Analogical Reasoning: An Algorithm Comparison for Natural Language Processing

Kara Combs
Wright State University, USA
Combs.171@wright.edu

Trevor J. Bihl
Air Force Research Laboratory, USA
Trevor.Bihl.2@us.af.mil

Subhashini Ganapathy
Wright State University, USA
Subhashini.Ganapathy@wright.edu

Drue Staples
Applied Research Solutions, USA
Dstaples@appliedres.com

Abstract

There is a continual push to make Artificial Intelligence (AI) as human-like as possible; however, this is a difficult task. A significant limitation is the inability of AI to learn beyond its current comprehension. Analogical reasoning (AR), whereby learning by analogy occurs, has been proposed as one method to achieve this goal. Current AR models have their roots in symbolist, connectionist, or hybrid approaches which indicate how analogies are evaluated. No current studies have compared psychologically-inspired and natural language processing (NLP)-produced algorithms to one another; this study compares seven AR algorithms from both realms on multiple-choice word-based analogy problems. Assessment is based on selection of the correct answer, "correctness," and their similarity score prediction compared to the "ideal" score, which is defined as the "goodness" metric. Psychologically-based models have an advantage based on our metrics; however, there is not a clear one-size-fits-all algorithm for all AR problems.

1. Introduction

Artificial Intelligence (AI) implies that machines can exhibit human-like reasoning, decision making, and problem-solving [1]. However, a considerable gap exists between AI capabilities and hype [2]. The entertainment industry portrays AI based on its "strong" definition, in which AI can completely mimic human thought processes [1]. In reality, the vast majority of what we consider to be AI is "weak," meaning that it has been programmed with a very specific objective in mind and is incapable of developing other abilities on its own.

Learning is a significant barrier in AI systems and many algorithms are narrow in that they can only analyze classes or groups they have been trained on [3]. Biological intelligent agents have this learning ability, which current AI systems overwhelmingly lack [4]. For AI agents to embody such biological characteristics of intelligence, they need to be able to reason and learn from novel scenarios [5]. One avenue being explored in hopes of advancing a step closer to "strong" AI is analogical reasoning (AR) [4] [6]. Analogies allow information about a familiar situation to be translated and interpreted in the context of a novel scenario [7]. Reasoning by analogy is common in biological intelligence development (as such with children), and is one hypothesis about how humans gain new knowledge [8]. Artificial AR methods have been developed by leveraging concepts from biological intelligence.

One AI method used to solve AR problems is Natural Language Processing (NLP). NLP allows machines to "understand" language as a human would [9]. Within NLP are vector space models (VSMs), which create word embeddings, that allow for geometrical manipulation on variables formerly considered to be nominal [10]. Recently, through these advances with NLP techniques, AR can compute similarity as measured between VSMs [11] [10].

Overall, this paper examines a variety of AR models while providing a broad comparison of performance with discussion of the algorithms' results. While prior comparisons between AR models exist [12] [13] [14] [15], performance on algorithms with psychology inspiration and those without has yet to appear in the literature. The results of this study show how algorithms from these two branches compare on our correctness and analogy "goodness" metrics.

2. Background

NLP, a subset of text mining, aims to allow machines to understand text similar to that of the human brain [9]. NLP focuses on understanding text, meanwhile, it does not always interpret meaning, which is potentially why it struggles with new information. However, by focusing on analogies, AR

URI: https://hdl.handle.net/10125/79493
978-0-9981331-5-7

provides for improvements in current NLP methods by incorporating context for unknown words without having to explicitly train a model on such [16].

Analogy problems take on many forms such as drawing parallels between lengthy stories to sentence-based forms to simple word comparisons [17]. Currently, available solutions and approaches to AR depend on how the problem is posed and the type of analogies being considered. Identifying common links within an analogy is the subject of AR, which has three primary processes: (1) retrieval, (2) mapping, and (3) evaluation [7]. At the heart of AR, in its psychological sense, research is focused on how the mapping process takes place and the best hypothesis for how it occurs in humans [18].

2.1. Forms of Analogies

In general, an analogy consists of two parts, the "base" or "source" (familiar scenario) and the "target" (unfamiliar scenario). Common analogy problems are of the word form shown in Equation 1 where *A* and *B* form the "base" of the analogy and *C* and *D* form the "target" [17],

$$A{:}B :: C{:}D. \quad (1).$$

Examples of word-based analogies, originally from Sternberg and Nigro [19] and modified in Morrision et al. [20], are shown in Figure 1. In addition to the *A*, *B*, *C*, and *D* words shown in Equation 1, there is also *D′* ["D prime"], which we are calling the "distractor," is contrasted with the "correct" *D*. Posed as *A:B::C:?*, the test subjects had a choice between *D* and *D′* based on which best completes the analogy.

Relationship	Example				
	A	B	C	D (Correct)	D' (Distractor)
Antonym	STOP	GO	EAST	WEST	DIRECTION
Synonym	NEAR	CLOSE	FIX	MEND	TAPE
Category - Subordinate	LION	ANIMAL	CHRISTMAS	HOLIDAY	EASTER
Category - Superordinate	DAY	SUNDAY	CLOTHES	SHOES	WEAR
Functional	BIRD	FLY	RABBIT	HOP	BUNNY
Linear Ordering	JANUARY	FEBRUARY	FIRST	SECOND	LAST

Figure 1. Analogy Categories and Examples

Ideally, AR models would be able to seamlessly consider semantics, structure, or both. However, an understanding of the AR methods' mechanics is needed to further comprehend their capabilities. These inherently follow the AI schools of thought.

2.2. Analogical Reasoning Model Types

At a high level, artificial AR is an AI approach and understanding it requires a general knowledge of the AI schools of thought: symbolist, connectionist, and dynamicist [21] [22]. These schools of thought differ largely on how intelligence is understood and conceptualized through artificial means. Briefly, symbolicism considers the mind to be a computer/logic system, connectionism considers the mind to be a neural network, and dynamicism considers the mind a watt governor [21]. These ideas are briefly described in Table 1. Given that biological mental processes likely follow a combination of these approaches (or something yet to be discovered), hybrid AI paradigms are also of interest as discussed by Eliasmith [23].

AR models, similarly, are structured according to these paradigms, but largely, they follow two: symbolist and connectionist (with some models being hybrids) [12] [13]. In AR applications, symbolist approaches consider each element of an analogy to be separate and independent from one another similar to a top-down approach [12]. Originally, the first AR methods were symbolic, beginning with Evan's 1963 ANALOGY model for visual AR problems [12]. Later in 1989, Gentner's word-based structure mapping theory (SMT) would be turned into the influential AR model, the structure mapping engine (SME) (part of the Many Are Called but Few Are Chosen (MAC/FAC) program) [24] [25]. Several symbolic models followed, such as the Incremental Analogy Machine (IAM) and Heuristic-Driven Theory Projection (HDTP) [13].

Though AR's origins started with symbolist models, currently there is a push toward

Table 1. General differences across AI paradigms, adapted from [21] [22]

Paradigm	SYMBOLISM	CONNECTIONISM	DYNAMICISM
Metaphor	Symbol system	Neural system	Dynamical System
Example	Mind as Computer	Mind as Brain	Mind as Watt Governor
Mechanism	Logical	Electrical	Mechanical
Description	Syntactic	Functional	Behavioral
Representation	Localist	Distributed	Continuous
Organization	Structural	Connectionist	Differential
Adaptation	Substitution	Tuning	Rate Change
Processing	Sequential	Parallel	Dynamical
Structure	Procedure	Network	Equation
Mathematics	Logic, Formal Language	Linear Algebra, Statistics	Geometry, Calculus
Space/Time	Formal	Spatial	Temporal

connectionism [12]. These models are characterized by elements that are associated using a bottom-up approach; many do this in a distributed fashion. The first connectionist model was Holyoak and Thagard's 1989 Analogical Constraint Mapping Engine (ACME), though its methods followed symbolist ideals more so than today's standard for connectionism [26]. However, some more recent models include Structure Tensor Analogical Reasoning (STAR) [27] [28], Learning and Inference with Schemas and Analogies (LISA) [29], Discovery Of Relations by Analogy (DORA) [30], and Bayesian Analogy with Relational Transformations (BART) [31] [32]. STAR is a tensor-product-based parallel distributed processing model embedded in a neural network [27], a framework popular for many AR models to come. LISA uses a neural network to process analogies while modeling a human's short-term and long-term memory [29]. DORA focuses on improving and incorporating self-supervised learning (SSL) into LISA [30]. SSL has enabled role-fillers to fire asynchronously; whereas, in LISA once fired, all corresponding semantic units are activated [30]. Additionally, VSMs have been included in the connectionist paradigm due to operating in a distributed fashion. Latent Relation Analysis (LRA) was one of the first VSMs created in 2006 (see [33]); however, since then, the creation of Word2vec, Global Vectors (GloVe), 3CosAvg, and LRCos, as well as many others, has been accomplished.

