

Understanding Narratives through Dimensions of Analogy

Thiloshon Nagarajah , Filip Ilievski and Jay Pujara

Information Sciences Institute, University of Southern California

{tnagaraj, ilievski, jpujara}@isi.edu

Abstract

Analogical reasoning is a powerful reasoning tool that enables humans to connect two situations, and to generalize their knowledge from familiar to novel situations. Cognitive Science research provides valuable insights into the richness and complexity of analogical reasoning, together with implementations of expressive analogical reasoners with limited scalability. Modern scalable AI techniques with the potential to reason by analogy have focused on the special case of proportional analogy, and have seldom been applied to understand higher-order analogies. In this paper, we aim to bridge the gap on the task of narrative understanding, by: 1) formalizing six dimensions of analogy based on mature insights from Cognitive Science research, 2) annotating a corpus of fables with each of these dimensions, and 3) defining four tasks with increasing complexity that enable scalable evaluation of AI techniques. Experiments with language models and neuro-symbolic AI reasoners on these tasks reveal that state-of-the-art methods can be applied to reason by analogy with a limited success, motivating the need for further research towards comprehensive and scalable analogical reasoning by AI. We make all our code and data available.

1 Introduction

Understanding narratives requires command of qualitative causal relationships between events, entities, and quantities. Qualitative Reasoning (QR) models have been developed to formalize such relationships in accordance with human mental models [Forbus, 2011], and recent work has demonstrated their ability to model social reasoning [Tomai and Forbus, 2007; Dehghani *et al.*, 2008]. Yet, building scalable cognitive AI architectures requires qualitative representations to be combined with mechanisms that enable efficient and effective generalization and reuse of knowledge, through analogy.

Cognitive Science research dictates that humans develop the ability for analogical reasoning in the first several years of their lives, and rely on this ability to understand, explain, or imagine situations across domains [Holyoak *et al.*, 1995]. In law, practitioners rely on precedents to make decisions in

novel complicated cases. In education, teachers seek familiar situations to capture the student intuition about novel concepts and processes. In politics, leaders craft their arguments based on the lessons learned from impactful events that are commonly known amongst their people. In business, innovators get inspired from the design of processes and objects that have similar function or purpose, being it found in nature or man-made.

Analogical reasoning can manifest in many ways. Some may serve to help describe an object, for example a fire engine may be red like a tomato. Others may be figurative, comparing a lover to a summer's day. Analogies may draw on relationships between entities, such as comparing the Sun and the Earth to the Earth and the Moon, based on a *revolves* relation. Analogies may also describe events such as comparing a natural disaster to a war, or convey causal information, such as the contributors to the Great Depression and the financial crisis. Given the diversity of these analogies, building AI agents to understand and generate analogies must possess a variety of skills for relational awareness.

While a rich body of Cognitive Science work on analogical reasoning exists [Gentner, 1983; Holyoak *et al.*, 1995; Tversky, 1977], modern AI techniques have not been adapted to this challenge at scale. As machine learning is the dominant AI paradigm in the past decade, AI models largely base their predictions on prior experiences. However, they generally lack analogical inference mechanisms, i.e., mechanisms to distill high-level patterns, such as entire situations or events, from exemplar data. This prevents neural methods from learning efficiently with little data and understanding, imagining, and explaining novel situations [Chen *et al.*, 2019]. For example, large language models like BERT [Devlin *et al.*, 2018] and GPT-3 [Brown *et al.*, 2020] struggle with understanding analogies between proportions [Ushio *et al.*, 2021b], and they are likely to perform poorly on more complex analogies between two situations or domains. Qualitative representations [Forbus, 2011] facilitate the continuous modeling of situations, which is essential for narrative understanding, but they require mechanisms to generalize to a wider set of examples. Seminal cognitively inspired architectures [Forbus *et al.*, 1995; Aamodt and Plaza, 1994; Forbus and Hinrichs, 2017] are able to draw analogies between situations, yet, adapting them to understand narratives at scale is an open challenge.

To bridge this gap between cognitive insights and AI algorithms, we develop a novel taxonomy of six dimensions of analogy. We apply this taxonomy to annotate analogies in pairs of fables, and we perform extensive evaluation and analysis of state-of-the-art AI methods across the different analogical categories. Concretely, our contributions are:

1. We motivate and formally define six dimensions of analogy that are essential for AI models to understand and reason qualitatively with in situations such as stories.
2. We annotate 44 story pairs with each of the six analogical dimensions. We also correlate these annotations with literal similarity and emotional arc similarity.
3. We define four analogical tasks on this dataset and evaluate several representative state-of-the-art methods based on language models and frame sequences, noting that these are insufficient to capture variants of analogical similarity.

2 Dimensions of Analogical Reasoning

2.1 Insights from Cognitive Science

People are born without an analogical reasoning skill, yet, this skill has been shown to develop over time with the expansion of knowledge [Gentner, 1988]. Holyoak *et al.* specify three levels of analogy, each tied approximately with human age as a proxy for this knowledge: 1) *Attributive* - A is like B because they share an attribute value (e.g., the ocean is like this car as both as blue). This analogical skill is formed in humans between age 1 and 2 [Gentner and Toupin, 1986]. 2) *Proportional* - A is like B because they have a similar role in their pairwise proportions. For example, car is like sofa because (car-larger than-motorcycle) and (sofa-larger than-chair). This is formed at human age of 3 to 4. 3) *Structural* - A is like B because their structures align, typically due to causality. Humans acquire structural reasoning at 5 - 6 years of age.

