Benchmarking Semantic Capabilities of Analogy Querying Algorithms

Christoph Lofi, Athiq Ahamed, Pratima Kulkarni, Ravi Thakkar
Technische Universität Braunschweig
38106 Braunschweig, Germany
lofi@ifis.cs.tu-bs.de

Abstract. Enabling semantically rich query paradigms is one of the core challenges of current information systems research. In this context, due to their importance and ubiquity in natural language, analogy queries are of particular interest. Current developments in natural language processing and machine learning resulted in some very promising algorithms relying on deep learning neural word embeddings which might contribute to finally realizing analogy queries. However, it is still quite unclear how well these algorithms work from a semantic point of view. One of the problems is that there is no clear consensus on the intended semantics of analogy queries. Furthermore, there are no suitable benchmark dataset available respecting the semantic properties of real-life analogies. Therefore, in this, paper, we discuss the challenges of benchmarking the semantics of analogy query algorithms with a special focus on neural embeddings. We also introduce the AGS analogy benchmark dataset which rectifies many weaknesses of established datasets. Finally, our experiments evaluating state-of-the-art algorithms underline the need for further research in this promising field.

Keywords: query processing, human-centered information systems, benchmarking, analogy processing, relational similarity, semantics of natural language.

1 Introduction

The increasing spread of the Web and its multitude of information systems call for the development of novel interaction and query paradigms in order to keep up with the ever growing amount of information. These paradigms require more sophisticated capabilities compared to established declarative SQL-style or IR-style keyword queries. Especially *human-centered* query paradigms, i.e. query paradigms which try to mimic parts of natural human communication as for example questions answering and verbose queries require sophisticated semantic processing. A central pattern in human communication which has received only little attention by the information systems research community are analogy queries [1]: in natural speech, analogies allow for communicating dense information easily and naturally by exploiting the semantic capabilities and knowledge of both communication partners. Basically, analogies can be used to map factual and behavioral properties from one (usually better known concept) to another (usually less well known) concept by using different types of similarity assertions, therefore transferring the semantic "essence" from one to another while dropping less important differences for the sake of brevity and simplicity. This is particularly effective for explaining and teaching (e.g., "The Qur'an is the 'Islam Bible'"), but can also be used for querying when only vague domain knowledge is available ("I loved my last vacation in Hawai'i. What place would be similar in East Asia?"). Analogical thinking plays such an important role in many human cognitive abilities that it has been suggested by psychologist and linguists that analogies are the "core of cognition" [2] or even the "thing that makes us smart" [3].

However, adapting this valuable concept of natural communication into information systems proves to be very challenging: on one hand, semantics of analogies very hard to grasp algorithmically as analogy processing heavily relies on human perception, abstract inference, and common knowledge. But furthermore, algorithms which claim to be able to mimic analogical reasoning are hard to evaluate due to the lack of benchmark sets and Gold standards. In its simplest form, the core of analogy processing is measuring *relational similarity* between two pairs of words, e.g., using the example from above, one could say that the Qur'an fulfills a similar role/relation for the Islam religion as does the Bible for the Christian belief. This example also highlights one of the challenges of capturing analogy semantics: of course, the role of the Qur'an is slightly different to the role of the Bible when examined in detail, but still similar enough for explaining either concept in a general discussion. While there are several datasets which are frequently used in researching analogy semantics between word pairs, they usually ignore the *definiteness* of relationships*,* a second core component of analogy semantics. The definiteness directly affects the usefulness of an analogy for transferring semantics during communication (e.g., the statement 'funny is to humorous as beautiful is to attractive' has a high degree of relational similarity as both words pairs share the same relationship "is a near-synonym of", but still this analogy is not useful for describing either concept as synonymy does capture the semantic essence.) Therefore, in this paper, we discuss the challenge of benchmarking analogy algorithms in detail:

- We define and discuss different properties of analogy semantics, and highlight their importance for the benchmarking process.
- We provide a brief survey of current state-of-the-art analogy algorithms, and highlight their different base assumptions.
- We introduce a new test set for benchmarking analogy algorithms. Our test set is systematically built by expanding existing benchmark sets, and by also incorporating crowdsourcing judgements in order to capture the human aspect of analogy semantics. While we do not seek to fully replace established benchmark datasets, our dataset introduces new qualities not exhibited by previous benchmarks. Especially, we provide a balanced set of test challenges with a wide range of different analogy challenges. This allows to analyse strength and weaknesses of different algorithms in on a more fine-grained level.
- As a proof of concept, we showcase and discuss the evaluation results for two current state-of-the art algorithms using our benchmark test collection, and briefly discuss the implications of the respective results.

