# Analogies between sentences: Theoretical aspects - preliminary experiments

Stergos Afantenos[1], Tarek Kunze[1], Suryani Lim[2], Henri Prade[1], and Gilles Richard[1]

[1] IRIT, Toulouse, France
[2] Federation University, Churchill, Australia
{gilles.richard, henri.prade, stergos.afantenos}@irit.fr
tarek.kunze@protonmail.com
suryani.lim@federation.edu.au

**Abstract.** Analogical proportions hold between 4 items $a$, $b$, $c$, $d$ insofar as we can consider that "$a$ is to $b$ as $c$ is to $d$". Such proportions are supposed to obey postulates, from which one can derive Boolean or numerical models that relate vector-based representations of items making a proportion. One basic postulate is the preservation of the proportion by permuting the central elements $b$ and $c$. However this postulate becomes debatable in many cases when items are words or sentences. This paper proposes a weaker set of postulates based on *internal reversal*, from which new Boolean and numerical models are derived. The new system of postulates is used to extend a finite set of examples in a machine learning perspective. By embedding a whole sentence into a real-valued vector space, we tested the potential of these weaker postulates for classifying analogical sentences into valid and non-valid proportions. It is advocated that identifying analogical proportions between sentences may be of interest especially for checking discourse coherence, question-answering, argumentation and computational creativity. The proposed theoretical setting backed with promising preliminary experimental results also suggests the possibility of crossing a real-valued embedding with an ontology-based representation of words. This hybrid approach might provide some insights to automatically extract analogical proportions in natural language corpora.

## 1 Introduction

Analogies play an important role in human reasoning, and are thus involved in natural language discourses. The study of analogical reasoning has a long history (see, e.g., Chap. 1 in [32]). It departs from case-based reasoning [29]. Beyond different classical works on analogical reasoning such as [15, 35, 14, 13, 16]), there has been a noticeable renewal of interest in analogical studies with a variety of approaches, ranging from reasoning [2], machine learning [23, 5] to word analogies [6, 11, 36, 37, 20, 27] and natural language processing [19, 12, 34]. These approaches have in common to deal with analogical proportions, i.e., statements of the form "$a$ is to $b$ as $c$ is to $d$" relating 4 items $a$, $b$, $c$ and $d$ [30].

Recently, some authors [10, 39] have started to study analogical proportions between sentences, motivated by question-answering concerns or the evaluation of sentence embeddings. This paper pursues this study, but first questions the modeling of analogical proportions that is used. Indeed, it is generally assumed that, as for the numerical proportions, the permutation of the central elements $b$ and $c$ preserves the validity of analogical proportions. This postulate is quite debatable for natural language items. For this reason, the paper proposes a postulate weaker than the stability under central permutation. Its main contributions are:

- i) A formal setting to deal with the notion of *analogical sentences*, where only an *internal reversal* property is assumed for the pairs $(a, b)$ and $(c, d)$.

- ii) A rigorous way to extend a finite set of analogical sentences with both valid and non-valid examples. In a machine learning perspective, this process is also known as data augmentation.

-iii) A set of preliminary experiments showing that this notion of analogical sentences is valid and could bring interesting perspectives in terms of applications.

The paper is structured as follows. Section 2 investigates the related works. Section 3 recalls the postulates characterizing analogical proportions and proposes a weaker set of postulates avoiding central permutation. This allows the identification of a rigorous method for enlarging a set of examples and counter-examples. Section 4 presents and discusses the experimental settings as well as the way we generate datasets of analogical sentences. Section 5 reports the results for experiments based on two datasets, showing that machine learning-based techniques are quite accurate to classify diverse analogical sentences. Section 6 points out some candidate applications, before concluding.

## 2 Related work

Although lexical analogies have been more thoroughly studied recently due to the advent of distributed representations ([6, 11, 36, 37, 20, 27, for example], to cite but a few papers), few works have focused on *sentence* analogies.

