A promising visual approach to solution of 82% of Winograd Schema problems via Tumbug Visual Grammar

Mark A. Atkins, Ph.D.
aginik@outlook.com

December 26, 2023

Abstract

This 2023 document is a wrapper that embeds the author's original 2022 article of the above title that has never been publicly available before. The embedded article is about Phase 1 (which is about Tumbug) and Phase 2 (which is about non-spatial reasoning) of the 5-phase Visualizer Project of the author, a project that is still in progress as of late 2023. The embedded article is currently being re-released by the author to supply more information about that project to the public, and for historical reasons. The embedded article was written before a much more thorough article about Phase 1 (viz., "Tumbug: A pictorial, universal knowledge representation method") became available in 2023, but the embedded article describes results from Phase 2 that have not yet been documented elsewhere.

1 The article

A promising visual approach to solution of 82% of Winograd Schema problems via Tumbug Visual Grammar

Mark A. Atkins

Abstract

A new type of approach is presented for solving common-sense reasoning problems from the Winograd Schema. The proposed approach uses three novel knowledge representation methods, all visual and all integrated: (1) Tumbug, which a universal, visual grammar that is a type of labeled, directed graph, (2) SOAV, which is a system level extension to OAV, where OAV is the familiar Object-Attribute-Value triple of object-oriented programming, and (3) a concepedia, which is a computer-readable visual encyclopedia of concepts that effectively contains commonsense reasoning rules. The approach should apply to 82% of the Winograd Schema problems, though the approach has not yet been computer simulated to confirm this claim. The approach consists of using Tumbug-SOAV notation for both the problem and the concepedia, then of applying one image matching function per Winograd Schema problem, which makes the approach extremely fast, automatic, and reflexive. This direct approach to commonsense reasoning avoids indirect, short-term, technical tricks.

1 INTRODUCTION

Real progress in the field of artificial intelligence (AI) depends on the critical but currently underdeveloped, currently unapplied subfield called artificial general intelligence (AGI). In the quest to produce AGI one solid approach is to tackle one of the well-known, longstanding subfields of AGI that are believed to be key. One such subfield is commonsense reasoning (CSR), which currently all computers and their programs still do very poorly ([1], p.1). Within CSR there exists a list of benchmark problems called the Winograd Schema (WS), which is a very convenient testbed for any new AGI approach, which this article's proposed approach is. For brevity the set of 150 problems in the version of the WS dataset (from [2]) that was used in this study is called "WS150" in this document (cf. [3], p. 6). A small percentage of these problems are called "Broad Sense" because they deviate from Terry Winograd's intended required problem characteristics [2].

The proposed approach is based on a striking discovery by the author in 2022 is that there exists a single visual matching algorithm that can solve the vast majority of WS150 problems. The knowledge representation method (KRM) used is called "Tumbug," a convenient written form of "TUMBVG," or "Temporal Universal Model-Based Visual Grammar" [4], developed by the author. The Tumbug KRM is combined with the SOAV KRM to produce the Tumbug-SOAV KRM, where "SOAV" means "System-OAV," where "OAV" (Object-Attribute-Value) is a term used in object-oriented programming (OOP). SOAV is the author's higher-level extension of OAV that considers "systems" of objects, connected by arrows and other icons to represent relationships between the objects. Tumbug Visual Grammar differs from Tumbug Textual Grammar, the latter of which lacks diagrams in the usual sense.

2 RECOMMENDED CATEGORIZATION OF WS150 PROBLEMS

This section lists all WS150 problems within categories relative to the proposed Tumbug-SOAV approach. So far, only the Non-Spatial Reasoning problems can be consistently solved by the Tumbug-SOAV algorithm.

—DISCRETE REASONING—

Non-Spatial Reasoning: 1, 3, 4, 5, 6, 7, 8, 9, 12, 13, 15, 16, 17, 18, 19, 20, 21, 23, 24, 25, 26, 27, 28, 29, 32, 33, 35, 36, 37, 38, 39, 41, 42, 43, 44, 46, 47, 48, 49, 50, 51, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 67, 68, 69, 70, 71, 73, 76, 78, 79, 80, 81, 82, 83, 86, 87, 88, 89, 90, 91, 92, 93, 95, 96, 97, 98, 99, 100, 102, 104, 105, 106, 108, 109, 110, 111, 112, 113, 114, 115, 116, 117, 118, 119, 121, 122, 123, 124, 125, 126, 127, 128, 129, 130, 131, 132, 133, 134, 135, 136, 138, 139, 140, 141, 142, 143, 144, 145, 146, 147, 148, 149
Spatial Reasoning, Algorithmic: 2, 10, 22, 34, 52, 53, 66, 72, 77, 84, 101, 103, 120
Spatial Reasoning, Non-Algorithmic: 11, 14, 30, 31, 40, 45, 54, 74, 107
Spatial Reasoning, Special Function Evaluation: 75, 85, 137, 150

—CONTINUOUS REASONING—

Simulation: 94

These categories and subcategories are those suggested by the author. There exists only one simulation problem in the list, Problem 94, yet this one problem represents an extremely important class of problem, a problem whose solution and foundations are particularly difficult to store, model, and demonstrate on computer.

At an intermediate stage of categorization are the Spatial Reasoning problems. Although such problems are single-step problems like the Non-Spatial Reasoning (Tumbug-SOAV) problems, they involve spatial reasoning that makes them distinctly different in character than the Non-Spatial Reasoning problems. Each of these Spatial Reasoning problems is typically novel and tends to require an original solution from scratch, based on the specifics of the problem. The author developed algorithms for several of these otherwise Non-Algorithmic problems since they have common themes whose essence can be captured a simple, compact way, but in real life such algorithmic solutions are probably rare. One attribute that makes the Spatial Reasoning problems unique from Non-Spatial Reasoning problems is that the algorithmic solutions do not generalize in an obvious way as the Tumbug-SOAV problems did. Discovery of a general algorithm for solution of Spatial Reasoning problems would likely be a tremendous boost to CSR theory. Regarding overall percentage of problem types in WS150, 123/150 = 82.0% of the WS150 problems appear solvable by the one-step visual algorithm Tumbug-SOAV, and 27/150 = 18.0% of the WS150 problems require solution by some other method.

