

Arrows are the Verbs of Diagrams

Malihe Alikhani and Matthew Stone
Computer Science, Rutgers University
malihe.alikhani, matthew.stone@rutgers.edu

Abstract

Arrows are a key ingredient of schematic pictorial communication. This paper investigates the interpretation of arrows through linguistic, crowdsourcing and machine-learning methodology. Our work establishes a novel analogy between arrows and verbs: we advocate representing arrows in terms of qualitatively different structural and semantic frames, and resolving frames to specific interpretations using shallow world knowledge.

1 Introduction

Natural communication is multimodal—people get their ideas across not just through words but through gestures, diagrams, illustrations, and even practical activity. Research in discourse has found suggestive evidence for interpretative parallels across modalities; the challenge now is to substantiate these parallels at large scale, by developing broad-coverage cognitive models of multimodal communication.

This paper illustrates the promise of new data sets to further such a general understanding of visual communication. In particular, Kembhavi et al. (2016) have curated, annotated and released a sizeable corpus of scientific diagrams in diverse domains. We draw on formal analysis, crowdsourcing, and machine-learning experiments to carry out a systematic study of the interpretation of arrows in these diagrams. The central insight in our work is to give a novel account of arrows using existing *linguistic* concepts and methodology—specifically, ideas developed for explaining verb meaning in context.

We begin in Section 2 with a tour of work on multimodal communication, which has made a strong case for parsing diagrams to organize basic conventional elements into recursive structural relationships, and for interpreting such parses by combining compositional meaning with discourse-based inference. We use this framework in Section 3 to motivate a linguistically-inspired approach to the interpretation of arrows: we characterize arrows into four qualitatively different structures, analogous to verb subcategorization frames; we associate each structure with a meaning, analogous to verb frame semantics (Baker et al., 1998); and we describe the contextual supplementation some structures require, by analogy to the co-compositionality of generative lexicon theory (Pustejovsky, 1998, GLT).

The remainder of the paper offers empirical evidence in support of our approach. Our first experiments, presented in Section 4, assess our semantic frames for arrows in light of crowd workers’ judgments about the Kembhavi et al. (2016) data set. People label our categories with high agreement, and the categories account for the overwhelming majority of items in the corpus. Moreover, we find that arrows are normally used in one sense per diagram, much as lexical items exhibit one sense per discourse (Gale et al., 1992).

Our next experiments, presented in Section 5, assess our approach to contextual interpretation: certain frames must be supplemented with a salient relationship. Assuming one sense per diagram, we train a machine-learning model that predicts this relationship from the textual content of the diagram. The performance of the method corroborates our hypothesis (analogous to GLT) that these relationships can and should be resolved based on relatively shallow encyclopedic knowledge.

Our contributions are primarily formal—to characterize the knowledge needed to model diagrams. Fundamental challenges remain to apply such ideas in practice. We conclude in Section 6 with an

This work is licensed under a Creative Commons Attribution 4.0 International License. License details: <http://creativecommons.org/licenses/by/4.0/>

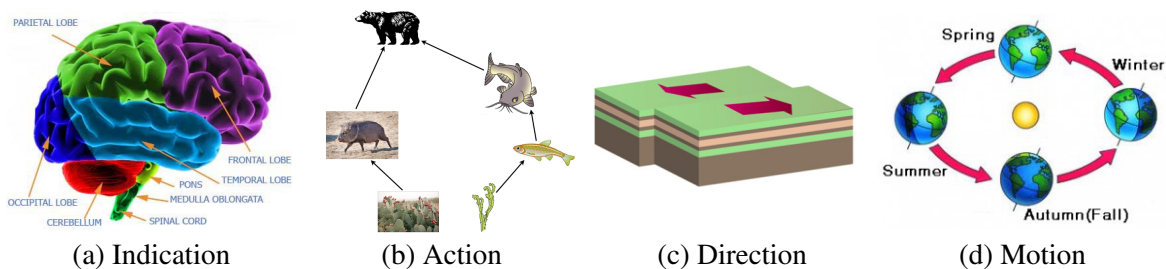


Figure 1: Examples of diagrams with arrows from our four different semantic frames. Diagram (a) has indication arrows, diagram (b) has action arrows, diagram (c) has direction arrows, and diagram (d) has motion arrows. See Section 3.2.

appraisal of our contributions: we suspect many other abstract pictorial elements can be described using linguistic concepts, enabling more expressive systems for diagram understanding and generation.

2 Related Work

Schematic pictorial content abounds in everyday communication, from informal interactions around a whiteboard to figures in formal scientific and engineering documents. Consequently, diagrams have been studied from the standpoint of design, as in Horn (1998); psychology, as in Larkin and Simon (1987); computer graphics, as in Agrawala et al. (2003); computer vision, as in Alvarado and Davis (2004); and common-sense reasoning, as in Forbus et al. (2011)—as well as natural language (NL) technology.