According to [Gentner and Markman, 1997], analogy is drawn between two domains or situations. A successful analogy requires three criteria to be fulfilled [Holyoak *et al.*, 1995]. 1) *Systematicity* - the two analogical structures should be aligned - if they are not, comparison is meaningless. 2) *Relational similarity* - the two structures should have relational similarity, i.e., their fillers should play a similar role in the overall structure of the domain or the situation. 3) *Purpose* - as the two structures are not overall very similar, drawing an analogy is a creative process that is driven by a certain goal.

Analogy focuses on relational similarity, and as such it differs from other notions of similarity, such as literal similarity, mere appearance, and anomaly [Tversky, 1977]. *Literal similarity* between two objects A and B is proportional to the intersection of their features and inversely proportional to the features that differ ($A - B$ and $B - A$), where features include both attributes and relational predicates. *Mere appearance* focuses on similarity between attributes and not the relationships (e.g., moon - coin), while anomaly refers to a comparison with no attribute nor relational overlap [Gentner, 1983].

2.2 Our Taxonomy

Guided by the rich body of theoretical work, we devise six dimensions of analogy that we expect to be valid for narrative structures. Namely:

1. Shallow Attribute Analogy (SAA) - Characters in the two situations (here, stories) have explicitly stated *physical* attributes that are similar. For example, given story A where one of the protagonists is a brown fox and story B that features a brown deer, one can draw an analogy between the two stories based on shallow attribute correspondence: brown(fox) \leftrightarrow brown(deer).

2. Deep Attribute Analogy (DAA) - Characters in the two stories have *abstract* attributes that are similar. For example, given story A where an ass is depicted as naive and story B that features a naive deer, one can draw an analogy between the two stories based on deep attribute correspondence: naive(ass) \leftrightarrow naive(deer).

3. Relational Analogy (RA) - Character pairs in the two stories have a similar first-order relationship to each other. For instance, given a story A where a fox and a deer are friends, and story B where a conman is a friend with a countryman, we draw a relational analogy: friendswith(fox,deer) \leftrightarrow friendswith(conman, countryman).

4. Event Analogy (EA) - Two stories involve the same frame or similar event, e.g., both story A and story B involve a dangerous event.

5. Structural Analogy (SA) - At least two events, connected with causal links, can be matched across two stories. For instance, the merchant chases an ass and the ass runs as a result of it in story A; a tiger chases a rabbit and the rabbit runs away as a result in story B. Formally: chase(merchant,ass) & run(ass) & cause(chase,run) \leftrightarrow chase(tiger,rabbit) & run(rabbit) & cause(chase,run).

6. Moral/Purpose (MP) - Two stories have morals that are compatible, and can be aligned with each other. For instance, the moral of story A “Know thyself” corresponds to the moral of story B: “Know your worth”.

Suspecting that, unlike analogy, surface similarity is easily captured by state-of-the-art models, we also consider **Literal Similarity (LS)**, which indicates whether two stories use similar words such as the same character or the same location. For instance, lion in combat with a bull in the jungle \leftrightarrow lion walking with a fox in the jungle.

3 Analogical Reasoning Benchmark

We collected the primary dataset by scraping Aesop Fables from Litscape.com.¹ The dataset had 237 fables and corresponding single-sentence moral except for 36 fables without record of morals. Each fable was short with most stories having 3 - 8 sentences and 50 - 150 words. We performed two iterations of annotation to classify single stories and annotate analogical categories for pairs of stories.

¹<https://www.litscape.com/indexes/Aesop/Morals.html>

3.1 Single Story Annotation

In the first iteration, one annotator manually annotated each fable based on its moral with one of 15 classes. 8 of the classes were taken from the tags defined in the source website, yet most of the fables were not accurately tagged. The annotator introduced 7 more classes such that it best explains the moral. We annotated the moral rather than the story to be able to learn from the moral and generalize to the story. Each story can be annotated with multiple moral classes if they apply. This process resulted in 116 fables with 15 classes: CONSEQUENCE, CONTENT, DANGER, EFFORT, FLATTERY, FRIENDS, GREED, LAZY, LEARN, OPPORTUNITY, RESPECT, TRUE-NATURE, TRUST, WEAK and WORTHINESS. The class distribution of the fables is given in Figure 1.

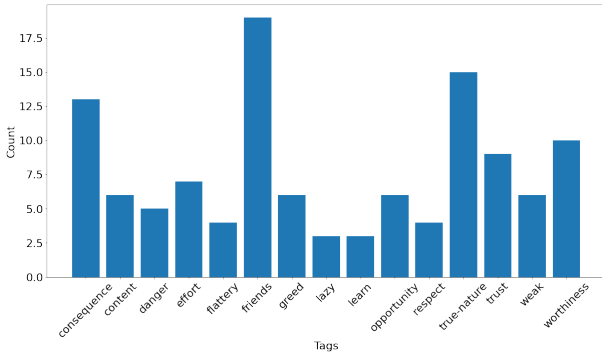


Figure 1: Moral distribution of stories.