A few problems have the primary characteristic that they force the examinee to perform one of a set of Special Function Evaluations as the primary focus of the problem, namely ImageMatchingEvaluation for Problems 75 and 137, and TextComplexityEvaluation for Problems 85 and 150. It is possible that one day it may be decided that the Special Function Evaluation problems are not sufficiently indicative of CSR that they should be included as WS problems. One article from competitors in the Winograd Schema Competition (WSC) mentions that their approach did not work on the Special Function Evaluation problems, e.g., Problem 75 ([5], p. 1034), and another article mentioned their approach did not work on some of the Spatial Reasoning problems, e.g. Problem 34 [6], all of which corroborates the author's claim that the recommended categorization catches some fundamental differences in problem types.

3 THE PROPOSED APPROACH

3.1 Placement Of This Approach Within The Hierarchy Of CSR Methods

The proposed approach to CSR is knowledge-based (as opposed to web mining or crowd sourcing), but it does not fall into any of the three traditional categories of knowledge-based CSR (viz., mathematical, informal, or large-scale) ([7], p. 11), therefore the author created a new category called "visual" for the proposed approach, intended to compete with the mathematical category. The current working name for the proposed approach is "Tumbug-SOAV."

3.2 The Algorithm

The algorithm is that for each problem in WS150 to be solved:

1. (preliminary) Concepedia = the collection of heuristic CSR rules in the concepedia, stored in Tumbug-SOAV form

2. (preliminary) GivensAndQuestion = the givens and the question of the current Problem, stored in Tumbug-SOAV form

3. Attempt to match GivensAndQuestion against Concepedia.

4. IF a match is found THEN

    (a) Report match success.
    (b) MatchingRule = the rule in Concepedia where the match occurred
    (c) InstantiationDiagram = MatchingRule merged with GivensAndQuestion
    (d) Read the components off InstantiationDiagram that the question requests, via the diagrammed question's label "ID = ?".

5. ELSE if no match is found THEN

    (a) Report match failure.

6. (if the concepedia is designed to learn:) Store GivensAndQuestion in Concepedia.

3.3 Solutions Of Specific Examples From The WS Problems

3.3.1 A System-Object (SO) Problem

Consider the following problem from WS150: *(Problem 3) Joan made sure to thank Susan for all the help she had [given/received]. Who had [given/received] help? Possible answers: {Susan, Joan}.* If these two sentences (one sentence is the question) are converted into Tumbug then Figure 1 results. In this document the convention is to choose the first word of the pair, in this case "given" instead of "received." Thin arrows represent information flow. Thick rounded arrows represent physical contact. Dashed circles represent pronouns or ambiguity in the physical-informational nature of an object (such as of "help"). Since the text of the example uses the ambiguous
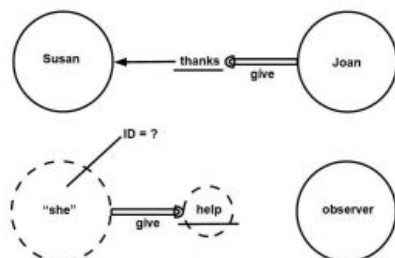
Figure 1: Problem 3's givens and question.

Figure 2: Problem 3's concepedia rule.

Figure 3: Problem 3's matching process.

pronoun "she," *pro tem* the label of that circle is "she" in the first figure, and the dashed border of the circle there represents a pronoun instead of a noun, per Tumbug convention. In Tumbug a nearly 45-degree line is typically used as a flag on the diagram component to be identified, accompanied by the label "ID = ?". Floating horizontal lines underneath a label represent information content, which is non-physical.

Note that the pair of diagrams above cannot provide a reasonable way of arriving at an answer. This is where CSR comes into play. CSR provides default behavior and/or default values based on commonly observed actions in the real world of which the above diagrams are unaware. CSR is similar to possessing a video of an event whereas the diagrams above are like a single snapshot taken at one point during the event. In the real world, to know which actions would typically precede the snapshot and which would typically ensue after the snapshot, one would conceptually match the snapshot against the video to find where the snapshot and video match, then rely on the remainder of the video to give additional information, which can then be used to identify or annotate objects in the snapshot.

This example happens to rely on a common temporal pattern of social conduct, namely that if someone offers free help then an observer of the helper's action should give thanks to the helper. In general any set of rules would suffice: rules of a game, legal requirements, steps in a mechanical process, and so on. In the above example clearly the social rules consist of two time phases: (1) the helper gives help (while observer watches), (2) the observer gives thanks to the helper. Often the helper is directly helping the observer, in which case the observer is socially obligated to return thanks, though the scenario depicted below is slightly more general since it leaves open the possibility that a non-recipient is being helped. The commonsense expectations of this social rule are diagrammed in Figure 2.

In this example of a common event there are two phases. In general any number of phases may be involved, though typically the count in WS150 is around two through four. Counts can easily vary by opinion about which demar-
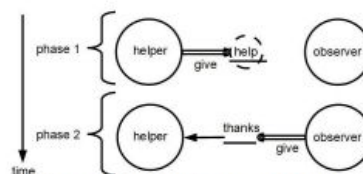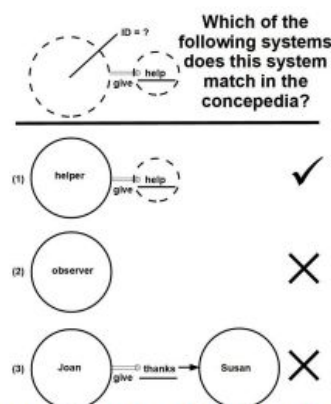
cation points in an event are important. Also note that no attributes of any objects were involved, only the objects/people and the transfers. In general an example may involve any number of attributes (i.e., adjectives) as well as any number of phases. Places can also have attributes, although place icons differ from people icons and object icons. Even the actions themselves may have any number of attributes (i.e., adverbs).

At this point the two systems are matched, against the givens-and-question against the concepedia, especially the system with the unknown ID, and systems are merged if they match. The objects (circles) of the diagrams must match, as well as the arrows and arrow directions. The uncertain object–the dashed circle with the "ID = ?" label attached–has its system matched against all (three) other systems of the CSR rule. It can be seen at a glance that the only system that matches is (1), and further examination shows that all existing labels in the key structure match (1), as well. The merged diagram has the "ID = ?" point to the circle labeled "Susan", which answers the question in an obvious way. Note that common sense knowledge overrides a "don't know" value during CSR matching.

