Towards Analogy-based Recommendation

Benchmarking of Perceived Analogy Semantics

Christoph Lofi
Web Information Systems - TU Delft
Mekelweg 4
Delft, Netherlands 2628CD
c.lofi@tudelft.nl

Nava Tintarev
Web Information Systems - TU Delft
Mekelweg 4
Delft, Netherlands 2628CD
n.tintarev@tudelft.nl

ABSTRACT

Requests for recommendation can be seen as a form of query for candidate items, ranked by relevance. Users are however often unable to crisply define what they are looking for. One of the core concepts of natural communication for describing and explaining complex information needs in an intuitive fashion are analogies: e.g., *"What is to Christopher Nolan as is 2001: A Space Odyssey to Stanley Kubrick?"*. Analogies allow users to explore the item space by formulating queries in terms of items rather than explicitly specifying the properties that they find attractive. One of the core challenges which hamper research on analogy-enabled queries is that analogy semantics rely on consensus on human perception, which is not well represented in current benchmark data sets. Therefore, in this paper we introduce a new benchmark dataset focusing on the human aspects for analogy semantics. Furthermore, we evaluate a popular technique for analogy semantics (word2vec neuronal embeddings) using our dataset. The results show that current word embedding approaches are still not not suitable to sufficiently deal with deeper analogy semantics. We discuss future directions including hybrid algorithms also incorporating structural or crowd-based approaches, and the potential for analogy-based explanations.

KEYWORDS

Analogy-Enabled Recommendation, Relational Similarity, Analogy Benchmarking

1 INTRODUCTION

In this paper we explore one method of efficiently communicating information needs or for explaining results of an inquiry used in natural human interaction, namely *analogies*. Using analogies in natural speech allows communicating dense information easily and naturally by implying that the "essence" of two concepts is similar or at least perceived similarly. Thus, analogies can be used to map factual and behavioural properties from one (usually better known concept, the source) to another (usually less well known, the target) concept. This is particularly effective for natural querying and explaining when only vague domain knowledge is available.

In this paper we consider two types of analogy queries, 4-term analogy completion queries (like "What is to Christopher Nolan as is 2001: A Space Odyssey to Stanley Kubrick?"), and 4-term analogon ranking (like "What is similar to '2001: A Space Odyssey to Stanley Kubrick'? a) The Fifth Element to Christopher Nolan, b) Memento to Christopher Nolan, c) Dunkirk to Christopher Nolan). Analogy completion queries might be seen an extension on classical critiquing in recommender systems which can be formulated in terms of "like x, but with properties y modified". In critiquing the feature (price) and the modification (cheaper) needs to be explicit, whereas in an analogy, the semantics are implicitly given by setting terms in relation, which is interpreted based on both communication partners' conceptualization of that domain (e.g., "The Fifth Element is like 2001: A Space Odyssey but by Scorsese" carries a lot of implicit information).

Analogies typically are represented as rhetorical figures of speech which need to be interpreted and reasoned on using the receiver's background knowledge - which is difficult for information systems to mimic. This complex semantic task is further complicated by the lack of useful benchmark datasets. Most current benchmarks are restricted in scope, usually focusing on syntactic features instead of relevant properties of analogies as for example their suitability for transferring information (which is central when one wants to use analogies for querying, or explaining recommendations).

Therefore, one of our core contributions is an improved benchmark dataset for analogy semantics which focuses specifically on the usefulness of an analogy with respect to querying and explaining, and providing it as a tool for guiding future research into analogy queries in recommender systems. In this work, we are focusing on general domain analogies. This allows us the choose the right technologies for future adaption in a domain-specific recommender system. In detail, our contributions are as follows:

- Discuss different properties of analogy semantics, and highlight their importance for querying and explaining recommendations.
- Introduce a new benchmark dataset systematically built on top of existing sets, rectifying many current limitations. Especially, we focus on perceived difficulty of analogies, and the quality and usefulness of analogies.
- Showcase and discuss the performance of an established word embedding-based algorithm on our test set.