Our work is most directly inspired by prototypes in multimodal communication that use NL techniques either to synthesize communicative presentations, as in André et al. (1993) and Feiner and McKeown (1993), or to interpret multimodal user input, as in Johnston (1998), Johnston and Bangalore (2005) and Bangalore and Johnston (2009). For example, Johnston (1998) argues for processing diagrams with a syntactic parser, so that systems can recognize their essential hierarchical relationships, and for using a constraint-based grammar formalism, so specific productions can inherit constraints from general rule schemas, as in HPSG (Pollard and Sag, 1994). The success of such efforts shows that NL architectures can effectively capture communicative acts across a diverse range of modalities and contexts. This leads us to ask how we might develop domain-general tools and methodologies that could support such multimodal architectures—perhaps following the successful application to text of wide-coverage grammars (Copestake and Flickinger, 2000, among others) and parsers (de Marneffe et al., 2006, among others).

To pursue this direction, we build on an emerging trend of applying the theory of NL semantics and discourse to multimodal communication, which includes work on the grammar of coreference, such as Schlenker and Chemla (2017) for gesture, and Abusch (2013) and Cohn (2013) for comics, and work on the grammar of coherence, such as Lascarides and Stone (2009) for gesture, and Bateman and Schmidt (2013) and Cumming et al. (2017) for film. Looking at arrows is a novel contribution to this literature.

Of course, we can find many loose parallels to our ideas in work on diagrams across many disciplines. For example, researchers from Alvarado and Davis (2004) to Tversky et al. (2000) have observed that diagrams are composed from a small taxonomy of abstract elements, including arrows, that are grouped together in hierarchical relationships. Researchers such as Alvarado and Davis (2004), Forbus et al. (2011), and Horn (1998) have argued that these elements take on diverse interpretations, which must be recovered in context using conceptual content from a specific domain. Researchers such as Agrawala et al. (2003) and Forbus et al. (2011) have emphasized that effective diagrams abstract cognitively essential information in consistent and recognizable ways.

Our findings are entirely compatible with such observations, but our contribution is to approach them using the tools and methods of NL research. Other ways of implementing the ideas include custom architectures based on cognitive design principles, as in Agrawala et al. (2003), or structured user interaction, as in Forbus et al. (2011), as well as general methods for visual recognition and grouping, as in Alvarado and Davis (2004) and Kembhavi et al. (2016). Such approaches may be effective, especially for purely

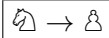



visual processing, but they make it difficult to compare and integrate the reasoning involved in interpreting text and diagrams in combination, as well as to develop models that can be used simultaneously for understanding, generation and mixed-initiative interaction—a key virtue of NL techniques.

Our work is made possible by Kembhavi et al. (2016). They have collected, annotated and distributed a data set of more than 5000 diagrams scraped from the web using science education key phrases as seeds for Google searches. We refer to this as the AI2D data set. Kembhavi et al. (2016) characterize diagram content by associating diagrams with high-level questions and answers. Meanwhile, their work formalizes the structure of diagrams using a graph-based representation that features a regimented set of elements and relationships. Their annotation protocol broke this process of building these representations down into a series of lightweight decisions that could be accomplished reliably by crowd workers on Amazon Mechanical Turk. This annotation scheme, while useful in describing the shallow organization and overall content of diagrams, says little about the semantic and pragmatic knowledge that connects form and content. Our work addresses one case of this open problem.


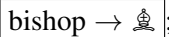
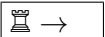
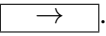
3 Towards a Formal Analysis of Arrows

We start by exploring how linguistically-inspired representations might capture the intuitive structure and interpretation of arrows. We draw on the AI2D examples in Figure 1, which are representative and diverse: Figure 1a maps areas of the human brain; Figure 1b shows part of the food chain of a northern forest; Figure 1c illustrates plate tectonics; and Figure 1d explains the seasons. Our observations about the form, meaning and interpretation of arrows in these diagrams lays the groundwork for the experiments we report in Sections 4 and 5 (and anticipates further formal investigations).

3.1 Formal Structure

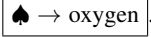
An arrow consists of a head, a tail and a body.¹ The head and tail can each be structurally linked to another constituent in the diagram, to create a larger unit containing the arrow and its affiliates. Thus the arrows of Figure 1a link the text at the tail to the pictorial element at the head; the arrows of Figure 1b link two pictorial elements; and the arrows of Figure 1c are anchored into the scene rather than connected to elements at the head or tail. We assume this structure in giving examples in running text: . This schematizes an arrow with  at the tail and  at the head. Where arrows are not understood to depict spatial information, we will use these schemas to abstract away from details of layout and rendering (so  would be equivalent).

Three classes of form are particularly important to our further investigations:

1. arrows with pictorial elements at both the head and tail, as in ;
2. arrows with pictorial elements at the head and text at the tail, as in ; and
3. arrows with an empty head or tail area (and perhaps pictorial elements elsewhere), as in  or .