Note that the algorithm as described does not specify procedural details because such details may be irrelevant if the plotted structure is represented as an image since all parts of the image could be matched in parallel: the circle matches, the arrow direction matches, and the label

Figure 4: Problem 3's instantiation. ID = Susan
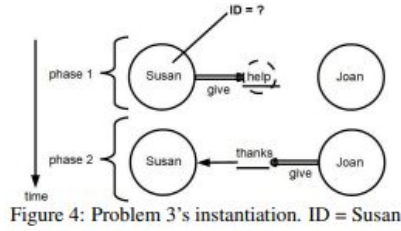
matches. For this reason the author uses the term "visual algorithm." Note that a visual algorithm, though weakly defined here, does not involve any manipulation of text or symbols as formal logic would. Such an approach to CSR is radically different from most or all earlier approaches. A more procedural, textual approach would be relatively straightforward to code, if desired, such as representing the Tumbug diagrams as annotated linked lists, and then using standard graph matching algorithms.

A Tumbug diagram showing phases can be thought of as a strip of film with each phase in its own frame, where the frames are stacked and aligned so that identical features are likely to be aligned in 3D space during the matching process. The matching process then completes the analogy mentioned above, whereby a single frame of the film is located within the film by using parallel search methods that examine a given relative region of space across all film frames simultaneously, whereupon the film begins playing at the match point.

3.3.2 An Object-Attribute-Value (OAV) Problem

The above algorithm applies even if the diagrams contain attribute values. The next example from WS150 demonstrates this. *(Problem 21) I was trying to balance the bottle upside down on the table, but I couldn't do it because it was so [top-heavy/uneven]. What was [top-heavy/uneven]? Possible answers: {the bottle, the table}.* Using a variation of conventional OAV notation from OOP, except with attributes and values shown outside of the rectangle or circle that represents the object, with those objects placed into Tumbug diagram format, produces Figure 5. In Tumbug a thick, sharp arrow represents physical motion of an object. There exists only one phase in this problem, therefore phases need not be labeled. The prepositional verb "to try to" typically points to an "aggregation box," as it does here. Inside the aggregation box shows what is being attempted, which in this case is balancing a bottle. Since "balanced" is a state more than a permanent attribute, a two-axis state icon is used, labeled "balanced" and with the upside-down bottle's label inside. The rectangular box with the open top indicates literal spatial arrangements of any objects that have described locations, even if only relative locations. In this case the bottle object is atop the table object, and ex-
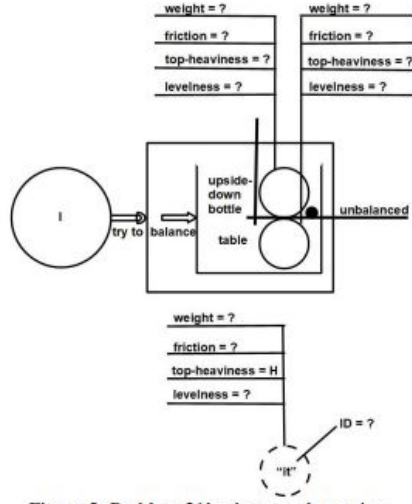

Figure 5: Problem 21's givens and question.

actly touching at the expected contact point for such a situation.

This example happens to rely on a common pattern of physics. The commonsense expectations of this rule are derived such that if an average person were asked for the attributes of an object that would tend to ensure that the object would be unbalanced, some variation of the following set would presumably be cited: low weight, low friction against the surface, high top-heaviness, low-to-medium levelness of the surface. Using L, M, H, VH, 0 for low, medium, high, very high, none, respectively, the above expectations could be assigned to any object that is represented as OAV triples. Applying this scheme to both mentioned objects produces one list of values for each object, where each object's attributes are [weight, friction, top-heaviness, levelness]...

upside-down bottle: [L, L, H, M]

table: [H, L, M, H]

...which produces Figure 6. The same matching process applies as before, but across objects instead of phases, using single objects instead of structures. This time not only must any circles, arrows, arrow directions, and any labels match, but also the attributes and attribute values of all the involved objects. Since the only object that matches perfectly is the upside-down bottle with attribute-value "top-heavy = H", the merged diagram have the "ID = ?" point to the circle labeled "upside-down bottle", which answers the question in an obvious way. Note that in this context any attribute (but not label) "?" ("don't know") or "-" ("don't
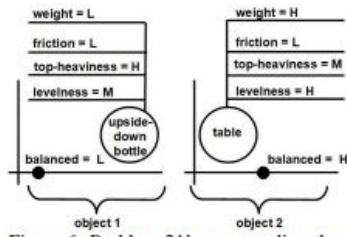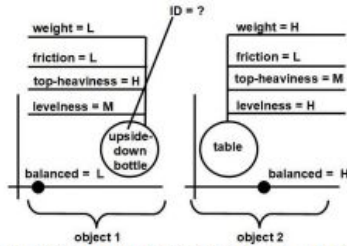
Figure 8: Problem 21's instantiation. ID = upside-down bottle

care" or "does not apply") attribute values automatically mismatch any known, specified, numerical value. The two-axis graph is automatically included in the concepedia as part of each object to capture the relationship between the object's attribute values and various states, in this case the state of being balanced. This is a mathematical relationship that can often be approximated by a continuous function. Discussion of such functions within Tumbug-SOAV is out of scope for this article.

3.3.3 A Timeline Problem

(Problem 78) Thomson visited Cooper's grave in 1765. At that date he had been [dead/travelling] for five years. Who had been [dead/travelling] for five years? Possible answers: {Cooper, Thomson}. This demonstrates a many-to-many matching situation, since only pieces of the givens-and-question diagram will match pieces of the single applicable rule in the concepedia. The single concepedia rule in Figure 10 contains two pieces of heuristic commonsense knowledge in visual form: (1) A person who is in a grave is dead, (2) A person who is in a grave and dead will always be in a grave and dead. Otherwise this problem is similar to the previous problems, and should be self-explanatory with Figure 9, Figure 10, Figure 11, Figure 12.

3.3.4 Overview Of Solutions To A Few Unusual Variations of Tumbug-SOAV Problems

WS150's Problem 6 requires relative speeds to be represented. This can be done either by using the length of object motion arrows to represent speed, as wind speed is commonly represented by vectors, or by using the attribute "speed" on the motion arrow.

Several problems such as Problem 33 require a timeline for visualizing the described situation. In others such as Problem 80 a timeline is useful but optional. One very simple version of a timeline that the author calls a "2-Point Status Timeline" is used by problems such as Problem 78. Another very simple version of a timeline is the "Influence Timeline," used in Problem 79. Event duration on such timelines can be regarded as a movable object (viz, a rigid
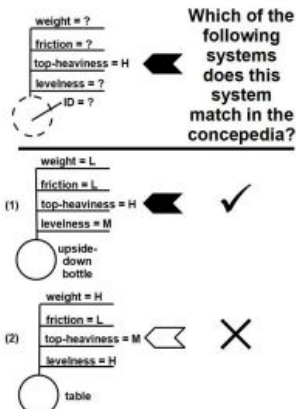

Figure 6: Problem 21's concepedia rules.