2 DISCUSSIONS ON ANALOGY SEMANTICS

The semantics of analogies have been researched in depth in the fields of philosophy, linguistics, and in the cognitive sciences, such as [11], [12], or [21]. There have been several models for analogical reasoning from philosophy (like works by Plato, Aristotle, or Kant [13]), while other approaches see analogies as a variant of induction in formal logic [21], or mapping the structures of relationships and properties of entity pairs (structure mapping theory [8]). However,

ComplexRec 2017, Como, Italy.

those analogy definitions are rather complex and hard to grasp computationally, and thus most recent works on computational analogy processing rely on the simple 4-term analogy model which 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 relationship between a_1 and a_2 is similar to the relation between b_1 and b_2, as for example in $[StarWars, Scifi] :: [ForrestGump, Comedy]$ (both are defining movies within their genres). This model has several limitations, as is discussed by Lofi in [15]: the semantics of "a high degree of relational similarity" from an ontological point of view is unclear, and the model ignores human perception and abstractions (e.g., analogies are usually not right or wrong, but better or worse based on how well humans understand them).

Therefore, in this paper we promote an improved interpretation of the 4-term analogy model [15], and assume that there can be multiple relationships between the concepts of an analogon, some of them being relevant for the semantics of an analogy (the *defining* relationships), and some of them not. An analogy holds true if the sets of defining relationships of both analogons show a high degree of relational similarity. For illustrating the difference and importance of this change in semantics, consider the analogy statement: $[StanleyKubrick, 2001 : ASpaceOdyssey] :: [TaxiDriver, MartinScorsese]$. Kubrick is the director of *2001: A Space Odyssey*, and Scorsese is the director of *Taxi Driver.* Both analogons contain the same "movie is directed by person" relationship, and this could be considered a valid analogy with respect to the simple 4-term analogy model. Still, this is a poor analogy statement from a human communication point of view because *2001: A Space Odyssey* is not like *Taxi Driver* at all. Therefore, this statement does neither describe the essence of 2001: A Space Odyssey nor the essence of Taxi Driver particularly well: both movies are iconic for their respective directors, but they do not for example describe that one movie is in the science fiction genre, and the other could be classified as a (violent) crime movie. 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 degree to which relationships are defining an analogon may vary with different contexts. 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 degree of how well the relationships shared by both analogons are defining them (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. Based on these observations, we define two basic types of analogy queries (loosely adopted from [16]) which can be used for analogy-enabled recommender systems:

(1) *Analogy completion* $? : [a_1, a_2] :: [b_1, ?]$
This query can be used to find the missing concept in a 4-term analogy. This is therefore the most useful query type in a future analogy-enabled information systems [15]. 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.

(2) *Analogon ranking multiple-choice*
$? : [a_1, a_2] ::?\{[b_1, b_2], [c_1, c_2] \ldots\}$
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 (e.g.,$[b_1, b_2]$; $[c_1, c_2]$) instead of freely discovering the missing analogon.

Previous approaches to processing analogies algorithmically cover prototype systems operating on Linked Open Data (LOD), as for example [5], but also approaches which mine analogies from natural text [18]. A very popular recent trend in Natural Language Processing (NLP) is training neuronal word embeddings [20]. The popular word2vec implementation [1] learns relational similarity between word pairs from large natural language corpora by exploiting the distributional hypothesis [9] using neuronal networks. Word-embeddings are particularly interesting for use in analogy-enabled recommender systems as they can be easily trained on big text corpora (like for example user reviews), and do not require structured ontologies and taxonomies like other approaches (e.g., [22]). In this paper, we evaluate the analogy reasoning performance of word embeddings using our new benchmark dataset which represents analogy semantics more naturally.

3 BENCHMARK DATASETS

For benchmarking the effectiveness of analogy algorithms, there are several established Gold standard datasets for general-knowledge analogies (i.e. not tailored to a specific domain). However, all of them are lacking in some respect, as discussed in the following sections.