Considering the benefits of both the symbolist and connectionist models, some research has investigated hybrid models that incorporate the best of both [12]. The first hybrid model was Copycat which had a unique domain of nonsensical strings (example: *ABC:ABD::PQR:{PQS, PQD, or PQR}*) [34]. Copycat later inspired the creation of an action-based analogy program called Tabletop [35]. The first generally accepted word/sentence-based hybrid model was created in 1994, called the Associative Memory-Based Reasoning (AMBR) model [36] [37], which was followed by Distributed Representation Analogy Mapper (DRAMA) [38]. Few hybrid models exist due to their complexity compared to the number of symbolist and connectionist models [13].

Following this reasoning, a general taxonomy of AR methods appears in Figure 2. While no known dynamicist AR method exists to date, this paradigm of AI is included for completeness.

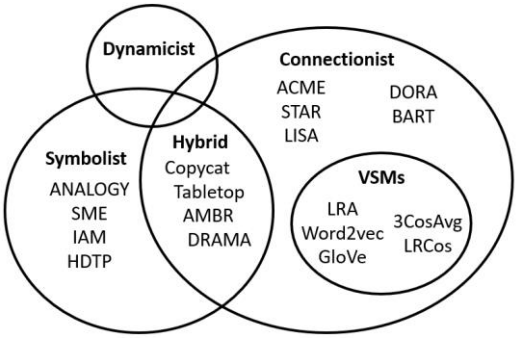

Figure 2. AR Models in the Context of AI Schools of Thought

To provide a more complete overview of the AR field, the general lineage of AR methods is presented in a temporal taxonomy in Figure 3. Notably, several of these algorithms are the subject of continuous research and revision. Many models are refined and improved upon over time, by the same or different investigators, creating a sense of linearity with respect to one another similar to a "family."

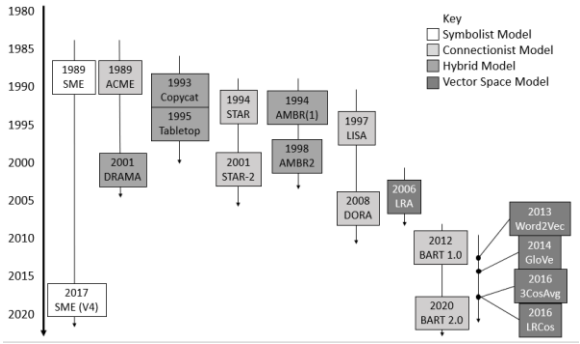

Figure 3. AR Model Timeline

3. Methodology

Several in-depth theory comparisons of various AR models exist [11] [12] [13] [14] [15]; however, algorithm performance on a common dataset with consistent metrics across more recent connectionist and hybrid models has yet to appear in the literature. Additionally, while comparisons have been made between AR methods, these comparisons are exclusively limited to those with psychological heritage or with VSM-classification. Thus, this study aimed to review AR methods from both backgrounds and selected methods that could solve simple word-based analogies.

3.1. Selection of AR Methods for Analysis

AR algorithms were selected for analysis based on their recency, previous success, and ease of

implementation on simple analogies. From the literature search, two psychological algorithms (DRAMA and BART) and four VSMs (Word2vec, GloVe, 3CosAvg, and LRCos) were selected. The lineages in Figure 2 show the most recently developed connectionist and hybrid models (based on our literature review) that were selected. As for the VSM selection, the baseline for new model performance has been Word2vec, e.g. [39], and GloVe, e.g. [40]; therefore, they were selected [10] [41]. With VSMs' ease of use and recent rise to popularity, many have been created, but 3CosAvg and LRCos were selected due to their early promising results (e.g. [11] [42]).

3.2. Psychological AR Models

As suggested earlier, psychological models have the potential for better applications compared to VSMs. Two models were selected for comparison in this study, but a review of other methods is presented for completeness. These models are considered to be "psychology-based" since their authors drew inspiration from and primarily worked in cognitive science at the time of their development(s).

3.2.1 SME. Due to its early prominence and symbolist nature, SME necessitates discussion. Originating in 1989, SME has been continually expanded with the most recent version (v4) being published in 2017 [43]. SMT posits that parts of an analogy should be mapped based on object relationships, which emphasizes structure [18]. The resulting SME mappings are measured with a structural evaluation score, which is the sum of match rule weights for the given base and target [24].

3.2.2. STAR. STAR was a connectionist model first created in 1994 and then, later expanded in 2001 in what was dubbed the "STAR-2" model [27] [28]. The original STAR model was the first distributed connectionist model, which is characterized by having representations exist over multiple units rather than just one (as in AMCE [26]) [27]. STAR-2's ability for hierarchically structured analogies allows it to solve problems the original STAR could not (such as the heat-flow/water-flow analogy e.g. [18]) in addition to an attempt to better mimic human capacity [28].

3.2.3. AMBR. Copycat was the first, and perhaps the most prominent, hybrid model, but limited in the sense of only applying to alphabetic strings [34]. AMBR was one of the first word-based hybrid AR model and was later expanded into AMBR2A and AMBR2B versions [36] [44]. AMBR was built upon the DUAL cognitive architecture, whose key distinction was small "dual agents" that form "coalitions" to complete tasks [45]. These dual agents allow for the retrieval, mapping, and transfer processes to occur in parallel [36]. AMBR2A [37] added a variety of new features, but in particular, allowed for decentralized representations [44]. AMBR2B modifications improved the constraint satisfaction network and recall from the system's long-term memory (LTM) [44].

3.2.4. LISA/DORA. Similar to STAR, LISA was based on a neural network and allowed knowledge sharing between its working memory and long-term memory [29]. LISA's performance was based on the difference between the correct mapping value and the highest incorrect mapping value [29]. LISA was the basis for the DORA model, which allowed for "asynchronous" firing as opposed to LISA's "synchronous" ability [30]. DORA's results were measured based on a "selectivity metric" (SM) associated with a semantic unit calculated by taking the average weight between the unit and relevant other units divided by the average weight between the unit and irrelevant other units plus one to help with standardization [30].

3.2.5. DRAMA. Despite using ACME as its basis, DRAMA has been generally accepted to be a hybrid model [46]. DRAMA uses holographic reduced representations (HRRs) (as discussed by Plate in [47]) and manipulates them through convolution and superimposition [46]. By nature, HRRs are influenced by noise, and experimental data shows that HRRs can yield results similar to human recollection [46]. DRAMA compares elements in the source and target by taking their dot product and dividing it by an arbitrary weight on semantics called the "semantic similarity" parameter, which is incorporated into the "activation" variable directly used to determine the analogy's final mapping [46].

3.2.6. BART. BART is one of the more recent AR models, which initially focused on solving comparative judgment problems [31]. BART draws inferences based on simple analogies, which makes it one of the few psychology-based models unable to solve sentence-based data. In this limited sense, BART uses bootstrapping to create "probabilistic weight distributions," which are then used to derive "importance-guided mappings."

In 2017, the creators of BART wanted to make the model more general which led to the creation of BART-g [48]. BART-g is still limited to simple analogies; however, it has the further ability to answer questions (such as "What is an animal larger than a dog?") that the original BART could not [48].