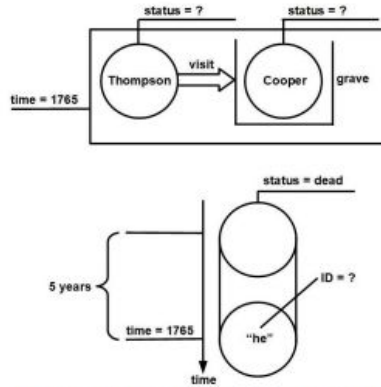Figure 7: Problem 21's matching process.

6

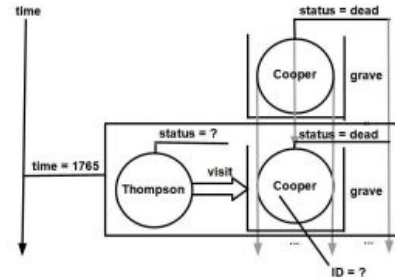Figure 9: Problem 78's givens and question.

Figure 12: Problem 78's instantiation. ID = Cooper
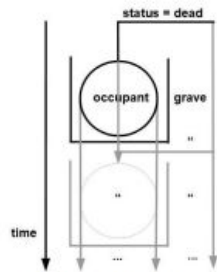

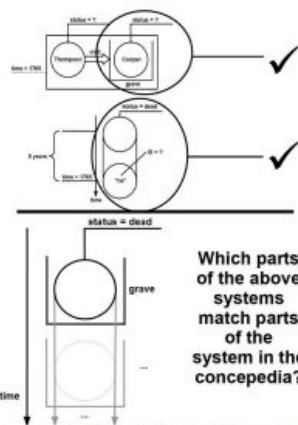Figure 10: Problem 78's concepedia rule
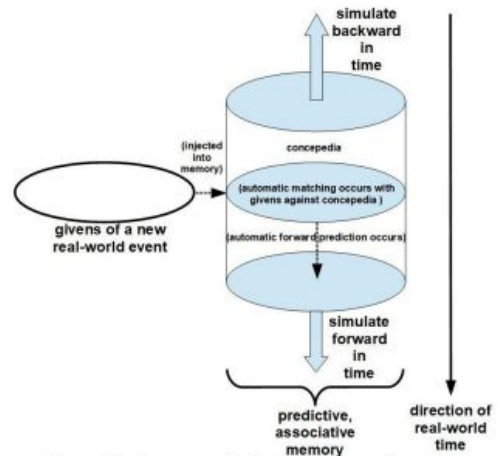

Figure 11: Problem 78's matching process.

Figure 13: A concepedia is essentially a simulator.

7

rectangle where one dimension is time), which for example simplifies the visualization of the amount of leaving late being exactly in sync to the amount of arriving late because the travel time is fixed.

Some Broad Sense problems such as Problem 26 require returning two identifications, some such as Problem 117 require identification of an object via a possessive adjective instead of via pronoun, and others such as Problem 141 use a noun (e.g., "fool") instead of an ambiguous pronoun (e.g., "he"). To Tumbug-SOAV's credit, such variations cause almost no extra problems, except the possessive adjective type of Broad Sense problems currently require existence of several unique types of arrows that the author collectively calls "genitive arrows."

Some problems such as Problem 46 require alternative ensuing scenarios to be included in concepedia rules. These require only a slight addition to the matching algorithm: the algorithm must consider every path in a tree as it would normally consider every listed phase of a common event's time evolution.

Some problems such as Problem 141 use or even require OOP style inheritance. Tumbug simply carries the inherited class information along with each object, with the inherited class considered a superset represented by a circumscribed circle. Some problems such as Problem 130 require recognizing illogical or impossible graph structures, such as a person carrying another person in the carried person's arms. These presumably require only simple, generic checks of the graph during the matching process.

4 CONCEPEDIAS

The heuristic CSR rules used by Tumbug-SOAV are fairly conventional, at least when written as text. However, some characteristics of Tumbug-SOAV rules that differ from typical production system rules are: (1) Ultimately all knowledge and information is stored in essentially visual/pictorial format. (2) The rules are often temporal, which means a temporal rule's components are ordered by increasing phase number that requires those components to be accessed in temporal order. (3) Rules often contain branches of possibilities. (Likelihood values may be assigned to each outcome, if desired.) (4) Knowledge about specific objects is much more extensive than is found in a dictionary, especially since it often describes temporal knowledge, relationships between objects, behavioral tendencies, motion characteristics, and very extensive values of unusual attributes such as texture, friction, price, and emotional value. (5) Abstract concepts are often stored as well as "corporeal" (i.e., non-abstract) objects.

The proposed solution for Tumbug-SOAV is to store such knowledge and information in a file somewhat like an encyclopedia in that each heading contains a great deal of

knowledge about the given object, and that the objects are stored as concepts instead of as corporeal (i.e., instantiated) objects. Also, the knowledge should be coded so as to ensure understandability for both human and machine, for reasons of interface manageability and machine efficiency. Since ultimately the proposed approach must be visual, this creates some extra challenges for both software and hardware, though discussion of these challenges is out of scope of this article. Such an encyclopeda is what the author defines as a "concepedia," which is a roughly an encyclopedia of concepts.

More generally, a concepedia can be thought of as the part of an intelligent machine where understanding occurs, based on only the few sparse clues provided by natural language input to the concepedia. The concepedia's vast stored knowledge fills in the many gaps left by the ambiguity of the words and phrases of natural language (cf. [8]). A concepedia can be considered a combination simulator, disambiguator, error corrector, and compiler as suggested in Figure 13.

5 ASSESSMENT OF THIS APPROACH

Kocijan et al. [3] assessed different approaches used so far to solve WSC problems, and mentioned the following five desiderata. Each desideratum below is listed with a comment about how Tumbug-SOAV is believed to obviate the mentioned pitfall. Page numbers refer to the aforementioned reference.

1. **The inference should be carried out automatically (p. 4).** Tumbug-SOAV excels at this because nearly all WS examples are solved via Tumbug-SOAV primarily with a single image-matching operation that is automatically triggered upon inputting a real-world conjecture into memory. Therefore Tumbug-SOAV does not require "commonsense reasoning of some depth and complexity" (p. 4). Tumbug is a novel implementation of a "reflexive reasoning" architecture (cf. [9]).