2 Foundations and Related Work

2.1 Analogy Semantics and Analogy Queries

Due to the ubiquity and importance of analogies in daily speech, there is long-standing interest in researching the foundations of analogy semantics in the fields of philosophy, linguistics, and in the cognitive sciences, such as [6], [7], or [8]. There have been several models for analogical reasoning in these fields, as for example very early definitions from the Greek philosophers Plato and Aristotele who propose a rather hard to formalize definition based on *shared abstractions* of two concepts [8], while the 18th century philosopher Kant defines an analogy as two pairs of concepts being connected by *identical relationshi*ps [9]. Other approaches see analogies as a variant of formal logics, i.e. analogy is seen as a special case of *induction* [8] or for performing *hidden inductions* [10]. Another popular model for analogies stems from the field of contemporary cognitive sciences and clarifies some of the vague concepts of Aristotle's view on analogies, and is commonly known as the *structure mapping theory* [11]. Structure mapping is assuming that knowledge is explicitly provided in form of propositional networks of nodes and predicates and claims that there is an analogy whenever large parts of the structural representation of relationships and properties of one object (the source) can be mapped to the representation of the other object (the target). This model resulted in several theoretical computational models, e.g. [12].

The aforementioned analogy definitions are rather complex and hard to grasp computationally. Therefore, most recent works on computational analogy processing rely on the simple *4-term analogy* model which is an extension of the analogy model given by Kant. Basically, a 4-term analogy is given by two sets of word pairs (the so-called *analogons*), with one pair being the source and one pair being the target. A 4-term analogy holds true if there is a high degree of *relational similarity* between those two pairs. This is denoted by $[a_1 : a_2] :: [b_1 : b_2]$, where the relation between a_1 and a_2 is similar to the relation between b_1 and b_2, as for example in $[Qur'an, Muslim] :: [Bible, Christian]$.

This model has several limitations and shortcomings, as we discuss in detail in [1]. For example, the actual semantics of "a high degree of relational similarity" from an ontological point of view is quite unclear, and many frequently used analogies cannot be mapped easily to the 4-term model (as for example the Rutherford analogy which sets a simplified model of atoms in relation to a simplified model of the solar system). Furthermore, the model ignores human perception and abstractions (e.g., the validity of analogies can change over time or even between different communication partner with different background knowledge). Still, this model for analogies is quite popular in computational analogy and linguistic research as it is easy to benchmark, and there exist several recent techniques which can approximate simple relational similarity quite well.

Therefore, in this paper, we argue for an improved interpretation of the 4-term analogy model which we introduced in [13]. The intuition underlying this model is that, basically, there can be multiple relationships between the concepts of an analogon.

However, not all of them are relevant for a semantically meaningful analogy– and furthermore, some of them should even be ignored. Therefore, the model introduces the set of *defining relationships*, and an analogy holds true if the defining sets of both analogons are relational similar. For illustrating the difference and importance of this change in semantics, consider the analogy statement $[Tokyo,\ Japan] :: [Braunschweig,\ Germany]$. Tokyo is a city in Japan, and Braunschweig is a city in Germany, therefore both analogons contain the same "city is located in country" relationship (and this could be considered a valid analogy with respect to the simple 4-term analogy model). Still, this is a poor statement from a semantic point of view because Braunschweig is not like Tokyo at all (therefore, this statement does neither describe the essence of Tokyo nor that of Braunschweig particularly well): the defining traits (relationships) of Tokyo in Japan should at least cover that Tokyo is the single largest city in Japan, and also its capital. There are many other cities which are also located in Japan, but only Tokyo has these two defining traits. Braunschweig, however, is just a smaller and rather unknown city in Germany, and there is nothing particularly special about it (therefore, the defining relationships of both word pairs are not very similar). The closest match to a city like Tokyo in Germany should therefore be Berlin, which is also the largest city and the capital city. Understanding which relationships actually define the essence of an analogon from the viewpoint of human perception is a very challenging problem, but this understanding is crucial for judging the usefulness and value of an analogy statement. Furthermore, the definiteness may vary with different contexts (e.g., the role of Tokyo in Japan in a general discussion vs. the role of Tokyo in Japan in a discussion about fashion trends: here Hamburg/Germany might be a better match than Berlin as Hamburg is often considered the fashion capital of Germany as Tokyo is the fashion capital of Japan).

In short, there can be better or worse analogies based on two core factors (we will later encode the combined overall quality with an *analogy rating*): the *definiteness* of the relationships shared by both analogons (i.e. are the relationships shared between both analogons indeed the defining relationships which describe the *intended semantic essence*), and the *relational similarity* of the shared relationships.

To further clarify the concept of definiteness, consider the analogon $[Bordeaux\ , France]$. Confronted with this word pair and asked for the most obvious relationships between 'Bordeaux' and 'France', most people would answer "Bordeaux is France's city of wine". Therefore, the analogy $[Bordeaux,\ France] :: [Nappa,\ United\ States]$ has a high degree of definiteness, as Nappa is also one of the most famous wine cities of the US. In contrast, $[Bordeaux,\ France] :: [Dallas,\ United\ States]$ would have a low definiteness: both analogons contain the same "city in country" relationship, and even more, both are indeed the 9th largest city of their respective country, but still, these relationships would not be the ones which come to people's mind – they are not an "analogy essence".

Based on these observations, we define three basic types of analogy queries (loosely adopted from [13]) which can be used for analogy-enabled information systems:

- *Analogy confirmation* $?: [a_1, a_2] :: [b_1, b_2]$
  This query checks if the given analogy statement is true, i.e. it checks if the defining relationships are similar enough (from a consensual human perspective).