We see this structural variation as analogous to the syntactic subcategorization of verbs: the idiosyncratic character of verbal complementation motivates a fine-grained account of syntactic combination. For example, the verb *get* can take arguments of a range of different syntactic categories, involving noun phrases, predicates and full sentences: “I got flowers for my mother”, “Get well soon”, “I get that this is difficult”. In a similar way, arrows have optional argument positions (head and tail, comparable to subject and object) which may be filled by elements of formally different categories (for example, with textual or pictorial elements).

¹We focus on simple arrows here of the form \rightarrow , excluding double-headed arrows \leftrightarrow and pure line segments (---). Arrows can also differ in rendering attributes such as width, perspective and color, which sometimes provide further cues to the role of the arrow in the diagram.

²Arrows with pictorial elements at the tail and text at the head also occur: . We hypothesize that these have analogous properties to purely pictorial arrows, but defer investigation to future work.

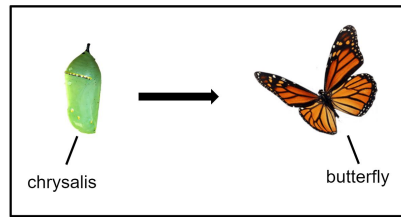


Figure 2: Part of a butterfly cycle of life diagram.

Just as with verbs, some structural possibilities for arrows seem not well-formed: it’s odd, for example, to have text at the tail of an arrow and nothing at its head, as in # problem →.³

As noted by many researchers (Alvarado and Davis, 2004; Johnston, 1998; Tversky et al., 2000), diagrams are hierarchically organized. We hypothesize that arrows can contribute to the overall structure of a diagram in diverse ways, just as verbs contribute in diverse ways to the structure of sentences and discourse. Of course, an arrow can establish the overall structure of a diagram, as in Figure 2—just as the main verb projects the root of a sentence. Arrows in diagrams can also be modified, for example with labels and other annotations; we show a few such examples in Figures 3 and 4. In addition, many arrows seem to attach to and modify existing elements, just as verbs can anchor relative clauses, which adjoin in as modifiers to existing structure. For example, we might analyze the indication arrows of Figure 1a as modifiers that expand on the presentation of individual brain regions, with the head of the arrow playing a role analogous to an extracted argument in a subject or object relative clause—compare *the region we call the frontal lobe*. Similarly, the direction arrows of Figure 1c are similar to locative relative clauses—compare *a place things are pushed eastward*. Finally, the arrows of Figure 1b or 1d seem broadly analogous to the use of sentence-final appositives to carry the discourse forward (Koev, 2012)—compare *the prickly pear is eaten by the boar, which is eaten by the bear*. Note, however, that the two-dimensional structure of diagrams allows for new structural possibilities, such as the circular dependencies exhibited in Figure 1d. Formalisms for encoding diagram structure involve generalized notions of syntactic combination (Johnston, 1998). Integrating our account of arrow structure into such a generalized syntactic framework remains a topic for future work.

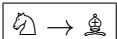
3.2 Meaning Frames


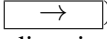
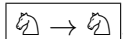
Arrows have qualitatively different meanings. One difference is whether the path has spatial meaning. For example, the arrows in Figure 1b indicate temporal or causal relationships with no spatial content. By contrast, the arrows of Figure 1c are rendered in perspective with the diagram’s tectonic plates to depict the direction of action in an earthquake.

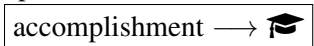
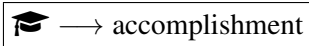
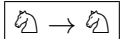
These qualitatively different interpretations are correlated with the structural description of an arrow. We see these constraints as analogous to the semantic frames assigned to verbs in databases like FrameNet (Baker et al., 1998). A semantic frame is a conceptual structure describing an event or a relation and the participants in it. For example, “Mary sold her car to John” and “John bought Mary’s car” describe the same semantic frame. We interpret each structural option with an appropriate underlying conceptual representation, or frame, and assign each argument an appropriate role in the frame. Our frames for arrows are INDICATION, ACTION, DIRECTION, and MOTION.

- The INDICATION frame, for arrows like those in Figure 1a, involves a SUBJECT (prototypically, a pictorial element at the head), and a LABEL (text at the tail), as in bishop → ♁. In indication, the label represents the name, location or category of the subject.
- The ACTION frame, for arrows like those in Figure 1b, involves a SOURCE (prototypically, a pictorial element at the tail) and a TARGET (prototypically, a pictorial element at the head), which are

³The distinctive status of text means that arrows with text at the tail are often equivalent to line segments with text at one end. In fact, only 19.09% of the labeling relations in AI2D are drawn with arrows; the rest are line segments. Other arrows cannot be replaced by segments.

understood to stand in a contextually-specified telic or agentive RELATION, as in . Such arrows may represent causation, change, consumption, or transfer of possession depending on the content of the diagram. Source and target may represent different objects or different states of the same object.