3.1 Mikolov Benchmark Dataset & Wordrep

The Mikolov Benchmark set [19] is one of the most popular benchmark sets for testing the analogy reasoning capabilities of neuronal word embeddings covering 19,558 4-term analogies with 14 distinct relationship types, focusing exclusively on analogy completion queries $[a, b] :: [c, ?]$. Nine of these relationships focus on grammatical properties (e.g., the relationship "is plural for a noun"), while five relationships are of a semantic nature (e.g., "is capital city of a country" like $[Athens, Greece] :: [Oslo, Norway]$). The benchmark set is generated by collecting pairs of entities which are members of the selected relationship type from Wikipedia and DBpedia, and then combining all these pairs into 4-term analogy tuples. The Wordrep dataset [7] extends the Mikolov set by adding more challenges, and expanding to 25 different relationship types.

A core weakness of this type of test set is that it does not include any human judgment with respect to the defining relationship types and the usefulness as an analogy for querying or explaining, but instead focuses only on "correct" relationships which do usually not carry any of the subtleties of rhetorical analogies, as e.g., "is city in" or "is plural of". In short, the Mikolov test set does not benchmark if algorithms can capture analogy semantics from a human perspective, but instead focuses purely on relational similarity for a very limited selection of relationship types. Thus, we feel that this benchmark dataset does not represent the challenge of real world analogy queries well.

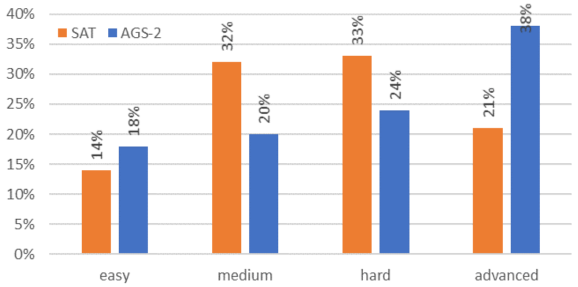

Figure 1: Difficulty of challenges for SAT and AGS-2 dataset

3.2 SAT Analogy Challenges

The SAT analogy challenge [14] plays an important role in real-world college admission tests in the United States to assess a prospective student's vocabulary depth and general analytical skills by focusing on analogon ranking queries. As the challenge's original intent is to assess the vocabulary and reasoning 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 cannot also be argued for). The SAT dataset contains 374 challenges like this one: *"legend is to map as is: a) subtitle to translation, b) bar to graph, c) figure to blue-print, 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, *solving these challenges is a 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%* [23]. An interesting aspect of the SAT analogy test set is that a large variety of different relationship types are covered, and that the contained challenges have a high degree of variance of difficulty from a humans' perspective.

Some analogy challenges are harder than others for the average person, usually based on the rareness of the used vocabulary and the complexity of the reasoning process required in order to grasp the intended semantics. *Understanding the difficulty of challenges is thus important when judging the effectiveness of an algorithmic solution, as this provides us with a sense for ”human-level performance”.*

However, while the SAT dataset can be obtained easily for research purposes, there is no publicly available assessment of the difficultly of different challenges. Lofi et al examined the performance of crowd workers recruited from Amazon Mechanical Turk when faced with SAT challenges [16]. They recruited 99 workers from English-speaking countries, and each challenge was solved by 8 crowd workers each (in average, each worker solved 29 challenges.) It turned out that workers are rarely in agreement, and that most challenges (> 67%) received 3 or 4 different answers from just 8 crowd workers. Only 3.6% challenges are easy enough to garner unequivocal agreement from all 8 crowd workers, while most challenges only had an agreement of 4 to 5 workers.