3.2 Story Pairs Annotation

In the second iteration, we introduced a framework of annotation under six analogy types, described in the previous section. Each fable could have more than one analogy type and each analogy annotation is binary. Besides annotating the binary value (yes/no) for each analogical type, we also included graph-like triples that justify the positive judgment for each analogical type, like *lion - friends with - ass*. This supporting evidence has to be formed as positive triples, negated triples are not accepted. Furthermore, our guidelines dictate that event analogy is a prerequisite for structural analogy since the structural analogy is comprised of at least two analogical events connected with a causal link. When two fables have opposite morals, we consider MP analogy as false. An example of this situation is 1: *a merchant gets greedy and kills goose, loosing his wealth*, wherein 2: *an ass was content with his food and saved himself from danger*. Both deal with greed, but the first shows the consequence of greed while the second shows a consequence of not being greedy. The full guidelines can be found online.²

To quantify the quality of annotation we computed Inter Annotator Agreement (IAA). We sampled three pairs of stories and three annotators were asked to annotate using the

²<https://docs.google.com/document/d/1OcFD8uSB1HzhYRIR0ubb06wda0daUOBaI7WlaxAXryY/edit?usp=sharing>

Table 1: Cohen Kappa IAA Scores.

Annotator	SAA	DAA	RA	EA	SA	MP	LS
1 VS 2	1.00	-0.5	0.39	1.00	0.00	1.00	0.39
1 VS 3	1.00	1.00	0.00	0.00	0.00	0.00	0.39
2 VS 3	1.00	-0.5	0.00	0.00	-0.5	0.00	1.00

six analogy types. We observed high agreements for SAA and LS, while DAA and SA gave lower agreements. RA, EA and MP were mostly close to random guessing. The primary reason for low agreement scores was that often analogical annotation used implicit attributes or events not included in the short fables, while only explicit annotation would yield very few analogical matches. This signals a need for further discussion and specification of the annotation guidelines and training. The Cohen Kappa IAA score between annotators 1, 2, and 3 is shown in Table 1.

4 Methods

We identified four main tasks to learn or evaluate analogies with the annotations. Each task is built from the results and understanding from the previous tasks and makes use of various characteristics of the annotation process. These tasks range from exploratory data analysis to using transfer learning to assess generality, which provide a robust pipeline for future research and serve as the initial benchmark. In task 1, we model the high level semantic themes of the fables by using their morals, and learn to cluster these thematic analogies to their tags. In task 2, we use different techniques using lexical and semantic structures of fables to generate candidate analogical fable pairs and evaluate the efficacy of these generative methods. Next, in task 3, we evaluate different learning approaches to identify the specific analogy types within pairs. Finally, in task 4, we evaluate the generality of our methods by trying to transfer the learned analogical relationship from a different dataset from a potentially unrelated domain. We present the task details, data, models, and evaluation in detail in the following subsections.

4.1 Task 1: Moral Clustering

Task description: At the most abstract level, analogical reasoning can be approximated by grouping similar items together. We approach this setting by clustering fables to their respective high-level morals. If two stories are about a fox being greedy and a merchant being greedy, first step of deriving this analogy is to identify that both stories belong to the moral tag GREED. Since morals in the dataset are one sentence summary of the fables, we experimented with methods to cluster the morals to one of 15 classes. To build these methods we used different characteristics of the story-moral pairs as below:

Language model baseline: Recent breakthroughs in NLP use contextual-language models to produce embeddings that represent documents. Word level language models have shown the ability to identify analogies like *man:woman::king:queen*, but work on sentence and document level modeling has mostly focused on NLI or paraphrase detection. We finetuned a RoBERTa [Liu *et al.*, 2019] model on the morals and their tags, and retrieved [CLS] hidden layer

of the morals as a sentence representation. This yielded poor performance as the number of classes were too large relative to the size of the dataset and some classes had very few training instances. As an alternative, we experimented with one-vs-one and one-vs-all classifiers. Embedded morals were passed to a 3-layer neural network with softmax output using a cross entropy loss function. We observed a low evaluation accuracy, suggesting such models are too complex for our limited data. Finally we built 15 one-vs-all logistic regression classifiers for each tag which performed better than neural network. After 100 iterations of under sampled one-vs-all classifiers resulted in accuracies ranging from 59% to 64% but a low F1 score of 0.3. Manual inspection of the variety of the moral tags suggest these tags did not adequately represent fables, leading to a different methodology for subsequent tasks.

Frame baseline: We observed that the stories under certain tags have similar common frames or frame sequences. For example, fables under FRIENDSHIP have *Travel-Collaboration* and fables under TRUST have *Manipulation* as common frames. We also noticed that stories tend to follow certain similar sequence of actions which could be captured with a sequence model. But to generalize the types of these actions, we choose to use frames parsed from FrameNet [Baker *et al.*, 1998] using the open-sesame frame parser [Swayamdipta *et al.*, 2017]. To answer the question whether the sequence of frames or even existence of frames could differentiate the type of the fable, we built different methods to classify morals. The simplest notion was to consider the number of occurrence of a frame and bi-grams of frames as numeric feature and learn with a neural network model. Then we considered sequence of frames embedded with RoBERTa as input for a Bi-LSTM model [Hochreiter and Schmidhuber, 1997]. We considered 376 frames found in all the stories as well as only the most common 15 frames per tag as features.

Results: The neural network classifier to predict one of 15 tags from frame count performed poorly. When all 376 frames were considered as network features, accuracy of 0.10 was attained. Using only the most common 15 frames increased accuracy slightly to 0.12. With under-sampled data, an one-vs-all classifier performed better than random guessing though F1 score was still low (0.35). Accuracies of the selected one-vs-all classifiers are given in Table 2. We leave it to subsequent work to analyze to what extent are the obtained results influenced by noise in the frame parser, noise in the annotations, or the usefulness of the frame representations.