In [39], the authors try to show how existing embedding approaches are able to capture analogies between sentences. They view analogies between pairs of sentences in very broad terms, which is reflected in the various corpora that they have constructed. For example, in order to create quadruples of analogous sentences they replace individual words with the corresponding words from the Google word analogy dataset [25]. Sentences that share more or less common semantic relations (entailment, negation, passivization, for example) or even syntactic patterns (comparisons, opposites, plurals among others) are also considered analogous. Using these datasets, analogies are evaluated using various embeddings, such as GloVe [28], word2vec [25], fastText [4, 26], etc. showing that capturing syntactic analogies which were based on lexical analogies from the Google word analogies dataset is more effective with their models than recognising analogies based on more semantic information.

In [10] the authors focus on the task of identifying the correct answer to a question from a pool of candidate answers. More precisely, given a question $q$ the goal is to select the answer $a_i \in A$ from a set $A$ of candidate answers. In order to do so they leverage analogies between $(q, a_i)$ and various pairs of what they call *"prototypical"* question/answer pairs, assuming that there is an analogy between $(q, a_i)$ and the prototypical pair $(q_p, a_p)$. The goal is to select the candidate answer $a_i^* \in A$ such that:

$$a_i^* = \arg min_i(||(q_p - a_p) - (q - a_i)||)$$

exploiting the properties of arithmetic proportion and analogical dissimilarities. The authors limit the question/answer pairs to $wh-$ questions from WikiQA and TrecQA. They use a Siamese bi-GRUs as their architecture in order to represent the four sentences. In this manner the authors learn embedding representations for the sentences which they compare against various baselines including random vectors, *word2vec, InferSent* and *Sent2Vec* obtaining better results with the WikiQA corpus.

Most of the tested sentence embedding models succeed in recognizing syntactic analogies based on lexical ones, but had a harder time capturing analogies between pairs of sentences based on semantics.

## 3 Formal framework

Analogies are often expressed in terms of analogical proportions. An analogical proportion over a set $X$ of items (Boolean vectors, real-valued vectors, words or even sentences in natural language) is a quaternary relation involving 4 elements, $a, b, c, d \in X$, often denoted $a : b :: c : d$ and should be read "$a$ is to $b$ as $c$ is to $d$", obeying postulates. Depending on the postulates, *strong* and *weak* forms of analogical proportions can be distinguished, leading to different Boolean and numerical representations.

### 3.1 Strong analogical proportions

Classically, analogical proportions are supposed to obey the 3 following first-order logic postulates (e.g., [18]) $\forall a, b, c, d \in X$,[1]which are satisfied by numerical proportions:

1. $a : b :: a : b$ (*reflexivity*);
2. $a : b :: c : d \to c : d :: a : b$ (*symmetry*);
3. $a : b :: c : d \to a : c :: b : d$ (*central permutation*).

These postulates have straightforward consequences like:

- $a : a :: b : b$ (*identity*);
- $a : b :: c : d \to b : a :: d : c$ (*internal reversal*);
- $a : b :: c : d \to d : b :: c : a$ (*extreme permutation*);
- $a : b :: c : d \to d : c :: b : a$ (*complete reversal*).

Among the 24 permutations of $a, b, c, d$, the previous postulates lead to 3 distinct classes each containing 8 syntactically different proportions regarded as equivalent due to the postulates. Thus $a : b :: c : d$ has in its class $c : d :: a : b$, $c : a :: d : b$, $d : b :: c : a$, $d : c :: b : a$, $b : a :: d : c$, $b : d :: a : c$, and $a : c :: b : d$. But $b : a :: c : d$ and $a : d :: c : b$ are not in the class of a:b::c:d and are in fact elements of two other classes.