- *Analogy completion* $?: [a_1, a_2] :: [b_1, ?]$
  This query can be used to find the missing concept in a 4-term analogy. (e.g., "What is for the Islam as is the Bible for the Christians?"). This is therefore the most useful query type in a future analogy-enabled information systems [1]. Solving this query requires identifying the set of defining relationships between a_1 and a_2, and then finding a b_2 such that the set of defining relationships between b_1 and b_2 is similar.
- *Analogon ranking* $?: [a_1, a_2] :: ?$
  Given a single analogon, this query asks for a ranked list of potential analogons which would result in a valid analogy such that the defining relationships between both analogons is similar. This query is significantly harder from a semantic perspective than completion queries, as there is less information available with respect to the nature of the defining relationships (as there is no on second analogon given, the intended analogon essence is harder to determine.)
- *Analogon ranking multiple-choice* $?: [a_1, a_2] :: ?\{[b_1, b_2], \ldots, [z_1, z_2]\}$
  A simpler version of the general analogon ranking query are *multiple choice ranking queries* as they are for example used in the SAT benchmark dataset (discussed below). Here, the set of potential result analogons is restricted, and an algorithm would simply need to rank the provided choices instead of freely discovering the missing analogon.

2.2 Algorithmic Analogy Processing

Unfortunately, despite the potential of analogies for querying an information system, developing analogy processing algorithms received only little attention by the database and information systems community. Few early exceptions tried to accommodate analogies in first principle-style knowledge-based systems, but used them only as fallback solutions when strict inference failed, e.g., [14]. Some other systems were based on specialized case-based reasoning techniques [15], introducing a measure of similarity into reasoning. Among early systems, most approaches relied on hand-crafted data sources as for example ontology-based approaches [16], or semi-manual structure-mapping approaches [12]. The first set of techniques which showed good performance and did not require extensive manual curation relied on different natural language processing (NLP) techniques. Especially pattern mining in large Web text collections with subsequent statistical analysis showed promising result, such as [17, 18].

A recent trend from the machine learning and computational linguistics communities is learning word embeddings using Deep Learning techniques. Word embeddings represent each word in a predefined vocabulary with a real-valued vector, i.e. words are embedded in a vector space (usually with 300-600 dimensions). Most word embeddings will directly or indirectly rely on the distributional hypothesis [19] (i.e. words frequently appearing in similar linguistic contexts will also have similar real-world semantics), and are thus particularly well-suited to measure semantic similarity and relatedness between words (which is one of the foundation of the 4-term analogy definition), e.g., see [20]. Early word embeddings were often based on dimensionality reduction techniques (like for example principal component analysis) applied to word-context-co-occurrence matrixes, e.g., [21]. However, in recent years, a new breed of neural approaches relying on Deep Learning neural networks have become popular. Early neural word embeddings trained a complex multi-layer neural network to predict the next word given a sequence of initial words of a sentence [22], or to predict a nearby word given a cue word [5]. Most of these early approaches were very slow, and it could take several months to train a single model. Recent algorithmic advancements could improve on this problem, and current approaches like the popular skip-gram negative sampling approach (SGNS) [5, 23] which uses a non-linear hidden layer neural networks can train a model in just few hours using standard desktop hardware. In this paper, we will focus exclusively on neural word embeddings in the evaluation section as they are the strongest technique available today.

The straight-forward application of word embeddings is computing similarity between two given words [20] by measuring the cosine similarity. However, many (but not all) word embeddings show some very interesting and surprising additional property: it seems that not only the cosine distance between vectors represents a measure for similarity and relatedness, but that also the difference vectors between a word pair carries analogy semantics [24]. For example, the difference between the vector for "man" and "king" seems to represent the concept of being a ruler, and the vector of "woman" plus this concept vector will result in "queen". While the full extend and reasons for this behavior is not fully understood yet, these semantics have been impressively demonstrated for several examples like countries and their capitals, countries and their currencies, or several grammatical relationships (see next section). To a certain extent, these semantics can be attributed to the distributional hypothesis: in natural speech, concepts carrying similar semantics will frequently occur in similar context. Therefore, the aforementioned concept vector should implicitly encode the defining relationships between two concepts as discussed in section 2.1 (i.e.: Tokyo/Japan and Berlin/Germany will likely occur in similar contexts in natural speech, while Braunschweig/Germany will likely appear in different context and will thus have a different concept vector). This interesting property is not yet well understood, and the extends of the semantic expressiveness of neural word embeddings are still unclear.

For example, a word embedding can be used to solve *analogy completion queries* as follows [5]: Given the query $[a_1, a_2] :: [b_1, ?]$, the word embedding will provide the respective word vectors $\overrightarrow{a_1}$, $\overrightarrow{a_2}$, and $\overrightarrow{b_1}$. Then, the vector $\overrightarrow{b_2}$ representing the query's solution can be determined by finding the word vector in the trained vector space V which is closest to $\overrightarrow{a_2} - \overrightarrow{a_1} + \overrightarrow{b_1}$ (see Figure **1**) with respect to the cosine vector distance, i.e. $\overrightarrow{b_2} = \arg \max_{\vec{x} \in V, \vec{x} \neq \overrightarrow{a_2}, \vec{x} \neq \overrightarrow{b_1}} \left(\overrightarrow{a_2} - \overrightarrow{a_1} + \overrightarrow{b_1}\right)^T \vec{x}$.