4.2 Task 2: Analogical Pair Generation

Task description: Our analogy framework was built to compare analogies in pairs of stories. But selection of stories to be paired cannot be done arbitrarily. The rationale behind choosing a pair of story has to be evaluated to be able to understand the implications of the method. Different sampling methods could affect the ability to learn analogy from fables. For instance, lexical methods could find the fables that share similar words, yet might not be a great method to find analogies. A semantic or structure bases method could do well with

Table 2: Accuracies of one-vs-all neural classifiers.

Tag	Accuracy
Consequence	0.64
Content	0.64
Danger	0.62
Effort	0.51
Flattery	0.59
Friends	0.54
Greed	0.64
Opportunity	0.59

Table 3: Analogies of story pairs.

	maximum	Semantic	Frames	Lexical	Random
Story count	44	9	10	12	13
Method average	6	2.22	1.90	3.00	1.54
SAA	1	0.00	0.00	0.08	0.08
DAA	1	0.44	0.60	0.58	0.31
RA	1	0.56	0.40	0.58	0.54
SA	1	0.22	0.10	0.25	0.00
MP	1	0.11	0.10	0.42	0.23
LS	1	0.33	0.00	0.50	0.15
SSS	1	0.56	0.70	0.58	0.23

identifying the similar buildup of stories which does not necessarily need to have similar lexical features. In this task, we define the different ways pairs could be created and methods to evaluate their efficacy.

Language model baselines: To build story pairs, we experimented with different sampling techniques under lexical, semantic, or structural methods. We noticed that certain stories have same characters and same interactions between characters, which could be captured well with lexical methods. But since we are more interested in the implicit analogies in the story rather than the explicitly stated similar words, we also considered semantic and frame similarities. For each story, we picked the most similar story using these methods and generated 348 story pairs. We added another 348 pairs of stories that were randomly selected to a final total of 696 story pairs. Implemented sampling techniques were as follows:

1. Lexical - We converted each word in the fable to Word2Vec [Mikolov *et al.*, 2013a] embedding and calculated the weighted average of the story with TF-IDF. We used cosine similarity between these story-level embedding to find the most matching fable pairs.
2. Semantic - We converted each fable to a RoBERTa [Liu *et al.*, 2019] embedding and retrieved the embedding value for the [CLS] token. We used cosine similarity on top of this embedding.
3. Frame - We parsed frames in each fable with open-sesame and calculated sentence edit distance between the frames to find the most similar frame sequence. Sentence edit distance is adopted from word edit distance algorithm by considering each word as a character in the original algorithm. Frame edit distance was scaled to fit different frame lengths. To have high variance, we kept only frames found in story 1 as features for edit distance.

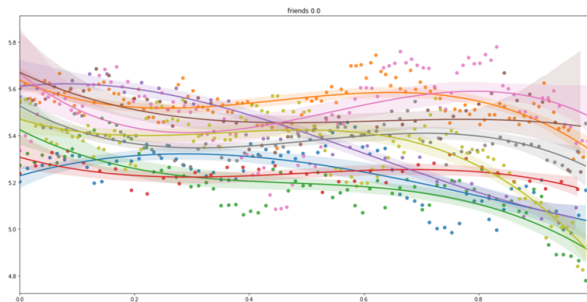


Figure 2: Fables with a story shape High - High - Low.

Story shape baseline: In addition to language models, we considered story shapes [Reagan *et al.*, 2016] to generate story pairs. Story shapes are arcs generated from sentiment scores of the fable as the story progresses. Though these arcs do not explicitly provide the information about the plot of stories, they give an overall characteristic curve of the emotions in the story. Table 4 lists the prototypical story shapes identified by Reagan *et al.*. Using the definition of these curves, we can tag the stories which have similar arcs as similar stories. We analysed how emotions in the story change over the duration of the story to classify the story into one of five common story shapes, using the arc types in prior work. To calculate emotional changes, we used the MTLabs [Dodds *et al.*, 2011a] vocabulary to assign happiness scores (hedonometric score) for each word in the story. A sliding window of 30 words was used to calculate the average of the happiness within the window and averages throughout the story was recorded. The HIGH or LOW points are calculated as average of the beginning (first 30% of the story), middle (middle 40%) and end (last 30%) of the story. The averages of the 30-word window typically lie between a hedonometric score of 5 to 7 [Dodds *et al.*, 2011b] and for fables the average of all the stories was 5.4. Based on the assumption 5.4 denotes a neutral tone, HIGH was taken as an average of above 5.4+0.2 and LOW as an average less than 5.4-0.2.

Results: We evaluated the strength of the 3 language model based sampling methods by assigning 0 or 1 for analogies and summing up over the methods. The scores are presented in table 3. The Word2Vec-based lexical method was found to have the highest analogy score overall. Frames based method gives high analogical similarity for DAA. The MP similarity, which is the most important analogy for this research was high with lexical sampling. Even the highest analogical scores were between 0.4 and 0.6, demonstrating that the most similar stories under different sampling methods do not necessarily share comparable analogies. It is also worth noting that RA and DAA was comparatively high even for the random sampling method, showing how any fable could be justified to have an DAA or RA analogy mainly due to the nature of the fable stories. This cautions of a possible high annotator bias that needs to be addressed with strict and elaborate annotation guidelines.