2. **The approach should not be easily solvable using word correlation (p.16).** The Tumbug-SOAV approach completely ignores the statistics of word frequency or word correlation because the visual search algorithm simply matches text verbatim, so obviously Tumbug-SOAV bypasses this pitfall.

3. **The approach should demonstrate some semblance of CSR, not merely technical tricks (pp. 7-8).** Because Tumbug-SOAV uses a visual grammar that the author believes is universal across all natural languages, and because Tumbug-SOAV deals with the actions of real-world objects, at the very least each scenario that Tumbug-SOAV visualizes could be run as a video simulation, which would confirm that the

correct corporeal actions on the correct corporeal objects are being used.

4. **Extraction of relevant information from the sentence should not be a bottleneck (p. 23).** This is unknown since Tumbug-SOAV has not been tested because Tumbug-SOAV has not been coded yet.

5. **The system should be able to generalize (p. 26).** This was mentioned in Problem 141 above, where "Joe" was considered an element of "fools" through inheritance, so the entire set "fools" would be moved along as the object "Joe" is moved around.

6 SUMMARY

The suggested Tumbug-SOAV approach to the WS is promising for many reasons, such as: (1) its single visual matching algorithm should already be able to solve 82.0% of WS150 problems, and current indications are that this percentage might be pushed up to 96.6% with a single generalization of the algorithm, (2) it elegantly meets four of the five criteria used to judge the quality of a WS solution [3], (3) Tumbug has already proved useful for foreign language learning [4], (4) its KRM opens up a new subfield of machine learning, one based on image-like data structures instead of numbers, (5) it is solidly based, partly on existing knowledge from OOP, (6) its functions mirror existing conventions of mathematical functions (e.g., piecewise-defined functions versus formulas, generic formulas versus explicit formulas), (7) it shows the relationship between various grammatical concepts (e.g., modal verbs versus propositional verbs, active tense versus passive tense, importance of the genitive case) in a clear-cut and visual manner that is rarely if ever mentioned elsewhere, (8) it forces a very logical categorization of WS problems that explains some of the non-generalization impediments that WSC competitors have been experiencing, (9) it is a new model and explanation for reflexive reasoning [9], (10) it handles Broad Case WS problems with ease, (11) it unifies two different types of human learning (viz., explicit and implicit), (12) it is a completely new and unexplored knowledge-based *type* of approach to CSR that opens up new ideas in the field of CSR and therefore in AGI, (13) it conforms to the popular modern belief that memory recall is more important than processing speed in biological brains (e.g., [8], pp. 67-69), (14) because images can hold multiple attributes at once, it can potentially perform matching in parallel, which should be much faster than following long chains of logical inferences.

7 FUTURE WORK

The author's main practical concern regarding the Tumbug-SOAV algorithm is that it has not been coded yet, there-fore empirical results are lacking. Coding such a simulation is planned, though the inherently visual nature of the algorithm and stored information will require either extensive use of image recognition algorithms or discretized workarounds. The main theoretical concerns of Tumbug-SOAV are the following:

1. How to generalize the algorithm to more types of WS150 problems, especially to the Spatial Reasoning type. In recent work the author found that two of the nine Spatial Reasoning Non-Algorithmic problems appear to be solvable by a more generalized version of Tumbug-SOAV. If this observation generalizes then it suggests that a slightly more general algorithm may be able to solve all 22 Spatial Reasoning problems, which would bring the number of solvable WS150 problems by a single algorithm up to 145/150 = 96.6%, which would be extremely good since human performance on the WSC is 92%-96%. The general strategy and hope here is that continued, step-wise generalizations of the Tumbug-SOAV algorithm will eventually lead to a single algorithm that can solve 100% of the WS150 problems, whereby a deep understanding of CSR will likely have resulted.

2. How to get real-world knowledge into a concepedia by either learning from real-world observations or by generalizing what is already in the concepedia. The author believes that both of these approaches are viable.

3. Some parser will eventually need to convert an arbitrary natural language sentence accurately and consistently into Tumbug and SOAV. This accomplishment alone would automatically solve many problems of sentence understanding, though the author has no plans to work on this particular interface problem.

References

[1] D. Ernest, *Representations of Commonsense Knowledge*. San Mateo, California: Morgan Kaufmann Publishers, 1990.

[2] E. Davis, "Collection of winograd schemas," 2018.

[3] V. Kocijan, E. Davis, T. Lukasiewicz, G. Marcus, and L. Morgenstern, "The defeat of the winograd schema challenge," *arXiv preprint arXiv:2201.02387*.

[4] M. A. Atkins, *Important things to know before learning any foreign language*. Unpublished, 2022.

[5] A. Sharma, "Using answer set programming for commonsense reasoning in the winograd schema challenge," *Theory and Practice of Logic Programming*, vol. 19, no. 5–6, pp. 1021–1037, 2019.

[6] A. Sharma, N. H. Vo, S. Gaur, and C. Baral, "An approach to solve winograd schema challenge using automatically extracted commonsense knowledge," in *2015 AAAI Spring Symposium Series*, 2015.

[7] E. Davis and G. Marcus, "Commonsense reasoning and commonsense knowledge in artificial intelligence," *Communications of the ACM*, 2015.

[8] J. Hawkins and S. Blakeslee, *On intelligence*. Macmillan, 2004.

[9] L. Shastri and V. Ajjanagadde, "From simple associations to systemic reasoning: A connectionist representation of rules, variables and dynamic bindings," January 1990.

2 arXiv reception

A promising visual approach to solution of 82% of Winograd Schema problems via Tumbug Visual Grammar

Mark A. Atkins

Abstract

A new type of approach is presented for solving common-sense reasoning problems from the Winograd Schema. The proposed approach uses three novel knowledge representation methods, all visual and all integrated: (1) Tumbug, which a universal, visual grammar that is a type of labeled, directed graph, (2) SOAV, which is a system level extension to OAV, where OAV is the familiar Object-Attribute-Value triple of object-oriented programming, and (3) a concepedia, which is a computer-readable visual encyclopedia of concepts that effectively contains commonsense reasoning rules. The approach should apply to 82% of the Winograd Schema problems, though the approach has not yet been computer simulated to confirm this claim. The approach consists of using Tumbug-SOAV notation for both the problem and the concepedia, then of applying one image matching function per Winograd Schema problem, which makes the approach extremely fast, automatic, and reflexive. This direct approach to commonsense reasoning avoids indirect, short-term, technical tricks.