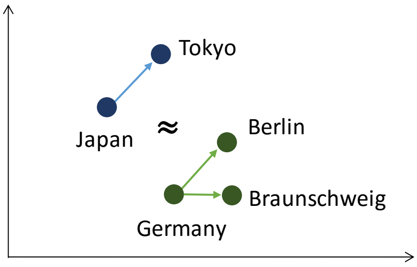

Figure 1: Example of Word Embedding Vectors reduced to 2-dimensions

$$\overrightarrow{Tokyo} - \overrightarrow{Japan} + \overrightarrow{Germany} \approx \overrightarrow{Berlin}$$

2.3 Test Collections

In the following, we highlight three commonly established benchmark collection for analogy queries, and discuss their strength and weaknesses.

SAT Analogy Challenges

The SAT analogy challenges [25] deserve some special attention due to their importance with respect to previous research and its role in real-world applications. The SAT test is standardized test for general college admissions in the United States. The test features major sections on analogy challenges to assess the prospective student's vocabulary depth and general analytical skills by focusing *on multiple-choice analogon ranking queries*. The analogy challenges contained are based strictly on relational similarity between word pairs, but do not further classify the nature of this relationship or the quality of the analogy itself. As the challenge's original intent is to assess the vocabulary skills of prospective students, it contains many rare words. In order to be able to evaluate the test without dispute, there is only a single correct answer while all other answers are definitely wrong (and can't also be argued for). As a very simple example, consider this challenge from the SAT-dataset: *legend* is to *map* as is: a) *subtitle* to *translation* b) *bar* to *graph* c) *figure* to *blueprint* d) *key* to *chart* e) *footnote* to *information.* Here, the correct answer is d) as a key helps to interpret the symbols in a chart as does the legend with the symbols of a map. While it is easy to see that this answer is correct when the solution is provided, actually solving these challenges seems to be a quite difficult task for aspiring high school students as the correctness rates of the analogy section of SAT tests is usually reported to be around 57%.

Unfortunately, this benchmark process measures the effectiveness of an algorithm only indirectly as an algorithm only needs to find the best answer – a task which is not too difficult as there is an unambiguous correct answer pair. In the design of our AGS dataset, we will relax this restricted design and introduce different degrees of result quality using a crowd-based analogy rating. This is rooted on the observation that analogies are usually not "correct" per se, but instead are more or less meaningful based on both definiteness and similarity of the involved relationships (see 2.1). Therefore, our dataset will also have a source word pair and multiple potentially analogous word pairs with an additional human judgement witch rates to the quality of the analogy (the *analogy rating* as discussed in section 3). Depending on the strictness of the intended benchmark (i.e. by adjusting the minimal analogy rating of analogon which should be considered as being correct), our dataset can therefore support challenges with a ranked list of multiple "correct" and "incorrect" pairs.

Mikolov Benchmark Dataset

For evaluating the improved continuous Skip-gram word-embedding presented in [26], Mikolov et al. created a large test set of analogy challenges covering 14 distinct relationships. The evaluation protocol is different compared to the SAT challenge set. The dataset contains 19,558 4-term analogy tuples, and each can be assigned to one of the 14 relationships contained in the dataset. The task of an analogy algorithm to be benchmarked is to predicted the missing element of a given incomplete 4-term analogy, i.e. to solve the analogy completion query $[a, b] :: [c, ?]$.

Nine of these relationships focus on grammatical properties (like for example the relationship "is plural for a noun", e.g., $[mouse, mice] :: [dollar, dollars]$ or "is superlative", e.g., $[easy, easiest] :: [lucky, luckiest]$, while five relationships are of a semantic nature (i.e. "is capital city for common country" $[Athens, Greece] :: [Oslo, Norway]$, "is capital city for uncommon country" $[Astana, Kazakhstan] :: [Harare, Zimbabwe]$, "is currency of country", "city in state", "male-female version" $[king, queen] :: [brother, sister]$. The test set is generated by collecting pairs of entities which are members of the selected relationship either manually or from Wikipedia and DBpedia, and then combining these pairs into 4-term analogy tuples. For example, for the "city in state" relationship, 68 word pairs like $[Dallas, Texas]$ or $[Miami, Florida]$ are collected, and then combined by a cross product. Interestingly, the dataset contains only 2,467 instead of the 4,556 possible tuples. It is unclear how or why this subset was sampled that way.

A core weakness of this type of test collection is that it focuses only on rather generic relationships, as e.g., "is city in" or "is plural of". The resulting 4-term statements do usually not focus on the defining relationships between the analogon terms, and therefore most of these analogy statements are semantically weak despite high relational similarity (see discussion in section 2.1). In short, this test set does not benchmark if algorithms can capture analogy semantics, but instead focuses purely on relational similarity. This is the core weakness of this dataset which we aim to rectify with our benchmark collection. Furthermore, by design, the Mikolov test set is only suitable for benchmarking analogy completion queries, and is less suitable for analogon ranking queries.

WordRep

In [27], the authors introduce the WordRep benchmark set. This dataset is based on the benchmark collection of Mikolov, and completes all missing tuples using the original word pairs. Furthermore, this test set merges some of the original categories and introduces 12 new ones, for a total of 25 categories. The word pairs for the new categories are derived automatically from both WordNet and Wikipedia. While this benchmark set is significantly larger and more complete than the Mikolov data set, it still shares the same properties, strength, and weaknesses.