- The DIRECTION frame, for arrows like those in Figure 1c, involves the DIRECTION indicated by the rendering of the arrow. This direction gives a contextually-specified ATTRIBUTE of a SUBJECT, which may be an element linked to the arrow (as in ) but may also be whatever lies where the arrow sits (as in ). The implicit meaning of such arrows has to be recovered using context. For instance, direction arrows can show the direction of traffic in street signs, the direction of the wind in weather maps, and the flow of material in mechanical drawings.
- The MOTION frame, for arrows like those in Figure 1d, involves a MOVER, which is rendered *twice*, both at the head and at the tail of the arrow, as in . The frame indicates a translation event in which the mover starts in the tail configuration and ends at the head configuration.

Minimal pairs underscore the qualitative distinctions among these interpretations, which must therefore rely on conventional knowledge, not just common sense. An element  probably indicates that graduation *amounts to* an accomplishment (the INDICATION frame), while the inverse element  probably means that graduation *leads to* accomplishment (the ACTION frame). The possibilities for ambiguity also support qualitative analyses of arrow meaning. For example, in some diagrams, the element  would represent an event in which a single person moves through space (the MOTION frame); in others, it would represent one in which a first person does something to a second (the ACTION frame). The experiments in Section 4 are focused at assessing whether (and how) people represent and resolve such ambiguities.

3.3 Resolving Interpretation

Designers like Horn (1998) enumerate the specific interpretations of arrows seen in Figure 1: the meaning *be eaten* of the arrows in Figure 1b, the meaning *driving force* of the arrows in Figure 1c, or the meaning *orbits* of the arrows in Figure 1d. Such alternatives cannot be listed in a wide-coverage approach; they have to be derived from the specific purpose and subject matter of the diagram. Work on verbs, such as the generative lexicon theory of Pustejovsky (1998), offers a suggestive model for how this derivation might go. Pustejovsky (1998) proposes that underspecified verbs (famously, *begin*) retrieve implicit actions via the *qualia structure*—the natural action associates—of their arguments. The broader lesson is that shallow associations can be a surprisingly good place to look for contextual interpretations. The experiments of Section 5 substantiate this suggestion by seeing what features are needed to recover the context-sensitive relation associated with the ACTION frame.

4 Distinguishing Semantic Categories

The linguistic approach of Section 3 raises empirical questions about qualitative ambiguities in diagrams. Do people distinguish between arrows that signal indication, action, direction and motion frames? Can we develop corpora that annotate our frame distinctions explicitly? AI2D annotations already distinguish indication arrows from other arrows, but do not encode our other distinctions, so we conducted a series of human-subjects experiments to address these questions. The studies were conducted with the approval of our human subjects review committee. The data is available electronically.⁴

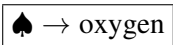
4.1 Materials

The annotations in the AI2D dataset suggest that 2077 diagrams have arrows, but the annotations are not always accurate. We removed 196 diagrams that do not have arrows. Further, to focus on clear cases, we removed 232 diagrams: those whose action arrows involved words, question marks or letters at the

⁴https://github.com/malihealikhani/Arrows_are_Verbs

arrow type	count		txt-obj	two-loc	empty	other
obj-obj	985	obj-obj	6	3	3	17
txt-obj	378	txt-obj		2	2	2
two-loc	138	two-loc			16	2
empty	70	empty				1
other	17					

Table 1: Results of experiment 1. Left, counts of diagrams tagged with a unique arrow category. Right, counts of diagrams tagged with exactly two arrow categories. We had only 5 diagrams tagged with three categories and no diagrams tagged with more.

head rather than pictorial elements (as in our  example, the majority of exclusions); those whose illustrations were cropped and incomplete, and cases such as “flash cards” that involved multiple renderings (a small number of further cases). This left us with 1634 diagrams from the AI2D dataset.

4.2 Methods

In the first experiment, our goal was to look at the structural features of arrows, as well as the semantic distinction between the MOTION and ACTION frames. (In pilot studies, we found that offering subjects a broad range of options let them categorize arrows more reliably.) In each interaction, subjects saw a diagram and were asked to answer if any arrows in this diagram are:

1. between two different objects
2. between a label and an object
3. between two different locations of an object
4. between empty areas
5. none of the above

Option 1 is indicative of arrows in the ACTION frame (obj-obj). Option 2 corresponds to the INDICATION frame (txt-obj); Option 3, to the MOTION frame (two-loc); and Option 4, to the DIRECTION frame (empty). Subjects were able to choose more than one answer, as diagrams sometimes include multiple kinds of arrows. (See Figure 3.)