A typical example of an analogical proportion over $X = \mathbb{R}$ is the arithmetic proportion defined as:

$$a : b :: c : d \text{ holds if and only if } a - b = c - d$$

easily extended to a real-valued vector space $X = \mathbb{R}^n$ with the same definition. From a geometric viewpoint in $\mathbb{R}^n$, it means that $(a, b, c, d)$ is a parallelogram [35]. In practice, it may be weakened into an approximate equality $a - b \approx c - d$ using some tolerance.

[1] In the following, we omit the universal quantifier for readability.

Considering now $X$ as the Boolean set $\mathbb{B} = \{0, 1\}$, various equivalent Boolean formulas satisfy the postulates of analogical proportion over $X$. One of them making explicit that "$a$ differs from $b$ as $c$ differs from $d$ (and vice-versa)" [24] is:

$$a : b :: c : d =_{def} ((a \wedge \neg b) \equiv (c \wedge \neg d)) \wedge ((\neg a \wedge b) \equiv (\neg c \wedge d))$$

It is easy to check that this formula is only valid for the 6 valuations in Table 1. As shown in [31], this set of 6 valuations is the *minimal* Boolean model obeying the 3 postulates of analogy. It can be seen on this table that 1 and 0 play a symmetrical role, which makes the definition *code-independent*. This is formally expressed in $X = \mathbb{B}$ with the negation operator as: $a : b :: c : d \to \neg a : \neg b :: \neg c : \neg d$.

| |
|---|
| $0 : 0 :: 0 : 0$ |
| $1 : 1 :: 1 : 1$ |
| $0 : 1 :: 0 : 1$ |
| $1 : 0 :: 1 : 0$ |
| $0 : 0 :: 1 : 1$ |
| $1 : 1 :: 0 : 0$ |

**Table 1.** Valid valuations for the strong Boolean analogical proportion

To deal with items represented by Boolean vectors, it is straightforward to extend this definition from $\mathbb{B}$ to $\mathbb{B}^n$ with:

$$a : b :: c : d =_{def} \forall i \in [1, n], a_i : b_i :: c_i : d_i$$

This is useful when words are represented by means of key terms appearing in their dictionary definition. For instance, using the 5 key terms $mammal$, $bovine$, $equine$, $adult$, $young$, Table 2 explains why $a : b :: c : d = cow : calf :: mare : foal$ can rigorously be considered as a valid analogy (we recognize patterns of Table 1, vertically, in Table 2). More generally, series of mutually exclusive properties such as $bovine$ / $equine$, $adult$ / $young$, as encountered in taxonomic trees, induce analogical proportions [3].

| | $mammal$ | $bovine$ | $equine$ | $adult$ | $young$ |
|---|---|---|---|---|---|
| cow | 1 | 1 | 0 | 1 | 0 |
| calf | 1 | 1 | 0 | 0 | 1 |
| mare | 1 | 0 | 1 | 1 | 0 |
| foal | 1 | 0 | 1 | 0 | 1 |

**Table 2.** A Boolean validation of $cow : calf :: mare : foal$

### 3.2 Weak analogical proportions

However, the central permutation postulate (3) is debatable for analogical proportions involving two *conceptual spaces* [2]. Indeed in the following example that involves nationalities and beverages, while "wine is to French people as beer is to English people" is acceptable, "wine is to beer as French people is to English people" sounds weird.

It is then legitimate *to abandon the central permutation postulate* and to replace it by the internal reversal property which is still a widely accepted property of analogical proportion, leading to a *weaker* definition of analogical proportion:

1 $a : b :: a : b$ (*reflexivity*);
2 $a : b :: c : d \to c : d :: a : b$ (*symmetry*);
4 $a : b :: c : d \to b : a :: d : c$ (*internal reversal*).

*Complete reversal* ($a : b :: c : d \to d : c :: b : a$) is still a consequence of this weaker set of postulates. Clearly a strong analogical proportion (in the sense of (1)-(2)-(3)) is also a weak proportion. According to postulates (1)-(2)-(4) $a : b :: c : d$ can be written only under 4 equivalent forms rather than 8: $a : b :: c : d$, $c : d :: a : b$, $d : c :: b : a$, and $b : a :: d : c$. Despite one might be tempted to have $a : a :: b : b$ (*identity*), this is no longer deducible from the postulates.