1 INTRODUCTION

Real progress in the field of artificial intelligence (AI) depends on the critical but currently underdeveloped, currently unapplied subfield called artificial general intelligence (AGI). In the quest to produce AGI one solid approach is to tackle one of the well-known, longstanding subfields of AGI that are believed to be key. One such subfield is commonsense reasoning (CSR), which currently all computers and their programs still do very poorly ([1], p.1). Within CSR there exists a list of benchmark problems called the Winograd Schema (WS), which is a very convenient testbed for any new AGI approach, which this article's proposed approach is. For brevity the set of 150 problems in the version of the WS dataset (from [2]) that was used in this study is called "WS150" in this document (cf. [3], p. 6). A small percentage of these problems are called "Broad Sense" because they deviate from Terry Winograd's intended required problem characteristics [2].

The proposed approach is based on a striking discovery by the author in 2022 is that there exists a single visual matching algorithm that can solve the vast majority of WS150 problems. The knowledge representation method (KRM) used is called "Tumbug," a convenient written form of "TUMBVG," or "Temporal Universal Model-Based Visual Grammar" [4], developed by the author. The Tumbug KRM is combined with the SOAV KRM to produce the Tumbug-SOAV KRM, where "SOAV" means "System-OAV," where "OAV" (Object-Attribute-Value) is a term used in object-oriented programming (OOP). SOAV is the author's higher-level extension of OAV that considers "systems" of objects, connected by arrows and other icons to represent relationships between the objects. Tumbug Visual Grammar differs from Tumbug Textual Grammar, the latter of which lacks diagrams in the usual sense.

2 RECOMMENDED CATEGORIZATION OF WS150 PROBLEMS

This section lists all WS150 problems within categories relative to the proposed Tumbug-SOAV approach. So far, only the Non-Spatial Reasoning problems can be consistently solved by the Tumbug-SOAV algorithm.

—DISCRETE REASONING—

Non-Spatial Reasoning: 1, 3, 4, 5, 6, 7, 8, 9, 12, 13, 15, 16, 17, 18, 19, 20, 21, 23, 24, 25, 26, 27, 28, 29, 32, 33, 35, 36, 37, 38, 39, 41, 42, 43, 44, 46, 47, 48, 49, 50, 51, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 67, 68, 69, 70, 71, 73, 76, 78, 79, 80, 81, 82, 83, 86, 87, 88, 89, 90, 91, 92, 93, 95, 96, 97, 98, 99, 100, 102, 104, 105, 106, 108, 109, 110, 111, 112, 113, 114, 115, 116, 117, 118, 119, 121, 122, 123, 124, 125, 126, 127, 128, 129, 130, 131, 132, 133, 134, 135, 136, 138, 139, 140, 141, 142, 143, 144, 145, 146, 147, 148, 149
Spatial Reasoning, Algorithmic: 2, 10, 22, 34, 52, 53, 66, 72, 77, 84, 101, 103, 120
Spatial Reasoning, Non-Algorithmic: 11, 14, 30, 31, 40, 45, 54, 74, 107
Spatial Reasoning, Special Function Evaluation: 75, 85, 137, 150

—CONTINUOUS REASONING—

Simulation: 94

Figure 1: October 3, 2022: Article submitted to arXiv, date stamped.

Figure 2: October 5, 2022: arXiv placed the article on hold.

MOD-18861 arXiv - important notification regarding submit/4525082

AS arXiv Support <jira@arxiv-org.atlassian.net>
To: You

Thu 10/27/2022 10:32 AM

—————

Reply above this line.

Dear author,

Thank you for submitting your work to arXiv. We regret to inform you that arXiv's moderators have determined that your submission will not be accepted and made public on[|http://arxiv.org] [arXiv.org|http://arxiv.org].

Our moderators determined that your submission does not contain sufficient original or substantive scholarly research and is not of interest to arXiv.

For more information on moderation policies and procedures, please see: [1]https://arxiv.org/help/moderation.

arXiv moderators strive to balance fair assessment with decision speed. We understand that this decision may be disappointing, and we apologize that, due to the high volume of submissions arXiv receives, we cannot offer more detailed feedback. Some authors have found that asking their personal network of colleagues or submitting to a conventional journal for peer review are alternative avenues to obtain feedback.

We appreciate your interest in arXiv and wish you the best.

Sincerely,

The arXiv Content Management & User Support Team
----
[1] https://arxiv.org/help/moderation

Figure 3: October 27, 2022: arXiv rejected the article.

3 AISTATS 2023 reception

Questions

1. Summary and Contributions: Briefly summarize the paper and its contributions
This paper aims to solve the Winograd schema challenge, a test of machine intelligence. The proposed visual approach wasn't properly discussed though.

2. Strengths: Please describe the strengths of the work according (but not limited) to the following criteria: soundness of the claims (theoretical grounding, empirical evaluation), significance and novelty of the contribution, and relevance to the AISTATS community.
Nothing stands out.

3. Weaknesses: Please describe the limitations of this work according (but not limited) to the following criteria: soundness of the claims (theoretical grounding, empirical evaluation), significance and novelty of the contribution, and relevance to the AISTATS community.
- The problem setting should be better introduced, for readers not familiar with this topic.
- There is no related work discussing other possible approaches, and the authors haven't compare their approach with existing ones.
- I am not convinced the results of the proposed approach is reproducible, as the discussion is hard to understand at times.

4. Correctness: Are the method and claims correct? Is the empirical methodology correct?
I am not sure about the correctness of the proposed approach, as its empirical tests were poorly done. There is no table/figure summarizing the quantitively results.

5. Clarity: Is the paper well written? Does it clearly state its contributions, notation and results?
The paper needs major revision as it is hard to follow at times.

6. Relation to prior work: Is it clearly discussed how this work differs from or relates to prior work in the literature?
No.

7. Additional Comments: Add your additional comments, feedback and suggestions for improvement, as well as any further questions for the authors.
I don't think the paper is ready for publishing.

8. Reproducibility: Are there enough details to reproduce the major results of this work?
No, the results are not reproducible in the current form of the paper

9. Assumptions and limitations: Does the paper explicitly and clearly state the main assumptions and limitations of the work?
Only some assumptions and limitations are discussed, or they are not clear enough

10. Societal impact: Does the paper discuss the societal impact of the work, including the impact that may arise from the misuse of the paper's contribution?
No, and it does not seem necessary to include discussion on potential societal impact

11. Code release: Do the authors promise to release code for this submission?
No

12. Score: Please provide an overall score for the submission
2 - Strong reject (I would be very upset if accepted)

13. Confidence score: Please provide a confidence score for your assessment of this
4 - You are confident in your assessment but not absolutely certain.