The goal of the second experiment was to see if people interpret empty arrows in line with the DIRECTION frame. In each interaction, subjects saw one of the 93 diagrams labeled as having empty arrows in Experiment 1, and answered whether any arrows in this diagram showed:

1. the direction of the movement of the objects
2. the different locations of the objects
3. none of the above

Option 1 and 2 explain the interpretation of the DIRECTION (dir) and MOTION (mov) frames respectively.

In all, we recruited 860 subjects through Amazon Mechanical Turk. subjects were all US citizens, high school graduates and had a computer to be able to participate in the experiments. Subjects gave written consent and received \$0.10 for the first experiment and \$0.30 for the second experiment for their participation in each human intelligence task (HIT), an estimated hourly rate of \$15.00. Each HIT asked subjects to characterize all the arrows in three separate diagrams. Each diagram has one data point in the dataset; in addition, we collected additional data collecting multiple responses for a random subset of the data (544 diagrams in Experiment 1, 47 in Experiment 2) to assess reliability.

4.3 Results

We followed Chen et al. (2005) in measuring inter-rater agreement. Their method is a generalization of Kappa statistics for cases that multiple raters are involved in a multi categorization task. For the first experiment, the overall kappa is 0.9744. For the second experiment, the overall kappa is 0.8517. This

type	count		mov	other
dir	72	dir	3	7
mov	9	mov		2
other	0			

Table 2: Results of Experiment 2. Left, diagrams with a unique response. Right, diagrams with two responses. No diagrams had all three responses.

Diagram category	No. of items	Top arrow	Arrow rate
Group A	353	txt-obj	97.73%
Group B	1007	obj-obj	94.44%
Group C	221	two-loc	84.31%
Group D	71	empty	91.57%

Table 3: Grouping diagrams by topic gives a strong indication of the structure and meaning of arrows.

indicates that subjects strongly agreed on their judgements, suggesting that the structural and interpretive distinctions that underwrite our semantic frames are clear to human interpreters.

This established, we can look at the distribution of annotations. Our first results appear in Table 1. Observe that only 38 diagrams seem to have an arrow classified in the “other” category; this suggests that our taxonomy has good coverage of frequent patterns. Further, the vast majority of the diagrams (1571 of 1634) feature exactly one kind of arrow; this is reminiscent of findings that words are used with one sense per discourse (Gale et al., 1992). This suggests that diagram understanding systems should be designed to look at a diagram holistically and prefer uniform interpretations for all the arrows.

Table 2 specifically focuses on the second experiment, which looked at whether DIRECTION is the appropriate semantic frame for empty arrows. The results are again consistent with our meaning frames. In particular, the 72 diagrams tagged as having just direction interpretations in Experiment 2 include 68 of the 70 diagrams tagged as having just empty arrows in Experiment 1. In addition, 12 out of 16 diagrams that were tagged with more than one option in Experiment 1 were reported to have multiple kinds of arrows in Experiment 2.

The AI2D dataset groups diagrams into different categories such as “Atom Structure” and “Moon Phase Equinox”. We made four larger groups of diagrams categories for diagrams in AI2D: Group A includes diagram that map out scientific structures; Group B includes diagrams that present interactions between elements; Group C explains changing geometric configurations; and Group D highlights momentary spatial relationships.⁵ These categories turn out to be highly correlated with a particular structure and interpretation of arrows: see Table 3. This provides further evidence that diagram interpretation should be guided holistically by general domain knowledge.

4.4 Exceptions

In an effort to understand the limitations of our taxonomy, we inspected the diagrams tagged as having multiple types of arrow and additional types of arrow.

We found that when designers use arrows to represent different meanings, they almost always render those arrows with different shapes or colors to represent different meanings. Each kind of arrow therefore has a consistent interpretation in the diagram after all. The left and the middle diagrams in Figure 3 show two examples of diagrams of this kind. The right diagram however is an example of the few that do not follow this principle.

We designed an experiment with the same terms and conditions as Experiment 2 to test the hypothesis that different kinds of arrows in complex diagrams have different appearances. We chose all diagrams that were tagged in Experiment 1 with more than one option (excluding diagrams that were tagged with “other”) and asked participants to report how many different kinds of arrows with different appearances they see in each diagram. The result of the experiment shows that 21 out of 23 diagrams were reported to have more than one kind of arrow. This is in particular important in designing systems that can generate or understand diagrams. Such systems must allow each shape to get a different interpretation.

⁵Specifically, Group A is AI2D typesOf, partsOfA, atomStructure, partsOfTheEarth, circuits, solarSystem, rockStrata. Group B is Photosynthis, lifeCycles, foodChainsWebs, waterCNPCycle, volcano, rockCycle. Group C is moonPhaseEquinox. Group D is faultsEarthquakes, eclipses.



Figure 3: Authors typically render arrows with different interpretations differently. Left: thin circular arrows depict the moon’s orbit (motion) while thick straight arrows depict the force of gravity (direction). Right: red arrows represent the direction of the movement of the moon where the white arrows are between two different locations of the moon.