In 2019, the second version of BART (BART 2.0) was released with several improvements including using the SemEval-2012 Task-2 dataset to train BART 2.0 on other semantic relationships in addition to the comparative ones that BART 1.0 focused on [32].

3.3. VSMs for AR

As mentioned earlier, there has been an increase in the use of VSMs for AR [11] [41]. Word2vec specifically, but VSMs in general, have made exceptional progress in the field of auto-generation of semantics [9]. VSMs compile words/terms within documents to create a term-document matrix, later used to calculate various metrics such as the association between a pair of words or documents [49]. However, VSMs are limited in their abilities due to their lack of consideration of syntax- and semantic-related information, and their inability to identify analogies in sentence form [50]. The VSMs selected for this study can be customized with an alternative corpus; however, we used their default corpus, which was limited to the words the model was initially trained on. However, this limitation is addressed in FastText's model [51] and is the subject of other NLP-related research.

3.3.1. Word2vec. Word2vec has its roots in NLP and uses Skip-gram (a feed-forward neural network (NN) discussed more in [52]) as its internal mechanism (which alternatively can be switched with its Continuous Bag of Words (CBOW) in another variation) [53]. What has allowed Word2vec to make such a lasting impression is its ability to perform vector calculations on word problems. As discussed in [53] given the analogy *Spain:Madrid::France:?*, Word2vec can successful identify "Paris" through manipulating the original problem into:

$$v_{Paris} = v_{Madrid} - v_{Spain} + v_{France} \quad (1)$$

where, when attempting to calculate v_{Paris}, Word2vec uses a formula called 3CosAdd,

$$\underset{b' \in V}{arg\ max}\ (\cos(b', b - a + a')) \quad (2)$$

which considers the statement in Equation (2) in the general form: *a:a'::b:b'* where *b'* represents the attempted solution(s) to the problem *a:a'::b:b'*, not necessarily the *b'* corresponding to the (most) "correct" solution. The 3CosAdd method requires vector normalization and requires the words corresponding to *a*, *a'*, and *b* to be excluded from the space of possibility for *b'* [11].

3.3.2. LRA. In addition to the typical characteristics of a VSM, LRA allows the automatic derivation of corpus patterns and word pair synonyms and incorporates singular value decomposition [33]. LRA was applied to multiple-choice Scholastic Assessment Test (SAT) questions of the form: *A:B*::*C:D*, where *C* and *D* were presented in pairs among the choices [33]. LRA selects the best word pair based on a comparison of the source's (*A:B*) and the target's (*C:D*) "near analogies" and the commonalities amongst them [33]. The resulting frequencies are used to select the most correct answer to the given question.

3.3.3. GloVe. Unlike Word2vec, GloVe does not make use of a NN, but rather a "co-occurrence matrix" [54]. The creators of GloVe also introduce a new way to measure similarity,

$$\underset{b' \in V}{arg\ max}\ (cos\ (b', b) - cos(b', a) + cos\ (b', a')) \quad (3)$$

called 3CosMul or PairDistance, which replaces 3CosAdd in Word2vec's implementation. This method allows for more context to be considered by comparing the target, *b'*, individually with other elements of the analogy, *a, a'*, and *b*.

3.3.4. 3CosAvg and LRCos. In 2016, Drozd et al. [42] developed two alternatives to the standard 3CosAdd/3CosMul calculations used previously. 3CosAvg takes into consideration all vectors in the initial training set instead of just the *a:a'* pair [42]. This is achieved through

$$\underset{b' \in V}{arg\ max}\ (\cos(b', b + avg\_offset)) \quad (4)$$

$$avg_{offset} = \frac{\sum_{i=0}^{m} a'_i}{m} - \frac{\sum_{i=0}^{n} a_i}{n} \quad (5)$$

where Equation (4) has been corrected from its original presentation in [42] as identified in [55].

Though still using cosine similarity, LRCos factors in linear regression, as its name suggests. LRCos considers the probability that b' belongs to the target class that corresponds with a'. The corresponding formula for LRCos is thus

$$\underset{b' \in V}{arg\ max}\ P(b' \in target\_class)\cos\ (b', b) \quad (6).$$

4. Comparative Assessment and Evaluation

These models' success has been proven in their own analyses, but their outcomes compared to one

another in an AR context have only been tested in a limited sense. To broadly compare these algorithms, it was necessary to find applicable data and appropriate metrics.

4.1. Representative Example Data

As mentioned previously, there are several different formats for analogy problems [17]. For our apples-to-apples comparison, the Sternberg and Nigro dataset (originally used in [19]) was selected; however, due to availability, a modified version (from Morrison et. al [20]) was used. The modified version only provides two choices rather than the original four (as shown in Figure 1) to complete the *A:B::C:?* analogy. Within the dataset, there are five different analogy types: antonym, synonym, category (further broken down into subordinate and superordinate), functional, and linear ordering as identified in [19]. There are 40 antonym and 40 synonym analogies, which present opposite or alike words, respectively. There is a total of 40 categorical analogies with 35 being subordinate (specific to broad class) and 5 being superordinate (broad to specific class). The 41 functional analogies generally consider an object and an associated action or vice versa. Finally, there are 36 linear ordering analogies, which have a sequential relationship.

4.2. Performance Metrics

To facilitate this comparision, appropriate performance metrics were developed and determined to be correctness and analogy goodness. In general, correctness is the number of times the algorithm correctly selected *D* (over *D′*) divided by the total (also called "raw") or adjusted number of analogies as shown in Equations 7 and 8:

$$Raw\ \%\ Correct\ (RPC) = \frac{\#\ D\ was\ selectd\ over\ D'}{Total\ Number\ of\ Analogies} \quad (7)$$

$$Adj.\ \%\ Correct\ (APC) = \frac{\#\ D\ was\ selectd\ over\ D'}{Adjusted\ Number\ of\ Analogies} \quad (8).$$

The model's selection between *D* and *D′* is based on a comparison of their similarity metric explained in the next paragraph. The "raw" values are the total number of analogies in the overall set for a given relationship, and the "adjusted" values are the number of analogies that the given algorithm has the potential to answer correctly. In several instances, the model was unaware of the *A, B,* and/or *C* words' existence, which made the remainder of the analysis impossible. With that being said, the overall algorithm should not be penalized for this; however, if an algorithm has not encountered many words, it is also not ideal.

The similarity metric is a continuous value that measures how similar two words are. When calculating this, DRAMA uses the dot product between two word vectors, $\overrightarrow{v_1}$ and $\overrightarrow{v_2}$ (symbolized, $\overrightarrow{v_1} \cdot \overrightarrow{v_2}$); whereas, BART, Word2vec, GloVe, 3CosAvg, and LRCos use cosine similarity to compare the potential solution space. DRAMA's similarity scale ranges from [-1,1] instead of [0,1]; to normalize these values, DRAMA's similarity scores were modified per

$$sim_{DRAMA} = \frac{\overrightarrow{v_1} \cdot \overrightarrow{v_2}}{2} + \frac{1}{2} \quad (9)$$

which will be referred to as its "similarity metric" to normalize with the other models.

In its original setting, the dataset was constructed so that there was a "correct" answer among the four choices [19]. Understanding that *D* is the best choice amongst the other options, it is assumed that *A:B::C:D* is an "ideal" analogy (though individuals may differ on whether this is true). If *A:B::C:D* is, in fact, an ideal analogy, then the similarity ratio, sim_r (described in Equation (10), should theoretically equal one. The goodness metric evaluates how close the algorithm's predicted sim_r compares to an ideal analogy's similarity ratio. To calculate an analogy's goodness metric, the following steps take place:

i) Calculate the similarity score between *A* and *B*. sim_{AB}

ii) Calculate the similarity score between *C* and *D*, sim_{CD}

iii) Take the ratio between the similarity scores calculated above:

$$sim_r = \frac{sim_{AB}}{sim_{CD}} \quad (10)$$

iv) Take the difference between the similarity ratio for an "ideal" analogy, 1, and the ratio calculated above for the resulting analogy goodness measure,

$$Goodness = 1 - sim_r = 1 - \frac{sim_{AB}}{sim_{CD}} \quad (11).$$

5. Results

Results were obtained using the data from [19] and the correctness and goodness metrics for the algorithms: DRAMA, BART 1.0, BART 2.0, Word2Vec, GloVe, 3CosAvg, and LRCos.