3 The AGS Benchmark Collection

In the following, we will highlight the design and creation of our new AGS (Analogy Gold Standard) benchmark collection. It can be downloaded from http://www.ifis.cs.tu-bs.de/data/ags.

The semantics of real-world analogies rely on the perceived similarity and definiteness of relationships covered by the analogons from a human perspective (as discussed in section 2.1 and [13]). This core insight is ignored by all established analogy test sets presented in the last section, as they simply classify statements into "correct" and "incorrect" statements with respect to relational similarity (which, in fact, renders them

Table 1: Example challenge from AGS dataset

Source Analogon	Target Analogon	Analog Rating
sushi : Japan *(food/beverages)* *(defining)*	scallops : Italy	2.57
	currywurst : Germany	4.00
	tacos : Mexico	4.67
	curry : India	4.00
	tortilla : bat	1.00
	hamburger : pen	1.33

rather unsuitable for benchmarking analogy semantics despite their original claims, as the quality of analogies largely depends on how meaningful humans consider an analogy to be). This is the core weakness we aim to rectify with AGS by including consensual human judgements into all aspects of the collection's creation process. In order to allow for benchmarking all four query types introduced in section 2.1, we designed AGS as a collection of analogy challenges. Each challenge consists of one source analogon, and a choice of potential target analogons. Each target analogon has an *analogy rating* attached, which quantifies a consensual crowd-judgement of the analogies perceived quality (from 5: very good analogy to 1: not analogous at all). This rating is an implicit measure of both relational similarity and definiteness, as each measures is hard to elicit from humans individually. As an example, consider the challenge in Table 1. Here, the defining relationship is "is a stereotypical food of the country". Most challenges have 5-6 different target analogons. We specifically took care that some of these challenges have high analogy ratings, some have very low ratings, and some have middle-ground ratings (therefore, each challenge contains a ranked mix of good and bad analogies). Therefore, this design is particularly suitable for benchmarking multiple-choice ranking queries, but in contrast to the SAT-challenges we include also ambiguous and unclear analogies. Thus, algorithms have to decide for a proper ranking instead of simply identifying a single "correct" answer. Furthermore, AGS can also be used for analogy completion queries by only considering analogons with high analogy ratings (e.g. those with an analogy rating ≥ 4), and using the resulting 4-term statements in a similar fashion as the statements included in the Mikolov and WordRep benchmark sets (i.e., hide one concept from the 4-term statement to be guessed by the algorithm).

Each challenge of the AGS dataset is classified by a topical domain (i.e., what is the context of the intended semantics, as for example geography, or language and grammar. For a full list of included topic domains with the number of challenges and number of resulting analogies, refer to Table 2. Domains are different from the relationship types

Table 2: List of all topical domain categories used in AGS

Categories	#chall.	#analog.	Categories	#chall.	#analog.
Animals /Plants	10	128	House/Furniture/Clothing	17	169
Automobiles/Transportation	5	41	Humans/Human relations	3	28
Electrical/Electronics	4	42	Medicine/Healthcare	3	28
English grammar	11	121	Movies/Music	4	61
Food/Beverages	11	92	People/Profession	12	151
Geography/Architecture	9	144	Sports	4	33

Table 3: List of additional AGS classifiers and number of challenges

Definiteness		Knowledge		Domain	
Definite Relationship	Indefinite Relationship	Common Knowledge	Specific Knowledge	Intra-domain	Inter-domain
#37	#56	#89	#4	#93	#0

in the Wordrep collection: each of our challenges has an individual set of defining relationships). This allows us to drill-down benchmark results in case that certain algorithms show special strength and weaknesses based on the analogy's domain. We also classified each challenge with respect to the *definiteness of the source analogon*'s relationships. While we discussed that semantically meaningful analogies should always use defining relationships, many of the established benchmark datasets and algorithms focus only on relational similarity not caring if the relationships considered are defining or not. Therefore, in our dataset, we included a mix of "analogies" which use similar but not defining relationships (as, e.g., frequently used by the WordRep dataset), and semantically stronger analogies which use similar relationships which are also defining. Both classes are clearly marked in order to allow for experiments focusing on either subset. In addition, we introduced further classifications, as for example if a challenge focuses on *inter-domain* or *intra-domain analogies*. In our current version, the AGS dataset contains only intra-domain analogies, i.e. where both target and source are within the same topic domain. In future versions, we also expect inter-domain analogies which transfer the abstract essence of outwardly different analogons, as for example "The new X9000 tablet is the Ferrari of tablet computers" (i.e. $[X9000, tablets] :: [Ferrari, cars]$, which could carry the semantics that the X9000 is extremely fast and stylish, but also very expensive.) Finally, we also classified if we expect if common knowledge is sufficient to solve an analogy challenge, or if specific domain knowledge is needed (i.e. is it likely enough to build algorithms using general corpora like Wikipedia dumps or general Web crawls, or are specialized corpora needed like for example medical publications.)