Figure 4: Examples of diagrams tagged as having other kinds of arrows.

Diagrams with arrows that were labelled as having additional kinds of arrows don’t straightforwardly suggest omissions in our structural inventory. One case we might consider is arrows linked at their head to other arrows—a generalization of our INDICATION frame. The right diagram in Figure 4, includes such arrows. One arrow shows the direction of flow of the atmosphere; another arrow links that arrow to text explaining that the flow arises through the moon’s gravity. A few other diagrams that come with arrows that are annotating another set of arrows were tagged other as well.

However, most of the “other” examples may simply have been difficult to understand. Some of these diagrams, like the left diagram in Figure 4, have arrows whose path are cutting each other. Similarly, some have labels that are not in proper places—for example very far from the tail of the arrow—or other unnecessary and distracting elements, such as mountains and clouds in a food chain diagram. These observations suggest that certain visual features can increase users’ cognitive load in interpreting action diagrams. We suspect that viewers’ cognitive limitations are an important constraint in diagram generation—a suggestion we return to in the conclusion.

4.5 Limitations

The experiments of Section 4 suggest that people agree on the structural organization of arrows in diagrams, and they also assign arrows consistently into semantic frames based on the identity of depicted objects and the broad significance of the arrow. While these results support a linguistic approach to diagram organization, they require substantial development to inform automatic diagram understanding. Automatic methods have trouble recognizing structural links between diagram elements (Kembhavi et al., 2016). Successful practical methods for diagram parsing rely on sketch stroke order (Johnston, 1998; Alvarado and Davis, 2004), which is not available in AI2D. Inferring coreference among schematic pictorial elements is also a challenging open problem, since current machine learning models of computer vision are notoriously bad at line art.

5 Interpretation of Action Arrows

The result of the experiments in section 4 showed that the specific import of arrows varies across contexts, but is typically consistent within a diagram and strongly contained by its subject matter. This means that once we know that an arrow falls into the direction, motion or indication category we have the interpretation of the arrow. However, in order to come up with the exact relation that explains the action arrows, we need to take a further step. In this section, we look specifically at how to predict the specific relation associated with an action arrow. Previous work has focused on resolving these interpretations using handcrafted knowledge bases describing specific domains (Alvarado and Davis, 2004; Forbus et al., 2011). Our analogy with verbs suggests that shallow methods based on element statistics and simple machine learning methods may provide more practical methods for inferring the domain-specific meaning of arrows in particular cases.

We carried out computational experiments to assess this hypothesis. Our machine learning classifier takes as input the verbal information in a diagram, such as label text and other annotations, and predicts the likely interpretation of all its action arrows.

5.1 Dataset and experimental setup

To link action arrows with their interpretations, we ran a pilot study on Mechanical Turk with 56 diagrams (5% of the diagrams that were tagged with action arrows in Experiment 1). Each subject saw a single diagram and responded to the prompt: “If you remove the arrows in this diagram, what words would you replace them with to convey the same meaning?” The results were consistent with the hypothesis that each diagram uses arrows to convey a single relation; unsurprisingly, however, the responses used diverse vocabulary (Furnas et al., 1987), so could not be used directly as training data for a classifier.⁶ Instead, we associated diagrams with arrow meanings consistently based on AI2D categories: We used “turn into” for cycle of life and rock cycle diagrams; “eat/be eaten” for food chain web diagrams and “produce/receive” for photosynthesis diagrams.

AI2D includes multiple diagrams with similar content. We want to test whether we can learn general patterns for classifying arrows, not just memorize the features of individual examples. Thus, we organized our experiment in blocks using diagrams with different subject matter. We divided the data into three: we randomly split the receive class (it has only 93 datapoints), but we divide the two other groups, turn into and eats, based on three existing categories in AI2D (cycle of rocks, volcano, cycle of life) and three subcategories that we found in food web data (ocean, large birds, big cats). Since the number of blocks is small, we report cross-validation results leaving out one fold as test data.

5.2 Models

We used a bag of words (BoW) model with the SVM classifier as our baseline.

We compared the baseline to a convolutional neural network (CNN) with one convolution layer over pretrained word embeddings (Collobert et al., 2011; Kim, 2014). We used a max-over-time pooling operation and dropout regularization. See Table. 5. This setting enabled us to measure the effectiveness of considering distributional similarity in our classification task and thus the importance of generalizing based on shallow semantic knowledge.

To see how important it is to semantically understand the input by using sequential and structural information, we experiment with hierarchical attention RNNs (Yang et al., 2016). We made a look-up table for vocabulary, converted all words to integers and applied a one hot encoder. Next, we padded the input sequences, and fed them through Embedding, LSTM and Dense layers. Model Architectures are described in detail in Appendix A.