Analogy Relationship	Correctness Metrics													
	Raw Percent Correctness (RPC)							Adjusted Percent Correctness (APC)						
	DRAMA	BART 1.0	BART 2.0	Word2vec	GloVe	3CosAvg	LRCos	DRAMA	BART 1.0	BART 2.0	Word2vec	GloVe	3CosAvg	LRCos
Antonym	72.5%	42.5%	75.0%	42.5%	72.5%	40.0%	42.5%	72.5%	42.5%	75.0%	42.5%	72.5%	42.1%	44.7%
Synonym	80.0%	47.5%	76.3%	37.5%	55.0%	47.5%	50.0%	80.0%	47.5%	76.3%	41.7%	55.0%	50.0%	52.6%
Category	82.5%	42.5%	57.5%	47.5%	50.0%	57.5%	67.5%	82.5%	42.5%	57.5%	50.0%	51.3%	59.0%	69.2%
Subordinate	85.7%	42.9%	54.3%	54.3%	48.6%	57.1%	65.7%	85.7%	42.9%	54.3%	55.9%	48.6%	58.8%	67.6%
Superordinate	60.0%	40.0%	80.0%	0.0%	60.0%	60.0%	80.0%	60.0%	40.0%	80.0%	0.0%	75.0%	60.0%	80.0%
Functional	78.0%	58.5%	78.0%	56.1%	53.7%	61.0%	41.5%	78.0%	58.5%	78.0%	57.5%	55.0%	64.1%	43.6%
Linear Ordering	80.6%	63.9%	71.4%	52.8%	63.9%	47.2%	38.9%	80.6%	63.9%	71.4%	55.9%	65.7%	50.0%	41.2%
All	**78.7%**	**50.8%**	**71.6%**	**47.2%**	**58.9%**	**50.8%**	**48.2%**	**78.7%**	**50.8%**	**71.6%**	**49.5%**	**59.8%**	**53.2%**	**50.5%**

Figure 4. Correctness Metric

5.1 Correctness Results

Figure 4 presents the percentage correct using the raw and adjusted total number of analogies as the denominators as shown in Equations 8 and 9, respectively. While the APC is a fairer comparison, it is important to consider the difference between the RPC and APC values since if there is a large difference, this suggests that an algorithm lacks vital "vocabulary." An ideal algorithm would be able to identify every word so that it can at least attempt every analogy. DRAMA and BART 1.0 successfully attempted each problem; however, they were partially reliant on hand-coding, unlike the VSMs and BART 2.0, which were completely autonomous in our scenario.

Figure 4 presents each model's performance within each analogical relationship type. DRAMA had the best overall performance and outperformed the other algorithms on the synonym, category, and linear ordering relationships. However, BART 2.0 tied DRAMA's performance on functional analogies and had a slight advantage on those with an antonym relationship. DRAMA also had the highest performance for subordinate category problems; however, for the superordinate, BART 2.0 and LRCos tied one another. Since some of BART 1.0 and all of DRAMA's mappings require hand-coding to identify the words within the analogies, their RPC and APC correctness scores are the same. All of the models were trained enough to attempt at least 188 of the total 197 analogies.

It is clear that overall, DRAMA was the best model for the given dataset, followed by BART 2.0 and GloVe, respectfully, with the remaining algorithms having a similar performance around the 50% mark. At the top level, there was not a large difference in results between the RPC and APC scores; however, there was some shifting among the lower-ranking algorithms such as 3CosAvg and LRCos.

Despite DRAMA's exceptional performance, there is not a "one size fits all" algorithm regarding the different analogy relationships tested. Though valuable, overall correctness may not be appropriate for studies that consider a large number of potential answers for *D*, an area where VSMs perform better.

5.2 Goodness Results

In a comparison of the similarity metric, a heatmap of the analogy goodness measure scores for all of the considered data is shown in Figure 5. In the figure, an analogy goodness measure of 0.000 indicates that the given *A:B::C:D* is equivalent to an "ideal" analogy as discussed in 4.2 and shown in (11. An "average" analogy was determined to be 0.251 based on an average of the goodness score across all the algorithms. Anything with a score equal to or greater than 1.000 was considered a "poor" analogy. As mentioned earlier, the VSMs and default BART 2.0 were not trained on certain words, and a goodness score could not be calculated; these instances were denoted in black.

Looking at the average shown in the bottom row of Figure 5, the algorithms rank as follows based on the goodness metric:

1. LRCos (0.055)
2. 3CosAvg (0.078)
3. BART 1.0 (0.107)
4. BART 2.0 (0.220)
5. Word2Vec (0.417)
6. DRAMA (0.434)
7. GloVe (0.445).

When doing a broad visual overview, 3CosAvg and LRCos appear to be roughly tied followed by BART 1.0, BART 2.0, and the remaining models, which were tied on a different scale. In summary, LRCos provided the best possible comparison between analogies; however, it was followed relatively closely by 3CosAvg and BART 1.0, respectively.

6. Conclusions

The authors presented a review and analysis of analogical reasoning (AR) algorithms for word-based analogies. This review focused on 7 algorithms: DRAMA [46], BART 1.0 [31] & 2.0 [32], Word2vec [53], GloVe [54], 3CosAvg [42], and LRCos [42],

Key											
	0.000	Perfect Analogy		0.251	Average Analogy		1.000+	Poor Analogy		Unable to Attempt	