We further analyzed how the analogy types are correlated with each other and presented in Figure 3. We observed that MP is mostly correlated with DAA and that DAA was also

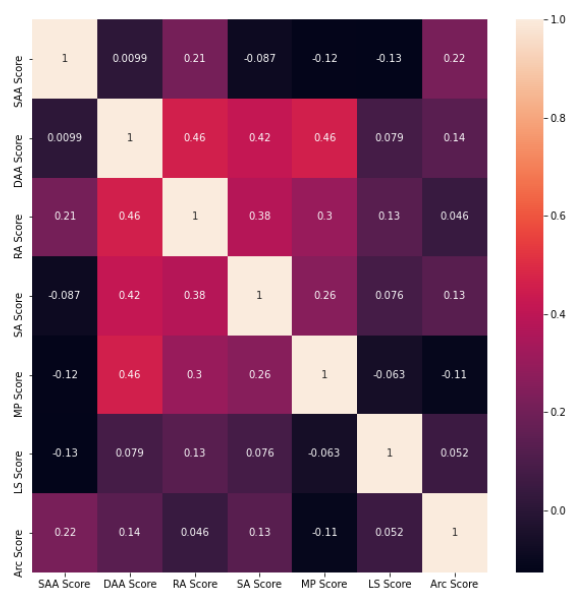


Figure 3: Correlation of analogies

Table 4: Exemplar story shapes.

Fable Arc	Arc type
High - High - Low (Figure 2)	Tragedy
Low - Middle - High	Rags to Riches
High - Low - High	Cinderella
Low - Middle - High	Rags to Riches
Low - Middle - Low	Oedipus

highly correlated with SA and RA. We note that the correlation scores were only as high as 0.46 and most of the correlations were under 0.2, showing how the analogies are not as correlated as one would expect to be.

The classifier to learn from story shapes yielded poor accuracy showing story arcs alone cannot be used to identify similar stories.

4.3 Task 3: Analogy Type Prediction

Task description: Building up from Task 2 of generating similar story pairs, the next natural step is to be able to predict analogy types. We could learn from the pairs of stories generated by different sampling methods to predict whether an analogical dimension holds for a story pair or not.

Baseline: We tuned a cross-encoder DistilRoBERTa-base [Sanh *et al.*, 2019] model to identify analogies. We formulated the problem as binary classification and built 7 classifiers: one for each of the six analogy types and one for literal similarity.

Results: Table 5 shows the obtained results together with the ration of the positive class for each analogical class. We also report F1-scores to account for the imbalance of certain classes. The cross encoder obtained high F1 performance for SAA, DAA, and RA. Yet, it performed only slightly better than random guessing for other analogy types. For analogy types with a balanced class distribution, the classifier

Table 5: Performance of Analogy Type Classifiers.

Analogy	Positive Class Ratio	Accuracy	F1
SAA	0.05	1.0	1.0
DAA	0.42	0.77	0.67
RA	0.51	0.77	0.77
SA	0.08	0.66	0.4
MP	0.22	0.77	0.43
LS	0.28	0.88	0.47

was able to result in high F1 score, and analogy types with skewed classes performed poorly. Since this experiment was performed only on 44 annotated pairs, we are optimistic that having a larger dataset may yield better performing models.

Below we provide an example of a story pair that has Deep Attributive, Relational, and Event analogies as identified by our classifier. Both soldier in story 1 and Hen in story 2 are caring, which is a deep attributive analogy. Their care and act of caring for horse or egg is relational and event analogies. These stories does not share a moral purpose as in story 1 the moral would be that *fickle care causes waste*, wherein story 2 it is to *be careful who you care for*.

Story 1: A HORSE SOLDIER took great pains with his charger. As long as the war lasted, he looked upon him as his fellow-helper in all emergencies, and fed him carefully with hay and corn. But when the war was over, he only allowed him chaff to eat and made him carry heavy loads of wood, and subjected him to much slavish drudgery and ill-treatment. War, however, being again proclaimed, the Soldier put on his charger its military trappings, and mounted, being clad in his heavy coat of mail. The Horse fell down straightway under the weight, no longer equal to the burden, and said to his master: You must now e'en go to the war on foot, for you have transformed me from a Horse into an Ass.

Story 2: A HEN finding the eggs of a viper, and carefully keeping them warm, nourished them into life. A Swallow, observing what she had done, said: You silly creature! Why have you hatched these vipers, which, when they shall have grown, will surely inflict injury on all of us, beginning with yourself?

4.4 Task 4: Analogical Transfer Learning

Task description: To evaluate the robustness of our methods, we design a transfer learning experiment, where the objective is to build a baseline that can learn analogies from one dataset and be able to sufficiently transfer that knowledge to a different dataset.