From a computational linguistics viewpoint (see for instance [6, 11, 21]), a proportion $a : b :: c : d$ is often understood as:

*for some binary relation $R$, $R(a, b) \wedge R(c, d)$ holds.*

where $R$ is a latent relation with a semantic or discourse connotation. This definition perfectly fits with postulates (1)-(2)-(4). As an example, consider the following pairs of sentences:

John sneezed loudly (a). Mary was startled (b).
Bob took an analgesic (c). His headache stopped (d).

where one could argue that $R$ is a kind of causal relation. Indeed, internal reversal holds because $R^{-1}(b, a) \wedge R^{-1}(d, c)$ holds. In the example above $R^{-1}$ would be an explanation relation although explicit discourse markers should be utilized:

Mary was startled (b), because John sneezed loudly (a).
Bob's headache stopped (d) because he took an analgesic (c).

In practice, the fact that $R(a, b) \wedge R(c, d)$ holds does not entail that there also exists a semantically meaningful $S$ such that $S(a, c) \wedge S(b, d)$ holds: that is why the central permutation postulate is not relevant here. It would make no sense in the above example.

The *minimal* Boolean model of postulates (1)-(2)-(4) is constituted with the 4 first lines of Table 1, still satisfying code-independence. A Boolean expression of the weak analogical proportion $a : b ::_w c : d$ is given by

$$a : b ::_w c : d = (a \equiv c) \wedge (b \equiv d)$$

Note that $a \wedge b \equiv c \wedge d$ and $a \vee b \equiv c \vee d$ are simple examples of formulas satisfying (1)-(2)-(4). But their Boolean models include more than the 4 first lines of Table 1 even if they are false for the last two lines of this Table.

When working on $X = \mathbb{R}$ rather than $X = \mathbb{B}$, a weak analogical proportion can be defined by:

$$a : b ::_w c : d \text{ holds } \begin{cases} \text{if } a = b = c = d \\ \text{if } a - b = c - d \quad \text{when } Cond \text{ is true} \end{cases}$$

where $Cond$ stands for $(a < b \ \wedge \ c < d) \vee (a > b \ \wedge \ c > d)$ with $b \neq c$. It removes the situation where $a - b = c - d = 0$ for $b \neq c$. It satisfies (1)-(2)-(4), but not (3) (e.g., we have $.3 < .9$ and $.7 < .8$ but not $.3 < .7$ and $.9 < .8$), nor $a : a :: b : b$ or $a : b :: b : a$. It has also the advantage to keep $a, b, c, d$ distinct, which would not be the case if $a : b ::_w c : d$ is defined by the straightforward option '$a = c$ and $b = d$'. Obviously in practice, $=$ could be replaced by $\approx$.

### 3.3 Analogical proportions and implication

A relation of particular interest between items such as sentences (or words) is the entailment relation: "$a$ entails $b$ as $c$ entails $d$". So we may wonder if a simplistic modeling of "$a$ entails $b$" in terms of material implication $(a \to b)$ would lead to a weak proportion.

In fact, $(a \to b) \land (c \to d)$ is not satisfactory as, in the Boolean universe, it is false for valuation $(1010)$ and true for the other patterns that a strong or weak analogical proportion fulfill (together with 4 other patterns: $(0001), (0100), (0111), (1101)$). Obviously this formula does satisfy neither central permutation nor internal reversal.

