborating on how one should carry out the action. The right image shows the result of the action that is described in the text.

In [1], we show that our question set can elicit reliable answers from non-expert contributors, and that it enables us to represent and study communicative inferences that involved in understanding imagery. Our questions are modeled after the inferences that connect sentences in text, and lead to many inferences, like the result inference of Figure 1b, that are familiar from text discourse. At the same time, our method allows us to discover inferences that are distinctive to imagery. For example, the image of 1a not only shows an example of how to lower a peach into boiling water (the focused peach in the spoon), but shows the result of doing so (the other peaches already being blanched in the background). Overlapping relations can be found in text, but these particular combinations of relationships distinctively exploit the ability of images to comprehensively depict a complex scene.

In addition to highlighting cases that involve deep and complex inferences, our approach also enables machine learning methods to draw robust inferences about visual content from the associated textual cues. For instance, we asked subjects to highlight parts of the text that are most related to the image.

Table 1 presents the top 5 unigram features of Naive Bayes classifiers as well as top 5 trigram features of NB-SVM related to the highlighted text and its complement. In other two questions, we asked if the text provides quantities that are not in the image or if the image depicts actions that are described in the text.

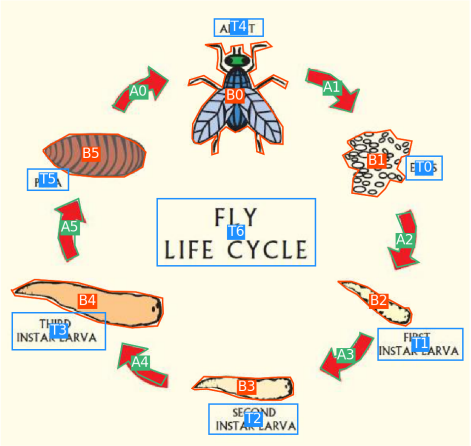
The input of the classifier is the text from the multimodal recipe dataset together with binary labels that describe whether the text describes quantities that are not in images or whether images depict actions in progress.

The results show that quantities and measurements are represented in natural language, whereas images visualize actions and processes that are denoted using action verbs in the accompanying text. [1] describes the details of annotations and the relevant machine learning experiments. Our computational analyses suggest that text content plus an understanding of multimodal coherence can provide strong guides towards understanding associated imagery or selecting imagery to support text content.

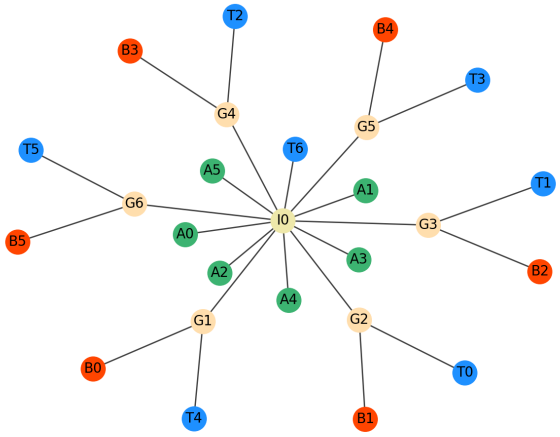
	Image depicts action in progress	
	unigrams	trigrams
1	add	added a beautiful
2	mix	put as much
3	place	skin off of
4	bread	cut side towards
5	make	blend and blend

	Image does not depict action in progress	
	unigrams	trigrams
1	1	do it clearly
2	cup	let cool for
3	minutes	recipe with direction
4	2	how slowly then
5	1/2	7 minutes on

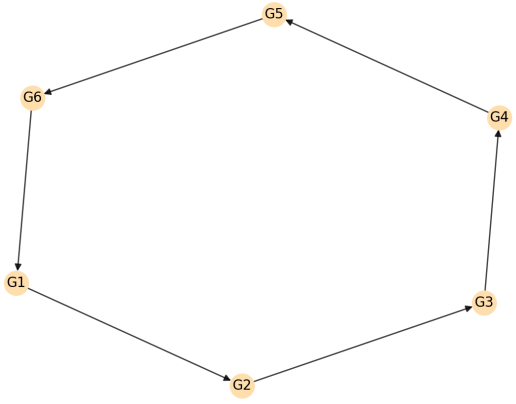
Table 1: Top five unigram features of Naive Bayes classifiers and trigram features of SVM with NB features (NBSVM)[13].



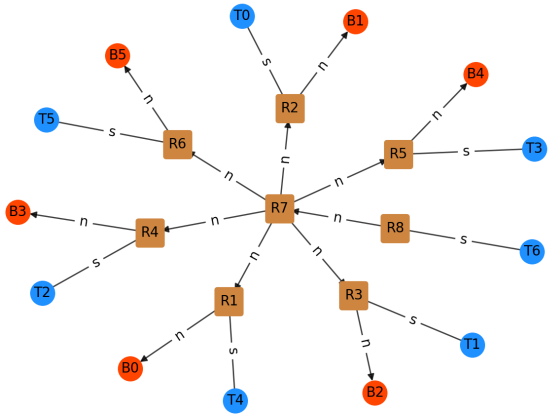
(a) Original crowd-sourced layout segmentation from AI2D. The colours indicate different element types such as text, illustrations and arrows. The element identifiers are carried over to the graphs.



(b) Compositionality: an acyclic graph representing the visual grouping of diagram elements based loosely on Gestalt principles and previous research on diagrammatic representation.