Baselines: We experimented with the following pipelines:

1. Fable2Fable - We used the 116 fables with their tags from our first annotation iteration to build a corpus of 544 story pairs whose morals belong to the same tag. To create a balanced dataset, we added 544 more story pairs that do not belong to same tag as negative samples. We finetuned a pre-trained cross encoder DistilRoBERTa model to predict whether the story pair belongs to the same tag. We evaluated the model on the 44 story pairs annotated with analogy types by predicting if the pair of story is analogical for each of the seven types.
2. AILA2Fable - We used the AILA [Bhattacharya *et al.*, 2019] dataset which contains legal cases and their precedent cases to build a positive and negative pairs. We identified 640 pairs of similar cases as positive samples and added another 640 unrelated cases as negative samples. As the legal cases are lengthy, we sampled 500

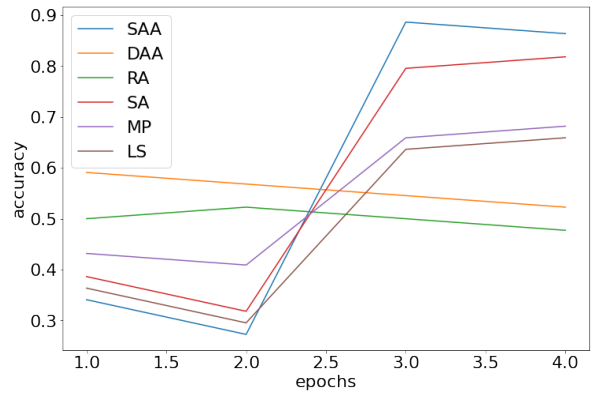


Figure 4: Accuracy of Fable2Fable.

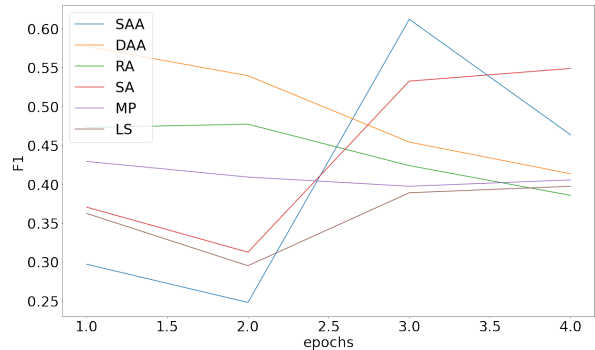


Figure 5: F1 score of Fable2Fable.

words from the middle of the document to support the Transformer-based cross encoder. The remaining training and evaluation setup was same as in Fable2Fable.

Results: Fable2Fable was built using DistilRoBERTa [Sanh *et al.*, 2019] cross encoder and we observed a substantial prediction power in SAA, SA, MP, and LS, with SAA reaching a highest accuracy of 88%, and SA reaching the highest F1-score of 55%. Curiously, we observed that the model predictions changed dramatically based on its training over the epochs. The changes in accuracy and F1 scores across the epochs are given in Figures 4 and 5, respectively. We note that SAA, SA, MP and LS performed better with more epochs, the performance of predicting DAA and RA analogy was affected negatively. One possible explanation for this different behavior across classes may lie in the class distributions, as DAA and RA are almost perfectly balanced whereas SA, MP and LS have low positive class ratios of 0.1, 0.2 and 0.23, respectively.

For the AILA2Fable pipeline, DistilRoBERTa was not able to learn sufficiently and we used a pre-trained model that was built for the paraphrasing task (paraphrase-MiniLM-L6-v2). We omit the results from this experiment, as the accuracy and the F1-score per class were close to random guessing. This signals that, while transferring labels for the same domain can be achieved to some extent with pre-trained language models, generalization from an unrelated domain is more difficult and may require a different architecture.

5 Related Work

Analogies in cognitive science Cognitive systems typically consist of multiple phases: retrieval, mapping, abstraction, and representation [Gentner and Forbus, 2011]. The structural mapping engine [Falkenhainer *et al.*, 1989; Forbus *et al.*, 2017] maps two structures to each other by combining local with structural similarity, ultimately being able to find multiple analogies for a given situation. Similarly, the MAC-FAC (Many Are Called but Few Are Chosen) [Forbus *et al.*, 1995] system starts with a cheap retrieval of possible analogies, followed by a structure mapping step. The Companion cognitive architecture [Forbus and Hinrichs, 2017], which leverages a subset of CYC [Lenat, 1995], strives to capture analogical and qualitative reasoning, and has been proven useful for textbook problem solving, moral decision-making, and commonsense reasoning tasks. Case-based reasoning (CBR) builds novel strategies based on prior experience through four steps: retrieval of similar cases, their reuse into a proposed solution, revision of this solution, and retention of the updated solution in the collection of prior cases [Aamodt and Plaza, 1994]. Unlike analogical reasoning, CBR is not meant to be used broadly, and it is typically meant to perform a specific task well, thus trading generality for efficiency or other performance measures [De Mantaras *et al.*, 2005]. Recent work has used case-based reasoning to answer knowledge base queries, by retrieving subgraphs that correspond to structurally similar queries and reasoning over them [Das *et al.*, 2022]. The development of these cognitively inspired architectures is a promising complementary direction. While it is currently unclear how to apply these methods on tasks based on natural language narratives such as the ones in our work, we expect that future research may be able to bridge this gap.