Analogy	DRAMA	BART 1.0	BART 2.0	Word2vec	GloVe	3CosAvg	LRCos
YES:NO::TRUE:FALSE	0.057	0.221	0.226	0.049	0.000	0.019	0.023
COOL:WARM::BLACK:WHITE	0.013	0.029	0.159	0.497	0.065	0.041	0.042
OFTEN:SELDOM::HAPPY:SAD	0.392	0.145	0.159	1.345	0.031	0.008	0.000
LOVE:HATE::HIT:MISS	0.495	0.098	0.234	0.804	0.415	0.020	0.125
STOP:GO::EAST:WEST	0.533	0.085	0.170	0.422	0.308	0.001	0.002
NARROW:WIDE::QUESTION:ANSWER	0.765	0.027	0.216	0.063	0.376	0.048	0.004
WILD:TAME::HARD:SOFT	0.008	0.068	0.178	0.210	0.287	0.081	0.001
STRAIGHT:BENT::FIND:LOSE	0.010	0.060	0.246	0.606	0.478		
UP:DOWN::POOR:RICH	0.068	0.147	0.250	0.058	0.807	0.003	0.004
EMPTY:FULL::BETTER:WORSE	0.057	0.062	0.308	0.154	0.460	0.106	0.002
WIN:LOSE::ABOVE:BELOW	0.101	0.222	0.208	0.276	0.315	0.008	0.002
LIKE:DISLIKE::WARM:COOL	0.098	0.120	0.239	0.400	0.680	0.047	0.051
RIGHT:WRONG::CALM:STORMY	0.390	0.026	0.291	1.617	1.195	0.114	0.088
FAST:SLOW::ON:OFF	0.120	0.010	0.166	0.173	0.212	0.032	0.031
FOOLISH:WISE::EARLY:LATE	0.051	0.121	0.217	0.542	0.376		
COME:GO::YOUNG:OLD	0.450	0.094	0.158	0.077	0.932	0.008	0.000
BEGIN:END::DARK:LIGHT	0.653	0.036	0.247	0.295	0.073	0.048	0.060
OPEN:CLOSE::SHALLOW:DEEP	0.119	0.210	0.222	0.034	0.294	0.194	0.072
BLACK:WHITE::BAD:GOOD	0.119	0.165	0.100	0.864	0.107	0.385	0.373
QUIET:LOUD::CATCH:THROW	0.020	0.186	0.240	0.086	0.305	0.042	0.090
JOHN:NAME::DINNER:MEAL	0.716	0.064	0.264	0.326	0.562	0.017	0.007
ROSE:FLOWER::CHURCH:BUILDING	0.542	0.124	0.309	0.042	0.474	0.049	0.075
BREAD:FOOD::BLACK:COLOR	0.185	0.063	0.222	0.118	0.127	0.097	0.094
RED:COLOR::HIT:ACTION	0.078	0.077	0.183	0.169	0.936	0.112	0.048
BEE:INSECT::LOVE:FEELING	0.110	0.056	0.228	0.623	0.316	0.014	0.010
NOON:TIME::WEST:DIRECTION	0.643	0.086	0.238	0.148	0.107	0.006	0.000
ENGLISH:LANGUAGE::BASEBALL:GAME	0.094	0.229	0.300	0.416	0.174	0.101	0.034
LION:ANIMAL::CHRISTMAS:HOLIDAY	0.019	0.156	0.211	0.459	0.571	0.053	0.005
CITY:NEW::YORKFISH:GUPPY	0.706	0.204	0.196			0.071	0.029
ANIMAL:CALF::CAT:SIAMESE	0.058	0.099	0.227		0.080	0.100	0.084
MOTHER:WOMAN::STREET:THOROUGHFARE	0.611	0.214	0.255		0.765	0.114	0.130
DAY:SUNDAY::CLOTHES:SHOES	0.586	0.149	0.242	0.200	0.047	0.057	0.123
TABLE:FURNITURE::APPLE:FRUIT	0.058	0.042	0.304	0.179	0.315	0.152	0.151
UNCLE:MAN::DOG:ANIMAL	0.552	0.022	0.208	0.576	0.251	0.001	0.013
SONG:MUSIC::PAINTING:ART	0.675	0.065	0.148	0.449	0.064	0.145	0.040
GOLD:METAL::SLOW:SPEED	0.086	0.071	0.211	0.569	0.400	0.019	0.020
RIVER:WATER::MOUNTAIN:LAND	0.704	0.125	0.212	0.239	0.713	0.033	0.032
CHAIR:FURNITURE::YOUNG:AGE	0.597	0.166	0.301	0.462	0.310		
TROUT:FISH::PIG:ANIMAL	0.814	0.250	0.231	0.028	0.199	0.418	0.449
G:LETTER::SEVEN:NUMBER	0.059	0.029	0.357	0.907	0.666	0.281	0.020
BIRD:FLY::RABBIT:HOP	0.548	0.071	0.118	0.048	3.344	0.186	0.112
HORN:PLAY::HORSE:RIDE	0.566	0.199	0.256	0.802	0.612	0.000	0.025
ROAD:CAR::SKY:PLANE	0.069	0.048	0.168	0.114	0.200	0.045	0.034
HOT:FURNACE::COOL:FAN	0.763	0.084	0.332	0.432	0.018	0.027	0.011
PLAY:GAME::GIVE:PARTY	0.681	0.075	0.216	0.300	1.598	0.132	0.020
WET:RAIN::DRY:SUN	0.634	0.146	0.194	0.174	0.917	0.043	0.031
BOOKS:READ::TOYS:PLAY	0.120	0.077	0.183	0.141	2.218	0.029	0.000
TASTE:MOUTH::TOUCH:HANDS	0.798	0.136	0.281	0.177	0.382		
SCHOOL:LEARN::STORE:BUY	0.013	0.002	0.189	0.329	0.201	0.123	0.055
PEOPLE:FOOD::CARS:GAS	0.305	0.120	0.277	0.104	0.104	0.002	0.017
BED:SLEEP::CHAIR:SIT	0.590	0.329	0.172	1.599	0.171	0.255	0.260
MORNING:BREAKFAST::EVENING:DINNER	0.394	0.058	0.095	0.109	0.199	0.081	0.117
FIRE:BURN::WIND:BLOW	0.060	0.104	0.204	0.128	0.209	0.065	0.045
STORY:TELL::SONG:SING	0.431	0.198	0.133	0.007	0.160	0.004	0.008
AIRPLANE:SKY::SHIP:OCEAN	0.434	0.045	0.178	0.000	0.543	0.001	0.001
BIRD:NEST::HORSE:STABLE	0.559	0.341	0.185	0.008	0.232	0.273	0.262
PLATE:EAT::CUP:DRINK	0.792	0.059	0.215	0.180	0.220	0.018	0.085
COOK:KITCHEN::BUY:STORE	0.434	0.009	0.164	0.740	0.011	0.285	0.004
CAR:DRIVE::BOAT:SAIL	0.178	0.223	0.140	0.174	0.115	0.065	0.090
READ:NEWSPAPER::SIT:CHAIR	0.660	0.197	0.205	2.247	0.186	0.010	0.120
MONTH:YEAR::INCH:FOOT	0.785	0.005	0.190	0.536	1.019	0.075	0.026
JANUARY:FEBRUARY::FIRST:SECOND	0.560	0.013	0.188	0.443	0.135	0.125	0.125
PALACE:CABIN::ROCK:PEBBLE	0.668	0.227	0.258	0.334	0.470	0.109	0.008
GOOD:BETTER::BAD:WORSE	0.427	0.098	0.078	0.141	0.218	0.119	0.021
WORSE:WORST::LOWER:LOWEST	0.663	0.032	0.102	0.260	0.173	0.007	0.000
PUPPY:DOG::CUB:BEAR	0.451	0.011	0.222	2.902	0.527	0.015	0.020
WARM:HOT::COOL:COLD	0.071	0.064	0.327	0.004	0.109	0.098	0.099
EGG:CHICKEN::SEED:FLOWER	0.804	0.050	0.258	0.414	0.702	0.029	0.069
SMALL:SMALLER::APPLE:FRUIT	0.034	0.035	0.253	0.234	0.612	0.138	0.136
BIRTH:LIFE::DAWN:DAY	0.747	0.033	0.254	0.243	0.036		
BREAKFAST:LUNCH::LUNCH:DINNER	0.004	0.025	0.232	0.021	0.061	0.097	0.041
MORNING:AFTERNOON::AFTERNOON:EVENING	0.067	0.027	0.082	0.052	0.141	0.012	0.003
INCH:FOOT::YARD:MILE	0.124	0.199	0.297	0.145	0.111	0.006	0.019
PRINCE:KING::PRINCESS:QUEEN	0.621	0.135	0.105	0.339	0.025	0.011	0.011
PERSON:CROWD::DROP:PUDDLE	0.817	0.219	0.323	1.037	1.978	0.044	0.003
NONE:SOME::MOST:ALL	0.678	0.060	0.129	0.101	0.203	0.029	0.035
RUN:WALK::SHOUT:TALK	0.331	0.181	0.269	0.387	1.074	0.122	0.056
FOURTH:FIFTH::APRIL:MAY	0.544	0.005	0.396	0.047	0.367	0.171	0.074
CAR:BUS::KNIFE:CLEAVER	0.415	0.020	0.191	0.465	0.680	0.148	0.006
EQUAL:SAME::QUIET:STILL	0.642	0.129	0.247	0.034	0.383	0.011	0.015
NEARLY:ALMOST::CAR:AUTO	0.662	0.011	0.263	1.038	0.528	0.066	0.058
EASY:SIMPLE::SHUT:CLOSE	0.528	0.172	0.256	0.144	0.629	0.017	0.008
END:FINISH::BROOK:STREAM	0.617	0.010	0.202	0.101	0.040	0.003	0.012
NEAR:CLOSE::FIX:MEND	0.087	0.222	0.258	2.243	0.730	0.051	0.051
CERTAIN:SURE::SHIP:BOAT	0.725	0.261	0.165	0.259	0.243	0.087	0.001
STEAL:ROB::PULL:DRAG	0.586	0.104	0.159	0.584	0.036	0.021	0.008
HUGE:ENORMOUS::PAMPHLET:BOOKLET	0.197	0.270	0.102		0.446		
QUARREL:FIGHT::BUILD:MAKE	0.730	0.063	0.290		0.603	0.004	0.019
THROW:PITCH::BEGIN:START	0.643	0.105	0.264	0.711	0.399	0.057	0.002
BIG:LARGE::WEAK:FEEBLE	0.454	0.201	0.151		0.007	0.382	0.484
FOREST:WOODS::STREET:ROAD	0.380	0.121	0.249	0.127	0.221	0.095	0.027
HAVE:POSSESS::HARD:DIFFICULT	0.684	0.007	0.186	0.030	0.499	0.051	0.008
LIBERTY:FREEDOM::FATHER:DAD	0.146	0.089	0.155	0.182	0.025	0.084	0.115
HELP:AID::HAT:CAP	0.612	0.002	0.272	0.402	0.380	0.011	0.001
MERRY:GAY::INTELLIGENT:SMART	0.053	0.075	0.333	0.353	0.925	0.004	0.002
CORRECT:RIGHT::OLD:AGED	0.708	0.122	0.240	0.273	0.133	0.009	0.001
REMAIN:STAY::SPEAK:TALK	0.429	0.136	0.145	0.396	0.142	0.047	0.029
RICH:WEALTHY::FAT:ROTUND	0.492	0.015	0.215		3.426	0.059	0.070
FAST:SPEEDY::DOCTOR:PHYSICIAN	0.882	0.055	0.151	0.267	0.387	0.056	0.000