14. Ethical Concerns: Does this submission raise potential ethical concerns? These include methods, applications or data that create or reinforce unfair biases and/or that have a primary purpose of harm or injury.
No

16. Code of conduct: While performing my duties as a reviewer (including writing reviews and participating in discussions), I have and will continue to abide by the AISTATS 2023 code of conduct, available at https://aistats.org/aistats2023/code-of-conduct.html
Agreement accepted

17. Confidentiality: I agree to keep the paper and supplementary materials (including code submissions and LaTeX source) as well as the reviews confidential. I also agree to delete any submitted code at the end of the review cycle to comply with confidentiality requirements.
Agreement accepted

Figure 4: AISTATS reviewer #1 assessment.

5. Check that the content of your submission, *excluding* references, is limited to **8 pages**. The number of pages containing references alone is not limited.

Figure 5: Rebuttal to reviewer #1. The introductory article has *346 pages*.

Questions

1. Summary and Contributions: Briefly summarize the paper and its contributions

The problem addressed in this paper is commonsense reasoning from the Wnograd Schema. The authors propose an approach that make use of three visual and integrated knowledge representation methods, including a universal, visual grammar that is a type of labeled, directed graph called Tumbug; System-level Object-Attribute-Value (SOAV), extending the standard Object-Attribute-Value triple of object-oriented programming at a system level; and a computer-readable visual encyclopedia of concepts that contains commonsense reasoning rules.

2. Strengths: Please describe the strengths of the work according (but not limited) to the following criteria: soundness of the claims (theoretical grounding, empirical evaluation), significance and novelty of the contribution, and relevance to the AISTATS community.

At the first glance this paper is interesting to read and following the authors' reasoning is easy to follow.

3. Weaknesses: Please describe the limitations of this work according (but not limited) to the following criteria: soundness of the claims (theoretical grounding, empirical evaluation), significance and novelty of the contribution, and relevance to the AISTATS community.

My main object with the paper is that it seems to me like a philosophical paper rather than an AI research paper, focusing solely on a theoretical aspect, and lacking a provable or empirical evaluation.

4. Correctness: Are the method and claims correct? Is the empirical methodology correct?

The claims in the paper is not verified through empirical evaluations nor does it provide any theoretical proves of correctness.

5. Clarity: Is the paper well written? Does it clearly state its contributions, notation and results?

The paper is written clearly.

6. Relation to prior work: Is it clearly discussed how this work differs from or relates to prior work in the literature?

The authors do a good job explaining how their work builds on existing approaches.

7. Additional Comments: Add your additional comments, feedback and suggestions for improvement, as well as any further questions for the authors.

I am very curious to how the proposed approach could be evaluated through, for example experimental evaluation. This could of strengthen the paper and would undoubtedly increase its relevance for the AI and ML community, as well as improving the content of the paper.

8. Reproducibility: Are there enough details to reproduce the major results of this work?

Not applicable: Paper of theoretical nature

9. Assumptions and limitations: Does the paper explicitly and clearly state the main assumptions and limitations of the work?

No, the assumption and limitations are not discussed

10. Societal impact: Does the paper discuss the societal impact of the work, including the impact that may arise from the misuse of the paper's contribution?

No, but it should be discussed due to the nature of the paper

11. Code release: Do the authors promise to release code for this submission?

No

12. Score: Please provide an overall score for the submission

3 - Clear reject (I vote and argue for rejection)

13. Confidence score: Please provide a confidence score for your assessment of this

4 - You are confident in your assessment but not absolutely certain.

14. Ethical Concerns: Does this submission raise potential ethical concerns? These include methods, applications or data that create or reinforce unfair biases and/or that have a primary purpose of harm or injury.

No

16. Code of conduct: While performing my duties as a reviewer (including writing reviews and participating in discussions), I have and will continue to abide by the AISTATS 2023 code of conduct, available at https://aistats.org/aistats2023/code-of-conduct.html

Agreement accepted

17. Confidentiality: I agree to keep the paper and supplementary materials (including code submissions and LaTeX source) as well as the reviews confidential. I also agree to delete any submitted code at the end of the review cycle to comply with confidentiality requirements.

Agreement accepted

Figure 6: AISTATS reviewer #2 assessment.

16

Figure 7: Rebuttal to reviewer #2: AISTATS 2023 guidelines for reviewers.

Figure 8: AISTATS reviewer #5 assessment. Request too vast for 8 pages.

Questions

1. Summary and Contributions: Briefly summarize the paper and its contributions
A promising visual approach to solution of 82% of Winograd Schema problems
via Tumbug Visual Grammar

In this paper the author suggests using a "Tumbug Visual Grammar" to solve Winograd Schema problems. They suggest that this allows for commonsense reasoning and will lead to artificial general intelligence. They show anecdotally how certain problems are solved by this system. They suggest the approach is automatic, does not use correlation, not "merely technical tricks", not "just neural networks" and able to generalize.

2. Strengths: Please describe the strengths of the work according (but not limited) to the following criteria: soundness of the claims (theoretical grounding, empirical evaluation), significance and novelty of the contribution, and relevance to the AISTATS community.
The paper is clear and easy to follow. As well, being able to solve Winograd schemas is an important and nontrivial problem.

3. Weaknesses: Please describe the limitations of this work according (but not limited) to the following criteria: soundness of the claims (theoretical grounding, empirical evaluation), significance and novelty of the contribution, and relevance to the AISTATS community.
Nonetheless, I am not convinced by this approach, aand do not believe it has any of the qualities suggested above. In particular, the approach is not "automatic" because it requires a complete and well specified grammar. It has no capability to learn and no capability to refine errors which must arise during specification or learning. This also precludes any meaningful generalization. While the approach does not use correlation, it does not need to since everything either matches or does not---correlations of 1 or 0! It is not clear what technical tricks are referred to, but in the broad sense, in the absence of a ghost in the machine, one may argue all we have (and are) are "technical tricks." Finally, while the approach does not use a neural network, it seems clear that a neural network could easily perform the kind of pattern matching done by this approach *if the grammar was given* as assumed here.