In this paper, we use those results to classify each SAT challenge by their crowd difficulty, and make this data publicly available. This allows us to discuss the performance of analogy algorithms in comparison to human performance (see section 4). To this end, we classified each challenge into one of four difficulty groups, easy for challenges which could be handled by most crowd workers (7-8 correct votes), medium (5-6 correct votes), difficult (3-4 correct votes), and advanced (0-2 correct votes). Our SAT difficulty ratings can be downloaded on our companion page [3], and the resulting difficulty distribution is shown in Figure 1.

Table 1: Example AGS challenge

Source Analogon	Target Analogon	Rating
sushi : Japan	scallops : Italy	2.57
	currywurst : Germany	4.00
	tacos : Mexico	4.67
	curry : India	4.00
	sombrero : Mexico	2.00
	hamburger : ship	1.33

3.3 Analogy Gold Standard AGS-1

Lofi et al. introduced a benchmark dataset which aims at rectifying the shortcomings of aforementioned sets, the Analogy Gold Standard AGS dataset [17]. In this dataset, each challenge can be used to benchmark both completion queries, but also ranking queries. AGS was systematically build using different seed datasets like the SAT dataset or the WordSim-353 dataset [6]. One of the core contribution of AGS was to include human judgments of how "good" an analogy is from a subjective point of view (as compared to previous test sets where analogy statements are either correct or wrong). The quality of an analogy is influenced by the relational similarity of the analogons, combined with how well the shared relationships represent the essence of the analogy (see section 2 for a discussion).

As previous experiments showed that human workers had problems to individually quantify those aspects, for AGS, a new *analogy rating* was introduced which implicitly encodes both relational similarity and representativeness, i.e. the analogy rating represents how "useful" and "good" the analogy is perceived by humans on a scale of 1 to 5. The judgments of at least 8 humans recruited via crowd-sourcing platforms was used for the AGS analogy ratings of each analogon pair in a challenge.

An example AGS challenge is given in Table 1: in this example, while sombreros are typically considered to come from Mexico, and in the source analogon sushi typically comes from Japan (i.e. there is similar relationship between both analogons), the resulting analogy $[sushi, Japan] :: [sombrero, Mexico]$ still has a low analogy rating because the the defining relationship in the source analogon $[sushi, Japan]$ is usually understood by people as "stereotypical food from a country" - and thus the analogy is deemed not useful by most humans.

3.4 Improved Analogy Gold Standard AGS-2

In this paper we introduce the AGS-2 dataset which significantly improves on AGS-1 with respect to several aspects. The AGS-2 dataset can be obtained at our companion page [3], thus providing tangible benefits to ongoing analogy-enabled information system research. Notably, the core improvements of AGS-2 are as follows:

- Added *difficulty rating* for each challenge based on the crowd feedback of 5 workers. The scale is similar to our difficulty rating for SAT challenges introduced in section 3.2: advanced, hard, medium, easy. The resulting difficulty distribution of AGS-2 is also shown in Figure 1.

- Extended *size and scope*: The initial seed analogons for creating AGS-1 were extracted from the the Simlex [10] and Wordsim datasets [6], resulting in 93 challenges overall. For AGS-2, this was extended using suggestions of potentially interesting analogy pairs by a subset of trusted crowd workers. We manually selected a subset of these suggestions, resulting in 168 challenges overall.
- *Improved balance* of analogy ratings: In AGS-1, challenges could be imbalanced with respect to the analogy ratings (i.e., a given challenge could contain many analogons with high analogy ratings, but only few with low ratings). For AGS-2 we ensure that each challenge covers 1-2 analogons each for high, medium, and low analogy ratings. Note that an analogy with high difficulty and high analogy rating might still not be understandable by many people (due to being difficult by using rarely known concepts or complex reasoning), but still will be accepted as a good analogy by the same people after the semantics are explained to them.

4 EVALUATION

In this section, we give a preview on how word-embeddings perform on analogy challenges of varying difficulty as previous work only relied on the aforementioned Mikolov dataset where they showed a comparably good accuracy of 53.3%. The embedding we evaluated is Gensim's skip-gram word2vec [2], trained on a Wikipedia dump. We use the SAT and AGS-2 datasets, and split the results by the new difficulty measurements introduced in section 3.2.