Overall, AGS contains 93 challenges classified by topic, specificity of knowledge required, definiteness of the relationships, and whether it is an intra- or inter-domain analogy (see Table 3). Each challenge includes multiple analogies with high, medium, and low analogy ratings for a total of 1040 analogies overall.

Creation of the AGS Benchmark Collection

A core goal of the design of our AGS benchmark collection is to integrate human judgements and human perception deeply into the collection's creation process. Therefore, we rely on a combination of established datasets which already include semantic human judgements, and augment this seed with additional crowdsourcing. As a starting point, we use the WordSim-353 [28] and Simlex-999 [29] benchmark sets. These established datasets are created to benchmark perceived relatedness between word pairs, and each word pair has been judged for relatedness by a large number of people. From these two datasets we selected word pairs as source word pairs for our AGS challenges based on their relatedness (assuming that such pairs will be diverse and semantically meaningful). We filtered the word pairs using expert judgements from our side with respect to the pair's potential to serve as a semantically meaningful analogon, leaving 93 pairs. For expanding a single word pair into proper AGS challenges with multiple analogies, we relied on using the CrowdFlower.com crowdsourcing platform to obtain suitable target analogons and analogy ratings. Crowdsourcing is a powerful technique to outsource small tasks requiring human intelligence to a large pool of people. However, a central challenge of crowdsourcing is controlling the result quality [30]. Therefore, according to the insights in [30], we split the rather complex task of creating our challenges in several smaller tasks which are easier to control using traditional quality control mechanisms like Gold questions, averaging, and majority voting. The first of these smaller steps was to classify each source pair into one of the 12 topic domains presented in Table 2. This is followed by a second crowdsourcing task where we ask crowd workers to provide target analogons expanding a given source pair. We used the topic domains to recruit crowd workers who felt particularly confident in that domain. For extending the source analogon, we asked specifically for some examples sharing the same essence (i.e. the same defining and similar relationships), but also for some bad analogons which are either unrelated or are related but have different essence. In a final crowdsourcing task, we ask multiple workers to assess each target analogon with respect to the analogy rating, averaging the individual judgements.

For each of the tasks described above, we only used native English speakers, and for each work package we combined the input or judgements of 5 different workers. Furthermore, each worker had to perform a quick pre-assessment task (confirming or rejecting presented analogies), and only crowd worker who could solve this task correctly were allowed to participate.

4 Benchmark Protocols and Benchmark Results

In this section, we define different benchmark protocols which can be used to benchmark a given analogy algorithm with respect to the query types identified in section 2.1. Each is covered by a brief pseudocode algorithm (which focuses on a single challenge. Of course, for a full benchmark, these algorithms need to be executed for each AGS challenge). We frequently rely on selecting "correct" and "incorrect" analogies from AGS challenges. This is realized by a user defined threshold for the analogy rating allowing to adjust the strictness of the benchmark. In future, we will introduce benchmark protocols which will further differentiate result quality using numerical analogy ratings instead of working with Boolean correctness. We omitted the test protocol for open analogon ranking queries in this paper as during our experiments, none of the currently available algorithms could handle that query type in a convincingly.

Analogy Confirmation Queries $?: [\boldsymbol{a_1}, \boldsymbol{a_2}] :: [\boldsymbol{b_1}, \boldsymbol{b_2}]$

This benchmark protocol evaluates performance of a given algorithm with respect to analogy confirmation queries, checking if it can distinguish "correct" analogies from "incorrect" ones. The correctness of AGS statements is based on the analogy rating of target analogons, i.e. those with an analogy rating exceeding a minimal threshold are considered correct, and those with a rating lower than a given threshold are considered incorrect (using Table 2 and minimal correct threshold of 4, and max incorrect threshold of 2, $[sushi, Japan] :: [tacos, Mexico]$ is a correct statement, $[sushi, Japan] :: [scallops: Italy]$ is excluded from the benchmark as it is neither clearly correct nor incorrect, and $[sushi, Japan] :: [hamburger, pen]$ is an incorrect statement.) The final result of this benchmark covering multiple challenges is the percentage of correctly confirmed statements contained in a given subset of AGS challenges.

`Analogy Confirmation Benchmark(Challenge c)`
Feature Required from Algorithm:
- Algorithm needs to be able to confirm or reject a given analogy statement

Parameters:
- $min\_correct$: Minimal analogy rating to consider a target analogon as "correct"
- $max\_incorrect$: Maximal analogy rating to consider a target analogon as "incorrect"

Output:
- $num\_success$, num_fail: Number of correctly and incorrectly processed statements

Protocol:
- correct statements = Combine source $c.source$ with all targets $t \in c.target$ which have $t \geq min\_correct$
- incorrect statements = Combine $c.source$ with all targets $t \in c.target$ which have $t < max\_incorrect$
- Check if algorithm confirms correct statements as analogies
- Check if algorithm rejects incorrect statements as analogies
- Return respective success and failure numbers

Analogy Completion Queries $?: [\boldsymbol{a_1}, \boldsymbol{a_2}] :: [\boldsymbol{b_1}, ?]$

In this benchmark protocol, we check if the given algorithm can complete "correct" analogy statements. The final result of this benchmark is the percentage of correctly completed statements contained in a given subset of AGS challenges.