Analogy	DRAMA	BART 1.0	BART 2.0	Word2vec	GloVe	3CosAvg	LRCos
ENTER:LEAVE::FLOAT:SINK	0.291	0.177	0.202	0.317	0.396	0.014	0.018
DIFFERENT:SAME::SHORT:LONG	0.726	0.146	0.169	0.166	0.041	0.147	0.146
LOUD:SOFT::FAT:THIN	0.461	0.112	0.210	0.324	0.185	0.020	0.019
START:FINISH::FAR:NEAR	0.127	0.014	0.215	0.706	0.400	0.011	0.004
BACK:FRONT::WET:DRY	0.699	0.014	0.184	0.354	0.354	0.006	0.004
NEVER:ALWAYS::LOWEST:HIGHEST	0.814	0.197	0.132	0.001	0.038	0.063	0.054
FEW:MANY::NOISY:QUIET	0.070	0.225	0.230	0.080	0.986	0.004	0.040
ADD:SUBTRACT::BEST:WORST	0.713	0.042	0.186	0.341	0.137	0.020	0.039
INSIDE:OUTSIDE::OVER:UNDER	0.717	0.008	0.165	0.337	0.537	0.014	0.013
GOOD:BAD::NEW:OLD	0.539	0.062	0.184	0.261	0.753	0.191	0.181
HUGE:TINY::CLEAN:DIRTY	0.869	0.172	0.171	0.220	0.198	0.314	0.311
SLOWLY:QUICKLY::WORK:PLAY	0.603	0.086	0.223	0.408	0.494		
HEAVY:LIGHT::LONG:SHORT	0.667	0.226	0.268	0.153	0.327	0.007	0.005
ASLEEP:AWAKE::LARGE:SMALL	0.692	0.127	0.269	0.476	0.227	0.019	0.009
BEFORE:AFTER::AFRAID:BRAVE	0.193	0.135	0.253	0.649	1.175	0.005	0.002
SUCCEED:FAIL::REMEMBER:FORGET	0.079	0.197	0.212	0.475	0.398	0.132	0.026
WEAK:STRONG::GROW:WITHER	0.011	0.001	0.290		1.142	0.211	0.012
DANGER:SAFETY::CRY:LAUGH	0.003	0.027	0.203	0.595	0.307	0.072	0.006
FORWARD:BACKWARD::FUTURE:PAST	0.716	0.038	0.218	0.022	0.055	0.016	0.004
HIRE:FIRE::START:FINISH	0.020	0.229	0.306	0.064	0.699	0.029	0.012
EGG:FOOD::SOUTH:DIRECTION	0.064	0.061	0.297	0.097	0.011	0.135	0.094
QUEEN:RULER::CIRCLE:SHAPE	0.323	0.087	0.231	1.633	0.051	0.016	0.032
MONTH:MAY::SEASON:WINTER	0.423	0.127	0.208	0.201	0.031	0.065	0.063
CORN:VEGETABLE::DOLLAR:MONEY	0.126	0.127	0.251	0.032	0.120	0.015	0.014
HAMMER:TOOL::SHORT:SIZE	0.107	0.299	0.340	0.212	0.045	0.061	0.058
ONE:NUMBER::BALL:TOY	0.037	0.024	0.375	0.192	2.767	0.001	0.002
FRUIT:ORANGE::DOG:POODLE	0.750	0.037	0.173	0.333	0.001	0.167	0.212
SILVER:METAL::PRINCE:NOBILITY	0.053	0.126	0.237	0.583	0.567	0.030	0.032
ARITHMETIC:SUBJECT::HEAVY:WEIGHT	0.041	0.112	0.310		0.756		
HOUSE:BUILDING::SMALL:SIZE	0.717	0.238	0.165	0.177	0.048	0.153	0.078
DOCTOR:TITLE::PRIVATE:RANK	0.008	0.243	0.271	0.200	0.386	0.027	0.064
KITCHEN:ROOM::NEAR:DISTANCE	0.413	0.015	0.254	0.185	0.925	0.047	0.017
PINE:TREE::DARK:ILLUMINATION	0.668	0.085	0.236	0.086	1.563	0.006	0.010
NEST:HOME::OUNCE:WEIGHT	0.486	0.103	0.225	0.602	0.359	0.009	0.009
FOOTBALL:GAME::BIG:SIZE	0.802	0.040	0.303	0.033	0.273	0.065	0.082
ROBIN:BIRD::JUNE:MONTH	0.760	0.077	0.219	0.086	0.759	0.012	0.047
KING:RULER::BOY:MALE	0.014	0.127	0.280	0.168	0.540	0.098	0.071
SWIMMING:SPORT::GUN:WEAPON	0.360	0.162	0.228	0.351	0.248	0.001	0.005
PACIFIC:OCEAN::KITCHEN:ROOM	0.237	0.087	0.256	0.108	0.022	0.003	0.004
ZINC:MINERAL::SHORT:HEIGHT	0.450	0.300	0.238	0.059	0.495	0.031	0.024
CLIMB:HILL::DIG:HOLE	0.850	0.099	0.196	1.248	0.215	0.000	0.005
EAR:HEAR::EYE:SEE	0.749	0.192	0.082	0.096	0.215		
DOOR:OPEN::STAIRS:CLIMB	0.083	0.146	0.204	0.129	0.077	0.003	0.018
ZOO:ANIMALS::FOREST:TREES	0.527	0.050	0.152	0.071	0.121	0.082	0.064
CUT:KNIFE::MIX:SPOON	0.218	0.000	0.184	0.136	0.074	0.013	0.012
SHOES:FEET::HAT:HEAD	0.043	0.138	0.312	0.595	0.469	0.036	0.081
NAIL:BUIL::DPEN:WRITE	0.094	0.269	0.240			0.002	0.012
WARM:SUN::WET:CLOUDS	0.515	0.133	0.173	0.054	0.124	0.010	0.005
CLOCK:TIME::NEWSPAPER:NEWS	0.042	0.131	0.282	0.264	0.290	0.005	0.003
CRY:HURT::SMILE:HAPPY	0.173	0.006	0.269	0.449	0.451	0.037	0.002
MOON:NIGHT::SUN:DAY	0.547	0.069	0.218	0.458	0.004	0.017	0.075
GLASS:DRINK::PAN:COOK	0.705	0.009	0.162	6.516	0.197	0.042	0.036
SLED:SNOW::CAR:ROAD	0.650	0.338	0.243	0.629	0.342	0.042	0.118
RAIN:SUMMER::SNOW:WINTER	0.217	0.030	0.135	0.185	0.437	0.066	0.059
FIGHT:SOLDIER::HELP:DOCTOR	0.545	0.000	0.151	0.407	0.043	0.023	0.005
BICYCLE:RIDE::BASEBALL:THROW	0.562	0.090	0.237	0.082	0.535	0.004	0.004
APPLE:EAT::ROSE:SMELL	0.431	0.132	0.278	0.172	2.719	0.006	0.008
DIRT:SOAP::PAIN:PILL	0.114	0.004	0.218	0.758	0.476	0.008	0.007
PLAY:GAME::WORK:JOB	0.883	0.097	0.134	0.130	0.199	0.145	0.054
CAT:PET::COW:MILK	0.377	0.077	0.155	0.322	0.221	0.481	0.138
BASEBALL:MARBLE::BUCKET:GLASS	0.819	0.143	0.226	0.340	0.607	0.056	0.006
CRAWL:WALK::WALK:RUN	0.109	0.151	0.267	0.121	0.116	0.141	0.005
LITTLE:TINY::LARGE:HUGE	0.116	0.019	0.264	0.025	0.359	0.377	0.372
BETTER:BEST::TALLER:TALLEST	0.742	0.119	0.177		0.397	0.146	0.023
TAP:STRIKE::BREAK:DESTROY	0.570	0.070	0.282	0.709	0.676		
YESTERDAY:TODAY::BEFORE:NOW	0.706	0.027	0.089	0.042	0.202	0.010	0.012
NEVER:SOMETIMES::OFTEN:ALWAYS	0.832	0.004	0.185	0.180	0.178	0.299	0.001
PAGES:BOOK::LETTERS:WORD	0.090	0.214	0.259	0.622	0.265	0.133	0.010
A LOT:A LITTLE::WET:MOIST	0.397	0.031	0.216			0.213	0.011
KITTEN:CAT::CALF:COW	0.714	0.036	0.214	0.249	0.014	0.074	0.075
ONE:DOZEN::PENNY:DOLLAR	0.055	0.026	0.238	0.030	1.373	0.002	0.056
FUTURE:PRESENT::PRESENT:PAST	0.951	0.049	0.286	0.053	0.436	0.420	0.005
LESS:LEAST::MORE:MOST	0.505	0.069	0.138	0.395	0.110	0.050	0.065
LAKE:OCEAN::BIG:BIGGER	0.315	0.202	0.222	0.177	0.418	0.118	0.111
CITY:COUNTY::STATE:COUNTRY	0.941	0.058	0.249	0.189	0.045	0.074	0.074
MODERATE:HEAVY::RAIN:DOWNPOUR	0.876	0.179	0.188		0.590	0.124	0.002
MOSTLY:SOMEWHAT::OVERCAST:PARTLY CLOUDY	0.795	0.105				0.071	0.066
EAGLE:HAWK::LARGE:MODERATE	0.904	0.182	0.207	0.142	0.791	0.040	0.028
STRANGE:ODD::DISH:PLATE	0.809	0.154	0.265	1.287	0.875	0.250	0.056
PERHAPS:MAYBE::CHAIR:SEAT	0.033	0.013	0.272	1.985	0.498	0.084	0.056
ALLOW:LET::CRY:WEEP	0.572	0.058	0.194	0.041	0.343	0.048	0.024
HURRY:RUSH::SAD:UNHAPPY	0.426	0.020	0.226	1.584	0.140	0.219	0.137
UNDER:BENEATH::PAIN:HURT	0.502	0.159	0.259	0.310	0.097	0.018	0.016
ENJOY:LIKE::FALL:AUTUMN	0.058	0.025	0.217	0.010	0.174	0.278	0.271
SLENDER:THIN::SICK:ILL	0.696	0.015	0.256		0.231	0.102	0.101
NOTICE:SEE::MURDER:HOMICIDE	0.237	0.059	0.161	0.445	0.419	0.111	0.024
OVER:ABOVE::PANTS:TROUSERS	0.040	0.161	0.174	0.454	0.572	0.001	0.005
FOOLISH:SILLY::HANDGUN:PISTOL	0.554	0.012	0.199	0.128	0.113	0.125	0.025
MOTOR:ENGINE::MIDDLE:CENTER	0.896	0.190	0.215	0.107	0.401	0.051	0.023
DISCOVER:FIND::DONKEY:ASS	0.664	0.202	0.264	0.231	0.614	0.031	0.041
HAPPY:CHEERFUL::TABLET:PAD	0.494	0.206	0.300	2.412	1.517	0.005	0.240
PART:PIECE::MOTHER:MOM	0.839	0.158	0.247	0.343	0.356	0.231	0.231
SOIL:EARTH::HAPPINESS:JOY	0.750	0.000	0.238	0.070	0.340	0.048	0.077
DEFRAUD:CHEAT::CHILD:KID	0.232	0.132	0.256		0.038	0.096	0.042
GIFT:PRESENT::FOOT:12 INCHES	0.473	0.149				0.018	0.054
DIFFICULT:HARD::COUGAR:MOUNTAIN LOIN	0.875	0.047				0.031	0.002
ILLEGAL:UNLAWFUL::HINDER:IMPEDE	0.733	0.078	0.140		0.325		
STRANGE:UNUSUAL::FEMALE:WOMAN	0.448	0.305	0.212	0.024	0.054	0.021	0.087
Average	**0.434**	**0.107**	**0.220**	**0.417**	**0.445**	**0.078**	**0.055**