As the SAT dataset only supports analogy ranking queries, for brevity we only report ranking query results in the following. We followed the test protocol outlined in [17]: we rank each analogon of a challenge by their relational similarity score with respect to the source as computed by word2vec, and consider a challenge successfully solved if the top-ranked analogon is the correct one (SAT) or has an analogy score higher than a chosen threshold (4.0 in our case for AGS-2).

The results are summarized in Figure 2, and are rather disappointing: for SAT, the average accuracy is 23.4% (random guessing achieves 20% as each challenge has 5 answer options, average human-level performance is 57% [23], while one of the current best performing analogy systems based on both distributional semantics and structural reasoning using ontologies (like DBpedia or WordNet) performs at 56.1% [22]). For AGS-2 the word2vec accuracy is 25% (random guessing achieves 16.7%). Interestingly, the performance of word2vec is consistent with respect to difficulty levels, and it performs worse than humans for easy challenges, but comparably well for advanced challenges. From this preliminary result we conclude that the analogy reasoning capabilities of neuronal embeddings, despite some of their advantages like ease of use and easy training just relying on text collections, are inferior than current anecdotal and empirical evidence suggests. However, specialized analogy reasoning algorithms have been shown to achieve up to 56.1% on SAT [4], and this has mostly been realized by also incorporating ontological knowledge (as suggested by [22]). Unfortunately, this could be challenging for some domain-specific recommender systems where such ontologies are not easily available, thus promoting future research to overcome this issue.

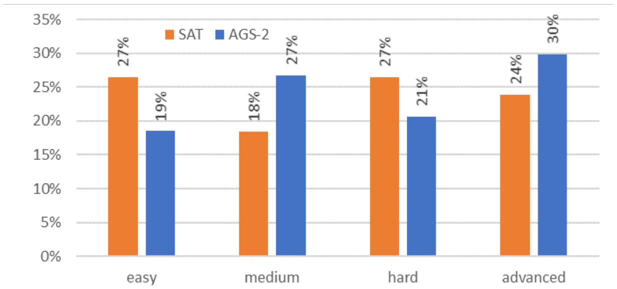

Figure 2: Accuracy of w2v solving SAT and AGS-2

5 SUMMARY AND OUTLOOK

In this paper, we introduced the challenge of analogy semantics for recommender systems. Analogies are a natural communication paradigm to efficiently state an information need or to explain a complex concept. This relies on exploiting the background knowledge and reasoning capabilities of both communication partners, a task challenging even for human judges. We introduced the AGS-2 benchmark data set overcoming many shortcomings of previous datasets, such as being usable for both analogy ranking and analogy completion queries, and is based on human perception for rating the quality and difficulty of a benchmark analogy. We evaluated the performance of neuronal word embeddings (using word2vec as a representative) on previous datasets and our new benchmark (AGS-2). While the method worked worse than human-level performance for simple queries, it was comparable for more complex queries with rare vocabulary and requiring extensive common knowledge.

In our next steps we plan to investigate the performance of hybrid methods, using both embeddings as well as structural reasoning to enable analogy queries for recommender systems. This particularly involves adopting analogy semantics to specific domains like music, books, or movies - while current analogy systems and benchmark datasets (including AGS-2) focus on common-knowledge analogies (which is slightly easier due to the ready availability of both large text corpora and ontologies).

In this context, we also plan to evaluate the performance of analogy based explanations for supporting people in making decisions about recommended items. This is a particularly interesting challenge as approaches based only on distributed semantics do implicitly encode analogy semantics, but have now explicit knowledge on the type of relationships which would be required to explain the semantics to a human user, thus again underlining the need for explicit information on the relationships used in an analogy.

REFERENCES

[1] 2013. Word2Vec. https://code.google.com/archive/p/word2vec/. (2013). Accessed: 2017-06-01.