We trained the models to minimize binary cross entropy. Text was fed in left to right and top to bottom based on the position in the diagram.

⁶44 responses consisted of a single verb (differing across subjects: turn into, metamorphose), and 5 included two verbs with similar meanings (develop, turn into), while 7 included nouns or short sentences.

	data size	CNN	RNN	SVM
fold 1	train: 1175	0.859	0.704	0.571
	test: 269			
fold 2	train: 1101	0.841	0.749	0.623
	test: 343			
fold 3	train: 1287	0.862	0.673	0.466
	test: 175			
Avg	train: 1187	0.842	0.708	0.553
	test: 262			

Table 4: Accuracy of predicting implicit RELATION for diagrams with ACTION arrows.

5.3 Results

Table 4 gives our experimental results. The CNN architecture works best. Its overall effectiveness corroborates our hypothesis that arrows pick up straightforward action associates of the objects they connect. The improvement over the SVM model suggests that distributional similarity is an effective proxy to the world knowledge needed to understand a diagram. The improvement over RNNs suggests that linguistic structure is not important for this task.

Error analysis revealed three major problems. The majority of the examples that our models failed to classify were the ones that contain labels that are shared between categories. For example, words such as “animal” and “plant” show up in both “turn into” and “eat” categories. We also had problems with diagrams labeled in languages other than English, and some diagrams with irrelevant labels, such as photosynthesis diagrams with “car” and “factory” labels.

Working with labels enable us to take into account the pictorial information that is available in a diagram without using a vision classifier. We can see that having a knowledge of the category of the diagram suffices for resolving the underspecification problems.

6 Conclusion

Arrows, we have argued, have a constrained form, meaning and interpretation, which can be theorized using linguistic concepts. This means, in particular, that we can ask subjects to report their interpretation of arrows in linguistic terms and use machine learning methods to resolve ambiguities in the interpretation of arrows in qualitative, linguistically-motivated ways. Vision researchers sometimes compare diagram understanding to scene understanding (Kembhavi et al., 2016). That may be right when it comes to grouping elements in diagrams together and recognizing pictorial elements—though much research on those topics is still needed. When it comes to organizing the content of a diagram as a whole, we think discourse modeling and other NL techniques also offer important tools.

Many of our findings echo designers’ advice about effective diagrams (Tufte, 2001). They are simple. They have limited figurative content. Most of them have one kind of arrow that’s easy to interpret based on general background knowledge. If they have multiple kinds the shape of the arrows are different so that the reader can distinguish between the two kinds of meanings that he has to infer. Our results therefore promise to help inform systems for diagram generation (Agrawala et al., 2003).

Our empirical work explores only a small number of the ways arrows vary—and diagrams include a wide range of other elements that we think can be studied with methods like ours. More generally, the minimal pairs we sketched in Section 3 suggest that we can operationalize interpretive constraints on multimodal communication with predictive, wide-coverage methods. We hope to contribute to such formalisms going forward.

7 Acknowledgement

This work was supported by NSF Award IIS-1526723 and has benefited from discussions with Doug DeCarlo, Gabriel Greenberg and anonymous reviewers.