Figure 5. Heatmap of Analogy Goodness Metric

which encompass the general state of the art in the field today. Previous comparisons, see [11] [12] [13] [14] [15], only considered a small subset of these algorithms. In addition to providing a broad review of algorithms and their capabilities, the authors further provided comparison metrics and a consistent dataset for analysis. In a broad sense, it appears that psychological models currently have a slight advantage over VSMs based on our defined metrics, correctness and analogy goodness. When concerned with the selection of the correct answer, DRAMA is the best overall model (78.7% correctness); however, the "best" model may depend on the relationship of a given analogy. When comparing models based on how "good" the similarity of an analogy is, LRCos has a small advantage over the other models (goodness score of 0.055). Overall, combining both metrics, the results show BART 2.0 and 3CosAvg tied at 1st, DRAMA and LRCos tied at 3rd, and then BART 1.0 (5th), GloVe (6th), and Word2Vec (7th). Thus, there is no "one size fits all" AR algorithm.

Further work in this field could look at similar metrics, with the addition of an analogy goodness metric to evaluate *A:B::C:D'* in addition to what we considered with *A:B::C:D*. Another interesting metric could consider the similarity score between *D* and *D'* and factoring that into the correctness metric since some of the *D'* options seem trickier than others when identifying the correct answer. Finally, the inclusion of more models (specifically psychological ones) would help give future investigations a more comprehensive overview of the strengths and weaknesses of AR models as a whole.

7. Acknowledgements

The views expressed in this paper are those of the author(s) and do not necessarily represent the views of any part of the US Government. This work was cleared for unlimited release under: AFRL-2021-3051 and AFRL-2021-3049.

8. Bibliography

[1] IBM Cloud Education, "What is Artificial Intelligence (AI)?," *IBM*, 3 June 2020. [Online].

[2] S. Shankland, "'AI is very, very stupid,' says Google's AI leader, at least compared to humans," *CNET*, 14 Nov. 2018.

[3] T. Ray, "Intel's neuro guru slams deep learning: 'it's not actually learning'," *ZDNet*, 23 Feb. 2019.

[4] T. Bihl and M. Talbert, "Analytics for autonomous c4isr within e-government: a research agenda.," *Hawaii Int'l Conf. on System Sciences*, pp. 2218-2227, 2020.

[5] S. Srivastava, "Defining AI: Reasoning, Interaction and Learning," *CIO*, 9 November 2017. [Online].

[6] B.-T. Zhang, "Hypernetworks: A Molecular Evolutionary Architecture for Cognitive Learning and Memory," *IEEE Computational Intelligence Magazine*, 3(3), pp. 49-63, 2008.

[7] D. Gentner and F. Maravilla, "Analogical Reasoning," in *International Handbook of Thinking & Reasoning*, New York, Psychology Press, 2018, pp. 186-203.

[8] K. J. Holyoak, "Analogy and relational reasoning," in *The Oxford handbook of thinking and reasoning*, New York, Oxford University Press, 2012, pp. 234-259.