[2] 2014. Gensim Word2Vec. https://radimrehurek.com/gensim/models/word2vec.html. (2014). Accessed: 2017-06-01.

[3] 2017. Analogy Semantics Companion Page. https://github.com/WISDelft/analogy_semantics. (2017). Accessed: 2017-08-01.

[4] 2017. SAT Analogy Questions: State of the Art. https://www.aclweb.org/aclwiki/index.php?title=SAT_Analogy_Questions_(State_of_the_art). (2017). Accessed: 2017-06-01.

[5] Danushka Bollegala, Tomokazu Goto, Nguyen Tuan Duc, and Mitsuru Ishizuka. 2012. Improving Relational Similarity Measurement using Symmetries in Proportional Word Analogies. *Information Processing & Management* 49, 1 (2012), 355–369.

[6] Lev Finkelstein, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin. 2001. Placing search in context: the concept revisited. In *Int. Conf. on World Wide Web (WWW)*. Hong Kong, China.

[7] Bin Gao, Jiang Bian, and Tie-Yan Liu. 2014. WordRep: A Benchmark for Research on Learning Word Representations. In *ICML Workshop on Knowledge-Powered Deep Learning for Text Mining*. Beijing, China.

[8] D Gentner. 1983. Structure-mapping: A theoretical framework for analogy. *Cognitive science* 7 (1983), 155–170.

[9] Z. Harris. 1954. Distributional Structure. *Word* 10 (1954), 146–162.

[10] F. Hill, R. Reichart, and A. Korhonen. 2014. SimLex-999: Evaluating Semantic Models with (Genuine) Similarity Estimation. *Preprint published on arXiv. arXiv:1408:3456* (2014).

[11] Douglas R. Hofstadter. 2001. Analogy as the Core of Cognition. In *The Analogical Mind*. 499–538.

[12] Esa Itkonen. 2005. *Analogy as structure and process: Approaches in linguistics, cognitive psychology and philosophy of science*. John Benjamins Pub Co.

[13] Immanuel Kant. 1790. *Critique of Judgement*.

[14] Michael Littman and Peter Turney. 2016. SAT Aanalogy Challange Dataset. (2016). https://www.aclweb.org/aclwiki/index.php?title=Analogy_(State_of_the_art)

[15] Christoph Lofi. 2013. Analogy Queries in Information Systems: A New Challenge. *Journal of Information & Knowledge Management (JIKM)* 12, 3 (2013).

[16] Christoph Lofi. 2013. Just ask a human? – Controlling Quality in Relational Similarity and Analogy Processing using the Crowd. In *CDIM Workshop at Database Systems for Business Technology and Web (BTW)*. Magdeburg, Germany.

[17] Christoph Lofi, Athiq Ahamed, Pratima Kulkarni, and Ravi Thakkar. 2016. Benchmarking Semantic Capabilities of Analogy Querying Algorithms. In *Int. Conf. on Database Systems for Advanced Applications (DASFAA)*. Dallas, TX, USA.

[18] C. Lofi, C. Nieke, and N. Collier. 2014. Discriminating Rhetorical Analogies in Social Media. In *Conf. of the Europ. Chapter of the Association for Computational Linguistics (EACL)*. Gothenburg, Sweden.

[19] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *International Conference on Learning Representations (ICLR)* (2013).

[20] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean. 2013. Distributed Representations of Words and Phrases and their Compositionality. *Advances in Neural Information Processing Systems* 21 (2013), 3111–3119.

[21] Cameron Shelley. 2003. *Multiple Analogies In Science And Philosophy*. John Benjamins Pub.

[22] Robert Speer, Joshua Chin, and Catherine Havasi. 2016. ConceptNet 5.5: An Open Multilingual Graph of General Knowledge. *AAAI Conference on Artificial Intelligence* (2016).

[23] P. Turney and M. Littman. 2005. Corpus-based learning of analogies and semantic relations. *Machine Learning* 60 (2005), 251–278.