References

- Dorit Abusch. 2013. Applying discourse semantics and pragmatics to co-reference in picture sequences. In Emmanuel Chemla, Vincent Homer, and Grégoire Winterstein, editors, *Proceedings of Sinn und Bedeutung 17*, pages 9–25, Paris.
- Maneesh Agrawala, Doantam Phan, Julie Heiser, John Haymaker, Jeff Klingner, Pat Hanrahan, and Barbara Tversky. 2003. Designing effective step-by-step assembly instructions. *ACM Trans. Graph.*, 22(3):828–837.
- Christine Alvarado and Randall Davis. 2004. Sketchread: a multi-domain sketch recognition engine. In *Proceedings of the 17th Annual ACM Symposium on User Interface Software and Technology, Santa Fe, NM, USA, October 24-27, 2004*, pages 23–32.
- Elisabeth André, Wolfgang Finkler, Winfried Graf, Thomas Rist, Anne Schauder, and Wolfgang Wahlster. 1993. WIP: the automatic synthesis of multimodal presentations. In Mark T. Maybury, editor, *Intelligent Multimedia Interfaces, the book is an outgrowth of the AAAI Workshop on Intelligent Multimedia Interfaces, Anaheim, CA, USA, August, 1991.*, pages 75–93. AAAI Press / The MIT Press.
- Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. The berkeley framenet project. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 1*, pages 86–90, Montreal, Quebec, Canada, August. Association for Computational Linguistics.
- Srinivas Bangalore and Michael Johnston. 2009. Robust understanding in multimodal interfaces. *Computational Linguistics*, 35(3):345–397.
- John Bateman and Karl-Heinrich Schmidt. 2013. *Multimodal film analysis: How films mean*, volume 5. Routledge.
- Bin Chen, D Zaebst, and Lynn Seel. 2005. A macro to calculate kappa statistics for categorizations by multiple raters. In *Proceeding of the 30th Annual SAS Users Group International Conference*, pages 155–30.
- Neil Cohn. 2013. *The Visual Language of Comics: Introduction to the Structure and Cognition of Sequential Images*. A&C Black.
- Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, 12(Aug):2493–2537.
- Ann A. Copestake and Dan Flickinger. 2000. An open source grammar development environment and broad-coverage english grammar using HPSG. In *Proceedings of the Second International Conference on Language Resources and Evaluation, LREC 2000, 31 May - June 2, 2000, Athens, Greece*.
- Samuel Cumming, Gabriel Greenberg, and Rory Kelly. 2017. Conventions of viewpoint coherence in film. *Philosophers' Imprint*, 17(1):1–29.
- Marie-Catherine de Marneffe, Bill MacCartney, and Christopher D. Manning. 2006. Generating typed dependency parses from phrase structure parses. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation, LREC 2006, Genoa, Italy, May 22-28, 2006.*, pages 449–454.
- Steven Feiner and Kathleen McKeown. 1993. Automating the generation of coordinated multimedia explanations. In Mark T. Maybury, editor, *Intelligent Multimedia Interfaces, the book is an outgrowth of the AAAI Workshop on Intelligent Multimedia Interfaces, Anaheim, CA, USA, August, 1991.*, pages 117–138. AAAI Press / The MIT Press.
- Kenneth D. Forbus, Jeffrey M. Usher, Andrew M. Lovett, Kate Lockwood, and Jon Wetzel. 2011. Cogsketch: Sketch understanding for cognitive science research and for education. *topiCS*, 3(4):648–666.
- George W. Furnas, Thomas K. Landauer, Louis M. Gomez, and Susan T. Dumais. 1987. The vocabulary problem in human-system communication. *Communications of the ACM*, 30(11):964–971.
- William A. Gale, Kenneth W. Church, and David Yarowsky. 1992. One sense per discourse. In *Proceedings of the Workshop on Speech and Natural Language, HLT*, pages 233–237, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Robert E. Horn. 1998. *Visual Language: Global Communication for the 21st Century*. MacroVU, Incorporated.

- Michael Johnston and Srinivas Bangalore. 2005. Finite-state multimodal integration and understanding. *Natural Language Engineering*, 11(2):159–187.
- Michael Johnston. 1998. Unification-based multimodal parsing. In *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, COLING-ACL '98, August 10-14, 1998, Université de Montréal, Montréal, Québec, Canada. Proceedings of the Conference.*, pages 624–630.
- Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. 2016. A diagram is worth a dozen images. In *European Conference on Computer Vision (ECCV)*.
- Yoon Kim. 2014. Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882*.
- Todor Koev. 2012. On the information status of appositive relative clauses. In *Logic, language and meaning*, pages 401–410. Springer.
- Jill H. Larkin and Herbert A. Simon. 1987. Why a diagram is (sometimes) worth ten thousand words. *Cognitive Science*, 11(1):65–100.
- Alex Lascarides and Matthew Stone. 2009. A formal semantic analysis of gesture. *Journal of Semantics*, 26(4):393–449.
- C. Pollard and I.A. Sag. 1994. *Head-Driven Phrase Structure Grammar*. Studies in Contemporary Linguistics. University of Chicago Press.
- James Pustejovsky. 1998. *The Generative Lexicon*. Bradford Books. MIT Press.
- Philippe Schlenker and Emmanuel Chemla. 2017. Gestural agreement. *Natural Language & Linguistic Theory*, Oct.
- Edward R. Tufte. 2001. *The Visual Display of Quantitative Information*. Graphics Press.
- Barbara Tversky, Jeff Zacks, Paul U. Lee, and Julie Heiser. 2000. Lines, blobs, crosses and arrows: Diagrammatic communication with schematic figures. In *Theory and Application of Diagrams, First International Conference, Diagrams 2000, Edinburgh, Scotland, UK, September 1-3, 2000, Proceedings*, pages 221–230.
- Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alexander J Smola, and Eduard H Hovy. 2016. Hierarchical attention networks for document classification. In *HLT-NAACL*, pages 1480–1489.

A Learning Experiment Parameters

Model learning code is available with the electronic resources accompanying this paper at https://github.com/malihealikhani/Arrows_are_Verbs. Key parameters for our machine learning experiments are provided in Table 5.

Model	Operation	Input	Kernel size	Output	Activation
CNN	Embedding	–	–	128	–
	Conv	Embedding	(3, 128)	100	Relu
	Dropout	Conv	–	300	rate=0.5
	Dense	Dropout	100	1	Sigmoid
RNN	Embedding	–	–	(20, 64)	Relu
	LSTM	Embedding	64	64	–
	Dense	LSTM	64	3	Tanh

Table 5: CNN and RNN architecture details.