[9] IBM Cloud Education, "Natural Language Processing," IBM, 2 July 2020. [Online]. Available: https://www.ibm.com/cloud/learn/natural-language-processing. [Accessed 27 April 2021].

[10] D. Chen, et al., "Evaluating vector-space models of analogy," arXiv preprint arXiv:1705.04416, pp. 1-6, 2017.

[11] A. Rogers, et al., "The (Too Many) Problems of Analogical Reasoning with Word Vectors," *Joint Conf. on Lexical and Computational Semantics*, pp. 135-148, 2017.

[12] B. Kokinov and R. M. French, "Computational Models of Analogy-making," *Encyclopedia of Cognitive Science*, vol. 1, pp. 113-118, 2003.

[13] D. Gentner and K. D. Forbus, "Computational Models of Analogy," *WIREs Cognitive Science*, v2, 266-276, 2010.

[14] R. P. Hall, "Computational Approaches to Analogical Reasoning: A Comparative Analysis," *Artificial Intelligence*, vol. 39, pp. 39-120, 1989.

[15] R. M. French, "The Computational Modeling of Analogy-making," *TRENDS in Cognitive Sciences*, 6(5), pp. 200-205, 2002.

[16] E. D. Liddy, "Natural Language Processing," *Encyclopedia of Library and Information Science*, 2nd ed., Marcel Decker, Inc, 2001, pp. 1-15.

[17] N. Ichien, et al., "Verbal Analogy Problem Sets: An Inventory of Testing Materials," *Behavior Research Methods*, vol. 53, pp. 1803--1816, 2020.

[18] D. Gentner, "Structure-mapping: A Theoretical Framework for Analogy," *Cognitive Science*, 7(2), pp. 155-170, 1983.

[19] R. J. Sternberg and G. Nigro, "Developmental Patterns in the Solution of Verbal Analogies," *Child Development*, 51(1), pp. 27-38, 1980.

[20] R. G. Morrison, et al., "A Neurocomputational Model of Analogical Reasoning and its Breakdown in Frontotemporal Lobar Degeneration," *Journal of Cognitive Neuroscience*, 16(2), pp. 260-271, 2004.

[21] C. Eliasmith, "Computational and Dynamical Models of Mind," *Minds and Machines*, 7, pp. 531-541, 1997.

[22] B. Zhang, "Hypernetworks: A molecular evolutionary architecture for cognitive learning and memory," *IEEE computational intelligence magazine*, 3(3), pp. 49-63, 2008.

[23] C. Eliasmith, *How to build a brain*, Oxford University Press., 2013.

[24] B. Falkenhainer and K. D. Forbus, "The Structure-mapping Engine: Algorithm and Examples," *Artificial Intelligence*, 41(1), pp. 1-63, 1989.

[25] K. D. Forbus, et al., "MAC/FAC: A Model of Similarity-based Retrieval," *Cognitive Science*, 19(2), pp. 141-205, 1995.

[26] K. J. Holyoak and P. Thagard, "Analogical Mapping by Constraint Satisfaction," *Cognitive Science*, vol. 13, pp. 295-355, 1989.

[27] G. S. Halford, et al., "Connectionist Implications for Processing Capacity Limitations in Analogies," *Advances in Connectionist and Neural Computation Theory*, vol. 2, pp. 363-415, 1994.

[28] W. H. Wilson, et al., "The STAR-2 Model for Mapping Hierarchically Structured Analogs," *The Analogical Mind*, pp. 125-60, 2001.

[29] J. E. Hummel and K. J. Holyoak, "Distributed Representations of Structure," *Psychological Review*, 104(3), pp. 427-466, 1997.

[30] L. A. A. Doumas, et al., "A Theory of the Discovery and Predication of Relational Concepts," *Psychological Review*, 115(1), pp. 1-43, 2008.

[31] H. Lu, et al., "Bayesian Analogy with Relational Transformations," *Psychological Review*, 119(3), pp. 617-648, 2012.

[32] H. Lu, et al., "Emergence of analogy from relation learning," *Proceedings of the National Academy of Science*, 116(10), pp. 4176-4181, 2019.

[33] P. D. Turney, "Similarity of Semantic Relations," *Computational Linguistics*, 332(3), pp. 379-416, 2006.

[34] D. R. Hofstadter and M. Mitchell, "The Copycat Project: A Model of Mental Fluidity and Analogy-making," *Advances in Connectionist and Neural Computation Theory*, vol. 2, pp. 205-267, 1995.

[35] R. M. French, *The Subtlety of Sameness: A Theory and Computer Model of Analogy-making*, MIT Press, 1995.

[36] B. Kokiov, "A Hybrid Model of Reasoning by Analogy," in *Advances in Connectionist and Neural Computation Theory*, vol. 2, Ablex, 1994, pp. 247-318.

[37] A. A. Petrov, *Extensions of DUAL and AMBR*, New Bulgarian University, Cognitive Science Department, 1997.

[38] C. Eliasmith and P. Thagard, "Integrating structure and meaning: A distributed model of analogical mapping," *Congitive Science*, 25(2), pp. 245-286, 2001.

[39] C. Amrit and J. Hek, "Clustering the results of brainstorm sessions: Applying word similarity techniques to cluster Dutch nouns," *Hawaii Int'l Conf. on System Sciences (HICSS)*, pp. 4232-4241, 2016.

[40] J. Pennington, et al., "GloVe: Global Vectors for Word Representation," in *2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Doha, 2014.

[41] J. C. Peterson, et al., "Parallelograms Revisited: Exploring the Limitations of Vector Space Models for Simple Analogies," *Cognition*, vol. 205, pp. 1-15, 2020.

[42] A. Drozd, et al., "Word Embeddings, Analogies, and Machine Learning: Beyond King - Man + Woman = Queen," *Int'l Conf. Computational Linguistics*, Osaka, 2016.

[43] K. D. Forbus, et al., "Extending SME to Handle Large-Scale Cognitive Modeling," *Cognitive Science*, vol. 41, pp. 1152-1201, 2017.

[44] A. A. Petrov, *Associative Memory-Based Reasoning*, Saarbrucken: LAP LAMBERT Academic Publishing, 2013.

[45] B. N. Kokinov and A. A. Petrov, "Integration of Memory and Reasoning in Analogy-Making," in *The analogical mind*, MIT Press, 2000, pp. 59-124.

[46] C. Eliasmith and P. Thagard, "Integrating Structure and Meaning: A Distributed Model of Analogical Mapping," *Cognitive Science*, 25(2), pp. 245-286, 2001.

[47] T. A. Plate, *Distributed Representations and Nested Compositional Structure*, Toronto, Ontario: University of Toronto, Department of Computer Science, 1994.

[48] D. Chen, et al., "Generative Inferences Based on Learned Relations," *Cognitive Science*, 41(5), pp. 1062-1092, 2017.

[49] D. Durbin, "The Most Influential Paper Gerard Salton Never Wrote," *Library Trends*, 52(4), pp. 748-764, 2004.

[50] N. Tomuro, "Vector Space and Probabilistic Retrieval Models," DePaul University, 2019. [Online]. Available: https://condor.depaul.edu/ntomuro/courses/575/notes/IR_Models-2.pdf. [Accessed 14 April 2021].

[51] P. Bojanowski, et al., "Enriching word vectors with subword information," *Transactions of the Association for Computational Linguistics*, vol. 5, pp. 135-146, 2017.

[52] T. Mikolov, et al., "Linguistic Regularities in Continuous Space Word Representations," *Conf. North American Chapter of the Association for Computational Linguistics: Human Language Technology*, Atlanta, 2013.

[53] T. Mikolov, et al., "Distributed representations of words and phrases and their compositionality," *Advances in Neural Information Processing Systems*, pp. 3111-3119, 2013.

[54] O. Levy and Y. Goldberg, "Linguistic Regularities in Sparse and Explicit Word Representations," *18TH Conf. on Computational Natural Language Learning*, Ann Arbor, 2014.

[55] E. Kafe, "Fitting Semantic Relations to Word Embeddings," *Global WordNet Conf.*, Wroclaw, 2019.