`Analogy Completion Benchmark(Challenge c)`
Feature Required from Algorithm:
- Algorithm needs to be able to complete an analogy statement $[a_1, a_2] :: [b_1, ?]$

Parameters:
- $min\_correct$: Minimal analogy rating to consider a target analogon as "correct"

Output:
- $num\_success$, num_fail: Number of correctly and incorrectly processed statements

Protocol:
- correct statements = Combine source $c.source$ with all targets $t \in c.target$ which have $t \geq min\_correct$
- From each correct statement, drop the second concept of the target analogon
- Check if the algorithm can correctly predict the dropped concept
- Return respective success and failure numbers

Analogon Ranking Multiple-Choice Queries $?: [\boldsymbol{a_1}, \boldsymbol{a_2}] :: ?\{[\boldsymbol{b_1}, \boldsymbol{b_2}], \dots, [\boldsymbol{z_1}, \boldsymbol{z_2}]\}$

In this simple ranking benchmark, we evaluate if a given analogy algorithm can select the best target analogon for a given source analogon from a limited list of candidates. In our future works, we will extend this protocol to consider rank correlation instead of focusing only on the best analogon. The final result is the percentage of correctly answered challenges.

`Analogy Ranking Multiple Choice Benchmark(Challenge c)`

Feature Required from Algorithm:

- Algorithm needs to be able to measure and quantify the quality of analogy between the analogon $[a_1, a_2]$ and $[b_1, b_2]$ (e.g., the relational similarity of the defining relationships of each analogon)

Output:

- $success$: Boolean result indicating if challenge was solved correctly

Protocol:

- statements = Combine source $c.source$ with all targets $t \in c.target$
- Measure quality of analogy for each statement
- If the statement with highest measured quality is also the statement with the highest analogy rating in AGS, return success. If not, return failure.

Benchmark Results

As a proof of concept, we present example benchmark results for two implementations of neural word embeddings in this section. We use our AGS collection with the aforementioned benchmark protocols. The algorithms under consideration are the well-known word2vec implementation by Mikolov et al. [26], and the Glove implementation by Pennington et al [23]. Both algorithms were trained on a dump of Wikipedia, using only the implementation's predefined default parameters. Besides benchmarks covering the full extent of AGS, we also focus on different classification aspects (like only focusing either definite challenges or indefinite ones, see classification in section 3).

We used the following parameters for our benchmark protocols: minimal analogy rating threshold for correct analogies of 4.0, and maximal threshold for incorrect analogies of 2.0. The results for completion and multiple-choice ranking queries are summarized in Table 4. Confirmation queries and open-rank queries are not directly supported by word-embedding based algorithms, and we therefore excluded them from this evaluation.

In general, the measured results of both word2vec and Glove on our AGS collection are rather weak. However, evaluations of the same algorithms on other benchmark sets like Wordrep showed slightly better results (see [27], around 0.25 overall accuracy). This can be explained by that fact that Wordrep uses some very limited set of relationships types which are, in comparison, rather simple in their semantic nature (e.g., plurals like $[fish{:}fishes] :: [pig, pigs]$, or simple "city is located in" relationships). Those datasets were mostly created by mining Wikipedia, DBpedia, or Wordnet for generating a large number of example analogons using the same relationships – and the same Wikipedia texts are used to train the neural word embeddings both Glove and word2vec use. In our AGS dataset, we use analogies as provided by real humans which are inherently more complex from a semantic point of view. Therefore, our results indicate that, while algorithm relying on word embeddings might be able to deal with simple relational similarity quite well, mastering semantically rich analogy processing still requires a significant amount of future research.

Table 4: Benchmark Results

Protocol	*Algorithm*	*Overall*	*Common Knowledge*	*Uncommon Knowledge*	*Definite*	*Indefinite*
Completion	Word2Vec	0.1786	0.1781	0.1851	0.225	0.1481
	Glove	0.1960	0.1941	0.2222	0.2562	0.1563
Ranking multiple choice	Word2Vec	0.3026	0.3030	0.30	0.3125	0.2954
	Glove	0.3289	0.3181	0.30	0.3437	0.3181

5 Summary and Outlook

In this paper, we presented and discussed the challenge of benchmarking analogy processing algorithms. Such algorithms will be an important building block of future human-centered information systems trying to understand the finer semantics of natural language. Unfortunately, despite the potential importance of analogies in future query paradigms, there is still no clear definition of the intended semantics of analogy queries. Therefore, we provided several discussions focusing on that topic, and derived a set of core properties of analogy semantics. Furthermore, we highlighted basic query types and discussed how they could be benchmarked. Based on these results, we created the AGS analogy benchmark dataset, which aims to rectify several shortcomings of existing benchmark datasets. Especially, our dataset is not automatically generated from structured data sources, but instead relies on crowdsourcing and a large number of human judgements. Furthermore, we explicitly focus on semantically rich analogies instead of limiting ourselves to the significantly weaker special case of only relationally similar word pairs. Finally, we designed the dataset in such a way that all our identified query types could be benchmarked with a single dataset.

From an algorithmic point of view, the recent years have brought several impressive advancements in language understanding, and especially a new breed of deep-learning neural embedding techniques showed very impressive results for challenges related to analogy processing like measuring semantic similarity or relational similarity. Unfortunately, these algorithms do not show strong results on our new benchmark dataset, thus further emphasizing the need for continued future research in this field.