Analogies in AI and NLP Several benchmarks for analogical reasoning exist in AI and NLP, e.g., based on psychometric analogy tests [Mikolov *et al.*, 2013b; Gladkova *et al.*, 2016]. These benchmarks focus on capturing the pairwise relations between two concepts (e.g., king - man), sometimes exemplified through the relation between another pair of concepts (e.g., queen - woman). Relational analogies can be learned to some extent from data, as shown in a recent study [Ushio *et al.*, 2021a], where BERT, RoBERTa, and GPT-3 have been evaluated on linguistic and cognitive assessment datasets equivalent to analogy problems on the Scholastic Aptitude Test (SAT). Structural analogies have been rarely employed in recent AI and NLP work [Lu *et al.*, 2019]. Elson and McKeown used frames from FrameNet to build a corpus with paraphrased Aesop fables by using static patterns. While such paraphrased stories can be seen as analogical, we expect that paraphrased story pairs can be easily picked up by latest language models. Efforts that combine analogical reasoning and language modeling exist [Ribeiro and Forbus, 2021], yet, it is unclear how to adapt them to the task of understanding narratives. In our study, we focus on six dimensions of analogy, thus targeting high-level structural and relational similarity, and explore how language models and semantic frames can come together to understand narratives.

6 Conclusions

This paper presented a framework for analogical reasoning by AI techniques. Guided by Cognitive Science theories, we devised six analogical dimensions with increasing complexity: shallow and deep attribute analogy, relational analogy, event analogy, structural analogy, and moral/purpose, and we considered literal similarity as a baseline dimension. We annotated fable narratives with moral tags and the existence of pairwise story analogy, resulting in a small comprehensive evaluation corpus. We defined four tasks over this data: clustering morals, analogical pair generation, analogical type prediction, and analogical transfer, each revolving around various analogical dimensions. We evaluated several representative AI techniques based on language models and semantic frames on these four tasks, observing that these techniques are applicable across tasks, yet, their ability to reason by analogy is limited, especially when this requires understanding of complex structures. We make all code and data available to facilitate subsequent work on analogical reasoning in AI: <https://github.com/usc-isi-i2/analogical-transfer-learning>.

Based on our results, we anticipate that explicit modeling of continuous qualitative aspects of narratives such as time, space, and quantity, holds the promise to improve the narrative understanding by AI through analogy in the future. Semantic natural language systems that automatically produce qualitative descriptions [Friedman *et al.*, 2021] provide an intuitive opening to integrate qualitative representations with analogical reasoning models. Future work should also develop scalable structure-aware methods for analogical reasoning, possibly by building on prior cognitive architectures. Finally, future work should aim to create an evaluation dataset that is orders of magnitude larger, and investigate clever ways to acquire large-scale training data for model development.

References

- [Aamodt and Plaza, 1994] Agnar Aamodt and Enric Plaza. Case-based reasoning: Foundational issues, methodological variations, and system approaches. *AI communications*, 7(1):39–59, 1994.
- [Baker *et al.*, 1998] Collin F. Baker, Charles J. Fillmore, and John B. Lowe. The Berkeley FrameNet project. In *COLING-ACL '98: Proceedings of the Conference*, pages 86–90, Montreal, Canada, 1998.
- [Bhattacharya *et al.*, 2019] Paheli Bhattacharya, Kripabandhu Ghosh, Saptarshi Ghosh, Arindam Pal, Parth Mehta, Arnab Bhattacharya, and Prasenjit Majumder. Overview of the FIRE 2019 AILA track: Artificial intelligence for legal assistance. In Parth Mehta, Paolo Rosso, Prasenjit Majumder, and Mandar Mitra, editors, *Working Notes of FIRE 2019 - Forum for Information Retrieval Evaluation, Kolkata, India, December 12-15, 2019*, volume 2517 of *CEUR Workshop Proceedings*, pages 1–12. CEUR-WS.org, 2019.
- [Brown *et al.*, 2020] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot

- learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.
- [Chen *et al.*, 2019] Kezhen Chen, Irina Rabkina, Matthew D McLure, and Kenneth D Forbus. Human-like sketch object recognition via analogical learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 1336–1343, 2019.
- [Das *et al.*, 2022] Rajarshi Das, Ameya Godbole, Ankita Naik, Elliot Tower, Robin Jia, Manzil Zaheer, Hannaneh Hajishirzi, and Andrew McCallum. Knowledge base question answering by case-based reasoning over subgraphs. *arXiv preprint arXiv:2202.10610*, 2022.
- [De Mantaras *et al.*, 2005] Ramon Lopez De Mantaras, David McSherry, Derek Bridge, David Leake, Barry Smyth, Susan Craw, Boi Faltings, Mary Lou Maher, MICHAEL T COX, Kenneth Forbus, et al. Retrieval, reuse, revision and retention in case-based reasoning. *The Knowledge Engineering Review*, 20(3):215–240, 2005.
- [Dehghani *et al.*, 2008] Morteza Dehghani, Emmett Tomai, Kenneth D Forbus, and Matthew Klenk. An integrated reasoning approach to moral decision-making. In *AAAI*, pages 1280–1286, 2008.
- [Devlin *et al.*, 2018] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.
- [Dodds *et al.*, 2011a] Peter Sheridan Dodds, Kameron Decker Harris, Isabel M. Kloumann, Catherine A. Bliss, and Christopher M. Danforth. Temporal patterns of happiness and information in a global social network: Hedonometrics and twitter. *PLoS ONE*, 6(12):e26752, dec 2011.
- [Dodds *et al.*, 2011b] Peter Sheridan Dodds, Kameron Decker Harris, Isabel M. Kloumann, Catherine A. Bliss, and Christopher M. Danforth. Temporal patterns of happiness and information in a global social network: Hedonometrics and twitter. *PLOS ONE*, 6(12):1–1, 12 2011.
- [Elson and McKeown, 2009] David Elson and Kathleen McKeown. Extending and evaluating a platform for story understanding. pages 32–35, 01 2009.
- [Falkenhainer *et al.*, 1989] Brian Falkenhainer, Kenneth D Forbus, and Dedre Gentner. The structure-mapping engine: Algorithm and examples. *Artificial intelligence*, 41(1):1–63, 1989.
- [Forbus and Hinrichs, 2017] Kenneth D. Forbus and Thomas R. Hinrichs. Analogy and qualitative representations in the companion cognitive architecture. 2017.
- [Forbus *et al.*, 1995] Kenneth D Forbus, Dedre Gentner, and Keith Law. Mac/fac: A model of similarity-based retrieval. *Cognitive science*, 19(2):141–205, 1995.
- [Forbus *et al.*, 2017] Kenneth D. Forbus, Ronald W. Ferguson, Andrew M. Lovett, and Dedre Gentner. Extending
- some to handle large-scale cognitive modeling. *Cognitive science*, 41 5:1152–1201, 2017.
- [Forbus, 2011] Kenneth D Forbus. Qualitative modeling. *Wiley Interdisciplinary Reviews: Cognitive Science*, 2(4):374–391, 2011.
- [Friedman *et al.*, 2021] Scott E Friedman, Ian H Magnusson, and Sonja M Schmer-Galunder. Extracting qualitative causal structure with transformer-based nlp. *arXiv preprint arXiv:2108.13304*, 2021.
- [Gentner and Forbus, 2011] Dedre Gentner and Kenneth D Forbus. Computational models of analogy. *Wiley interdisciplinary reviews: cognitive science*, 2(3):266–276, 2011.
- [Gentner and Markman, 1997] Dedre Gentner and Arthur B Markman. Structure mapping in analogy and similarity. *American psychologist*, 52(1):45, 1997.
- [Gentner and Toupin, 1986] Dedre Gentner and Cecile Toupin. Systematicity and surface similarity in the development of analogy. *Cognitive science*, 10(3):277–300, 1986.
- [Gentner, 1983] Dedre Gentner. Structure-mapping: A theoretical framework for analogy. *Cognitive science*, 7(2):155–170, 1983.
- [Gentner, 1988] Dedre Gentner. Metaphor as structure mapping: The relational shift. *Child development*, pages 47–59, 1988.
- [Gladkova *et al.*, 2016] Anna Gladkova, Aleksandr Drozd, and Satoshi Matsuoka. Analogy-based detection of morphological and semantic relations with word embeddings: what works and what doesn’t. In *Proceedings of the NAACL Student Research Workshop*, pages 8–15, 2016.
- [Hochreiter and Schmidhuber, 1997] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Comput.*, 9(8):1735–1780, nov 1997.
- [Holyoak *et al.*, 1995] Keith J Holyoak, Paul Thagard, and Stuart Sutherland. Mental leaps: analogy in creative thought. *Nature*, 373(6515):572–572, 1995.
- [Lenat, 1995] Douglas B Lenat. Cyc: A large-scale investment in knowledge infrastructure. *Communications of the ACM*, 38(11):33–38, 1995.
- [Liu *et al.*, 2019] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692, 2019.
- [Lu *et al.*, 2019] Hongjing Lu, Ying Nian Wu, and Keith J Holyoak. Emergence of analogy from relation learning. *Proceedings of the National Academy of Sciences*, 116(10):4176–4181, 2019.
- [Mikolov *et al.*, 2013a] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space, 2013.
- [Mikolov *et al.*, 2013b] Tomáš Mikolov, Wen-tau Yih, and Geoffrey Zweig. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013*

conference of the north american chapter of the association for computational linguistics: Human language technologies, pages 746–751, 2013.

- [Reagan *et al.*, 2016] Andrew J. Reagan, Lewis Mitchell, Dylan Kiley, Christopher M. Danforth, and Peter Sheridan Dodds. The emotional arcs of stories are dominated by six basic shapes. *CoRR*, abs/1606.07772, 2016.
- [Ribeiro and Forbus, 2021] Danilo Neves Ribeiro and Kenneth Forbus. Combining analogy with language models for knowledge extraction. In *3rd Conference on Automated Knowledge Base Construction*, 2021.
- [Sanh *et al.*, 2019] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *ArXiv*, abs/1910.01108, 2019.
- [Swayamdipta *et al.*, 2017] Swabha Swayamdipta, Sam Thomson, Chris Dyer, and Noah A. Smith. Frame-semantic parsing with softmax-margin segmental rnns and a syntactic scaffold, 2017.
- [Tomai and Forbus, 2007] Emmett Tomai and Ken Forbus. Plenty of blame to go around: a qualitative approach to attribution of moral responsibility. Technical report, NORTHWESTERN UNIV EVANSTON IL, 2007.
- [Tversky, 1977] Amos Tversky. Features of similarity. *Psychological review*, 84(4):327, 1977.
- [Ushio *et al.*, 2021a] Asahi Ushio, Luis Espinosa Anke, Steven Schockaert, and José Camacho-Collados. BERT is to NLP what alexnet is to CV: can pre-trained language models identify analogies? *CoRR*, abs/2105.04949, 2021.
- [Ushio *et al.*, 2021b] Asahi Ushio, Luis Espinosa Anke, Steven Schockaert, and Jose Camacho-Collados. BERT is to NLP what AlexNet is to CV: Can pre-trained language models identify analogies? In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3609–3624, Online, August 2021. Association for Computational Linguistics.