4. Correctness: Are the method and claims correct? Is the empirical methodology correct?
see above

5. Clarity: Is the paper well written? Does it clearly state its contributions, notation and results?
see above

6. Relation to prior work: Is it clearly discussed how this work differs from or relates to prior work in the literature?
see above

7. Additional Comments: Add your additional comments, feedback and suggestions for improvement, as well as any further questions for the authors.
As a result, although the author strives toward a worthy goal, I cannot recommend publication of this work as given.

8. Reproducibility: Are there enough details to reproduce the major results of this work?
No, the results are not reproducible in the current form of the paper

9. Assumptions and limitations: Does the paper explicitly and clearly state the main assumptions and limitations of the work?
Only some assumptions and limitations are discussed, or they are not clear enough

10. Societal impact: Does the paper discuss the societal impact of the work, including the impact that may arise from the misuse of the paper's contribution?
No, and it does not seem necessary to include discussion on potential societal impact

11. Code release: Do the authors promise to release code for this submission?
No

12. Score: Please provide an overall score for the submission
3 - Clear reject (I vote and argue for rejection)

13. Confidence score: Please provide a confidence score for your assessment of this
3 - You are fairly confident in your assessment.

14. Ethical Concerns: Does this submission raise potential ethical concerns? These include methods, applications or data that create or reinforce unfair biases and/or that have a primary purpose of harm or injury.
No

16. Code of conduct: While performing my duties as a reviewer (including writing reviews and participating in discussions), I have and will continue to abide by the AISTATS 2023 code of conduct, available at https://aistats.org/aistats2023/code-of-conduct.html
Agreement accepted

17. Confidentiality: I agree to keep the paper and supplementary materials (including code submissions and LaTeX source) as well as the reviews confidential. I also agree to delete any submitted code at the end of the review cycle to comply with confidentiality requirements.
Agreement accepted

Figure 9: AISTATS reviewer #6 assessment.

19

[AISTATS 2023] [Authors] Rejection Notification
MC
Microsoft CMT
To:You
Fri 1/20/2023 6:33 AM
Dear Mark A Atkins,

Thank you for submitting your work to AISTATS 2023. We regret to inform you that your submission #117 ("A promising visual approach to solution of 82% of Winograd Schema problems by using Tumbug Visual Grammar, System-Object-Attribute-Value diagrams, and a concept encyclopedia") has not been accepted to the conference.

We received over 2000 abstract submissions this year. Of the 1689 submissions that proceeded to review, 29% were accepted to the conference. This was a competitive year, and unfortunately there were many good submissions that we were not able to accept. Our decision however is final and we will not consider any appeals to overturn the outcome.

We did our best to make the decision process as thorough and fair as possible. 95% of submissions received at least 4 reviews from independent and qualified reviewers, with borderline or unclear cases further checked carefully by additional reviewers, an Area Chair, a Senior Area Chair and ultimately by the Program Chairs. Many factors played a role in the final assessment of a paper: not only numerical scores, but also the quality of author rebuttals and reviewer discussions. In the process, every aspect of the paper, including among others quality of the writing, representation of the state of the art, as well as novelty and potential impact in the AISTATS community, has been considered.

You will be able to check the reviews and meta-reviews on CMT later today (see detailed instructions below). We sincerely hope that the reviews and meta-review can help you improve your paper for a subsequent resubmission.

We thank you for submitting your work to AISTATS 2023, and we hope that you will continue to do so in the future. We also hope that you will still join us at the conference. Registration is already open at https://virtual.aistats.org/Conferences/2023 and the schedule will be announced soon.

Best regards,

Jennifer Dy and Jan-Willem van de Meent
AISTATS 2023 Program Chairs

Christian A. Naesseth and Davin Hill
AISTATS 2023 Workflow Chairs

*** Accessing reviews and meta-review ***
To view your reviews and meta-review:

Figure 10: January 20, 2023: AISTATS rejected the article.

4   Discussion

4.1   Likely underlying reasons for initial rejection

The likely underlying reasons the embedded article was so soundly rejected seem to be the following:

- *Problem #1:* The article was far too short (9 pages, in order to conform to conference regulations) to adequately describe either the Phase 1 foundations of this project (which alone took another article of 346 pages to describe!) or the Phase 2 results, much less both phases. None of the reviewers seemed to detect this underlying problem, and instead merely complained that the article was difficult to understand and that the article needed additional info.
  Solutions employed: (1) I began limiting the breadth of each new article about this project to a single phase. (2) I wrote and posted a new, full-sized article about only Phase 1, and included complete descriptions without regard to length. This new article took 11 1/2 months to write. (3) I began to post all later articles about this project only to sites that have no page limit.

- *Problem #2:* The article was submitted to a conference that emphasizes data science and machine learning, which is a relatively small subset of AI, so the larger issues of AI, CSR, and KRMs discussed in the article were evidently unfamiliar to the reviewers from the machine learning community.
  Solutions employed: (1) I began to post all later articles about this project only to archival sites rather than to conferences whose reviewers lack broad expertise in artificial intelligence.

- *Problem #3:* Many reviewers outright misunderstood what was written, such as making assumptions that were false, and not noticing that requested information was already mentioned.
  Solutions employed: (1) I included special paragraphs throughout the next article (about Phase 1) starting with the italicized phrase "Emphasized clarification:" to address those specific misunderstandings that reviewers made that were based on faulty logic, since these might be misunderstandings among future readers, also.

4.2 Some noticeable changes to Tumbug since the 2022 article

- "?" has been replaced by " " (blank) for most unknown values, partly to save writing effort, and partly to implement the "DON'T CARE" wildcard convention that describes the situation better.

- "?" to represent queries on the ends of slanted lines has been replaced by "DK" (= DON'T KNOW), to avoid misinterpreting the "?" symbol as the "?" wildcard of regular expressions.

- The term "Tumbug-SOAV" has been replaced by the simpler term "Tumbug."

- The term "SOAV" has been replaced by the term "SCOVA," to reflect new awareness of the "C" (= Change) type of component.

- The square cup icon to represent location has been replaced by an Aggregation Box, which is a square and is still a type of Location Box.

- Propositional attitudes (like with the label "try to") are now represented by an Aggregation Box with that textual label above the box, instead of treating a propositional attitude as an action verb.

- States (such as for the concept "balanced") are now represented with State Diagrams instead of graphical plots, even though the new convention can no longer represent fuzzy states.

- Graphical plots would now be contained in an Aggregation Box for purposes of clarity, to prevent graph axes or plotted lines from being confused with Attribute Lines or other types of lines.