(c) Connectivity: a cyclic graph describing connections signalled using arrows and other diagrammatic elements.



(d) Discourse structure: relations between diagram elements: IDENTIFICATION (R1–6), CYCLIC SEQUENCE (R7) and PREPARATION (R8). Edges indicate whether elements act as nuclei or satellites.

Figure 2: Diagram 2185 from the AI2D dataset [9] described using the annotation schema for diagrams proposed in [7].

3. Coherence in Diagrams

We describe an improved annotation schema for the AI2 Diagrams (AI2D) dataset [9], in which we draw on linguistic theories of discourse coherence to describe relations between diagram elements [8]. Specifically, we use Rhetorical Structure Theory, a theory of discourse structure that has been previously used for various tasks in computational linguistics [12]. We argue that the scope of discourse relations must cover the entire diagram, ranging from local to global relations, as exemplified by the relations between objects and their labels and object–label combinations in Figure 2a, respectively. These relations may be effectively represented using tree graphs, as shown in Figure 2d.

We apply the proposed discourse-based approach in a annotation schema that accounts for multiple diagrammatic structures, and create expert annotations on top of the crowd-sourced layout segmentations available in AI2D [7]. The schema, which is illustrated in Figure 2, uses three graphs with shared identifiers for diagram elements across layers to provide stand-off annotations for (1) compositionality, (2) connectivity and (3) discourse structure.

The annotation for compositionality, shown in Figure 2b, provides the foundation for describing connectivity in Figure 2c and discourse structure in Figure 2d. This means that

the descriptions of connectivity and discourse structure can build on groups of elements as necessary. By pulling apart these structures in diagrams, we seek to understand their individual contributions and how these structures vary depending on diagram type, such as the cycle shown in Figure 2a.

We are working towards building computer systems that can parse and semantically interpret diagrams, which builds on enhancements to the AI2D dataset [8, 7] and the previous work on the interpretation of arrows and sketch recognition models [2, 4]. We envision a diagram parser that detects constituents in a diagram, resolves discourse relations that hold between them and has access to the required encyclopedic knowledge to reason about how these representations relate to the world.

4. Conclusions

We have argued that studying inference in pictures in context requires not just understanding the content of pictures and text but also synchronized integration of modes. We have explored the potential of discourse coherence theory and natural language techniques for annotation and computational analyses of multimodal presentations. Future works involve using such corpora and analyses for building better models of diagrams and visual instructions. Our findings have direct implications for a wide range of applications, such as understanding, generation, summarization of multimodal documents and information retrieval.

5. Acknowledgement

The research presented here is supported by NSF Award 1526723 and through a fellowship from the Rutgers Discovery Informatics Institute. We would like to thank America's Test Kitchen for giving us the permission to include Figure 1 in this publication.

References

[1] M. Alikhani, S. Nag Chowdhury, G. de Melo, and M. Stone. A corpus of image–text discourse relations. In *Proceedings of the NAACL2019*. 1, 2

[2] M. Alikhani and M. Stone. Arrows are the verbs of diagrams. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3552–3563, 2018. 1, 4

[3] M. Alikhani and M. Stone. Exploring coherence in visual explanations. In *2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*, pages 272–277. IEEE, 2018. 1

[4] C. Alvarado and R. Davis. Sketchread: a multi-domain sketch recognition engine. In *ACM SIGGRAPH 2007 courses*, page 34. ACM, 2007. 4

[5] J. A. Bateman. *Text and Image: A Critical Introduction to the Visual/Verbal Divide*. Routledge, London and New York, 2014. 1

[6] R. A. Engle. *Toward a Theory of Multimodal Communication: Combining Speech, Gestures, Diagrams and Demonstrations in Instructional Explanations*. PhD thesis, Stanford University, 2001. 1

[7] T. Hiippala, J. Haverinen, T. Kalliokoski, E. Logacheva, A. Tuomainen, and J. A. Bateman. AI2D-RST: A multimodal corpus of school textbook diagrams. *Manuscript in preparation*, 2019. 3, 4

[8] T. Hiippala and S. Orekhova. Enhancing the AI2 Diagrams dataset using Rhetorical Structure Theory. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, pages 1925–1931, 2018. 3, 4

[9] A. Kembhavi, M. Salvato, E. Kolve, M. Seo, H. Hajishirzi, and A. Farhadi. A diagram is worth a dozen images. In *Proceedings of the 14th European Conference on Computer Vision (ECCV'16)*, pages 235–251, Cham, 2016. Springer. 3

[10] A. Lascarides and M. Stone. Discourse coherence and gesture interpretation. *Gesture*, 9(2):147–180, 2009. 1

[11] M. Taboada and C. Habel. Rhetorical relations in multimodal documents. *Discourse Studies*, 15(1):65–89, 2013. 1

[12] M. Taboada and W. C. Mann. Applications of Rhetorical Structure Theory. *Discourse Studies*, 8(4):567–588, 2006. 3

[13] S. Wang and C. D. Manning. Baselines and bigrams: Simple, good sentiment and topic classification. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2*, pages 90–94. Association for Computational Linguistics, 2012. 2

[14] R. Zellers, M. Yatskar, S. Thomson, and Y. Choi. Neural motifs: Scene graph parsing with global context. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'18)*, pages 5831–5840, 2018. 1