References

1. Lofi, C.: Analogy Queries in Information Systems – A New Challenge. J. Inf. Knowl. Manag. 12, (2013).
2. Hofstadter, D.R.: Analogy as the Core of Cognition. In: The Analogical Mind. pp. 499–538 (2001).
3. Gentner, D.: Why We’re So Smart. In: Language in Mind: Advances in the Study of Language and Thought. pp. 195–235. MIT Press (2003).
4. Veale, T.: Wordnet sits the sat: a knowledge-based approach to lexical analogy. In: Europ. Conf. on Artificial Intelligence (ECAI). , Valencia, Spain (2004).
5. Mikolov, T., Yih, W., Zweig, G.: Linguistic Regularities in Continuous Space Word Representations. In: Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language (NAACL-HLT). , Atlanta, USA (2013).
6. Dedre Gentner, Keith J. Holyoak, Boicho N. Kokinov eds: The analogical mind: perspectives from cognitive science. MIT Press (2001).
7. Itkonen, E.: Analogy as structure and process: Approaches in linguistics, cognitive psychology and philosophy of science. John Benjamins Pub Co (2005).
8. Shelley, C.: Multiple Analogies In Science And Philosophy. John Benjamins Pub. (2003).
9. Kant, I.: Critique of Judgement. (1790).
10. Juthe, A.: Argument by Analogy. Argumentation. 19, 1–27 (2005).
11. Gentner, D.: Structure-mapping: A theoretical framework for analogy. Cogn. Sci. 7, 155–170 (1983).
12. Gentner, D., Gunn, V.: Structural alignment facilitates the noticing of differences. Mem. Cognit. 29, 565–77 (2001).
13. Lofi, C., Nieke, C.: Modeling Analogies for Human-Centered Information Systems. In: 5th

Int. Conf. On Social Informatics (SocInfo). , Kyoto, Japan (2013).

14. Blythe, J., Veloso, M.: Analogical replay for efficient conditional planning. In: Nat. Conf. on Artificial Intelligence (AAAI). , Providence, Rhode Island, USA (1997).
15. Leake, D.: Case-Based Reasoning: Experiences, Lessons, and Future Directions. MIT Press (1996).
16. Forbus, K.D., Mostek, T., Ferguson, R.: Analogy Ontology for Integrating Analogical Processing and First-principles Reasoning. In: Nat. Conf. on Artificial Intelligence (AAAI). , Edmonton, Alberta, Canada (2002).
17. Bollegala, D.T., Matsuo, Y., Ishizuka, M.: Measuring the similarity between implicit semantic relations from the web. In: Int. Conf. on World Wide Web (WWW). , Madrid, Spain (2009).
18. Davidov, D.: Unsupervised Discovery of Generic Relationships Using Pattern Clusters and its Evaluation by Automatically Generated SAT Analogy Questions. In: Ass. for Computational Linguistics: Human Language Technologies (ACL:HLT). , Columbus, Ohio, USA (2008).
19. Harris, Z.: Distributional Structure. Word. 10, 146–162 (1954).
20. Lofi, C.: Measuring Semantic Similarity and Relatedness with Distributional and Knowledge-based Approaches. Database Soc. Japan J. 14, 1–9 (2016).
21. Ștefănescu, D., Banjade, R., Rus, V.: Latent Semantic Analysis Models on Wikipedia and TASA. In: Language Resources Evaluation Conference (LREC). , Reykjavik, Island (2014).
22. Mnih, A., Hinton, G.E.: A scalable hierarchical distributed language model. Adv. Neural Inf. Process. Syst. 1081–1088 (2009).
23. Pennington, J., Socher, R., Manning, C.D.: Glove: Global vectors for word representation. In: Conf. on Empirical Methods on Natural Language Processing (EMNLP). , Doha, Qatar (2014).
24. Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., Kuksa, P.: Natural language processing (almost) from scratch. J. Mach. Learn. Res. 2493–2537 (2011).
25. Littman, M., Turney, P.: SAT Aanalogy Challange Dataset, http://aclweb.org/aclwiki/index.php?title=SAT_Analogy_Questions_(State_of_the_art).
26. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed Representations of Words and Phrases and their Compositionality. Adv. Neural Inf. Process. Syst. 3111–3119 (2013).
27. Gao, B., Bian, J., Liu, T.-Y.: WordRep: A Benchmark for Research on Learning Word Representations. In: ICML Workshop on Knowledge-Powered Deep Learning for Text Mining. , Beijing, China (2014).
28. Finkelstein, L., Gabrilovich, E., Matias, Y., Rivlin, E., Solan, Z., Wolfman, G., Ruppin, E.: Placing search in context: the concept revisited. In: Int. Conf. on World Wide Web (WWW). , Hong Kong, China (2001).
29. Hill, F., Reichart, R., Korhonen, A.: SimLex-999: Evaluating Semantic Models with (Genuine) Similarity Estimation. Prepr. Publ. arXiv. arXiv14083456. 2014,.
30. Lofi, C., Selke, J., Balke, W.-T.: Information Extraction Meets Crowdsourcing: A Promising Couple. Datenbank-Spektrum. 12, (2012).