In that respect, a better option could be to consider

$$Imp(a, b, c, d) = [(a \to b) \land (c \to d)] \lor [(b \to a) \land (d \to c)]$$

which satisfies the postulates of a weak analogical proportion. However, in the Boolean universe, this formula is false only for 2 patterns $(0110), (1001)$ (which are known as maximizing analogical dissimilarity [23, 24]), and true for all the 14 remaining patterns. However, one can recover $a : b :: c : d$ as well as $a : b ::_w c : d$ from it. Namely,

$$a : b :: c : d = Imp(a, b, c, d) \land Klein(a, b, c, d)$$

where $Klein(a, b, c, d) = (a \equiv b) \equiv (c \equiv d)$ is an operator introduced by S. Klein [8], true for the 6 patterns that make true an analogical proportion, plus the two forbidden patterns $(0110), (1001)$. In the Boolean universe, weak analogical proportions are recovered as

$$a : b ::_w c : d = Imp(a, b, c, d) \land Par(a, b, c, d)$$

where $Par(a, b, c, d) = ((a \land b) \equiv (c \land d)) \land ((\neg a \land \neg b) \equiv (\neg c \land \neg d)$ is a Boolean operator named *paralogy* [31].

The above equalities show that one cannot have a pure implication-based view of analogical proportions. Moreover $Imp(a, b, c, d)$ is much weaker than $a : b ::_w c : d$ and cannot be a particular case of it.

### 3.4 Analogical proportions and sentences

We distinguish two types of analogies between sentences. The first one is a natural extension of word analogies. Starting from the fact that $a : b :: c : d$ is an analogy between lexical items $a, b, c$ and $d$, then sentences $s_1, s_2$ are analogous if they link $a, b$ and $d, c$ with the same predicate $R$ — so $R(a, b) \land R(c, d)$. A simple example would be: *French people drink wine ($s_1$) English people drink beer ($s_2$)* with *French:wine :: English:beer*. It is worth noticing that already a pair of sentences making a parallel between two situations and involving words or phrases in analogical proportions, may provide an argumentative support. For example, let us consider the two sentences *Polio vaccine protects against poliomyelitis. h1n1vaccine protects against influenza.* Clearly the pairs $(a, b) =$ (*Polio, poliomyelitis*) and $(c, d) =$ (*h1n1 vaccine, influenza*) make an analogical proportion $a : b :: c : d$ with respect to $vaccine$, $disease$, $virus$ and $bacteria$, as can be checked on first four lines of Table 3. If we also consider the sentence "BCG vaccine protects against tuberculosis" associated with the pair

| | $vaccine$ | $disease$ | $virus$ | $bacteria$ |
|---|---|---|---|---|
| a | 1 | 0 | 0 | 0 |
| b | 0 | 1 | 1 | 0 |
| c | 1 | 0 | 0 | 0 |
| d | 0 | 1 | 1 | 0 |
| a' | 1 | 0 | 0 | 0 |
| b' | 0 | 1 | 0 | 1 |

**Table 3.** The vaccine example

$(a', b') = (BCG\, vaccine,\ tuberculosis)$, $a' : b' :: c : d$ does not hold with respect to $virus$ and $bacteria$ (see Table 3), and then $(a', b')$ appears to be a poorer support for $(c, d)$ than $(a, b)$. It would be still worse if one considers the sentence "umbrellas protect against the rain" instead of $(a', b')$! Besides, note that in all the above sentences the same relation $R = protects\, against$ takes place, and both $R(a, b)$ and $R(c, d)$ hold. In case of synonyms used in different sentences, analogy may become a matter of degree (e.g.,[5]). When sentences have a different number of words, one may apply the approach by Miclet et al. [22] for aligning the key terms. Lastly, a parallel between 2 situations may involve 4 sentences making an analogical proportion, as in the example "some cetaceans eat phytoplankton; other cetaceans are carnivorous; some mammals eat grass; other mammals are carnivorous".

The second type of sentential analogies is not limited to $a, b, c$ and $d$ being extension of word analogies but instead they represent more complex sentences with $R$ being a latent relation linking $a, b$ and $c, d$. This type of sentential analogies was briefly evoked in subsection 3.2 wherein some examples involving causal relations were provided. Here is another example involving a relation that is more temporal in nature:

> Mary was working on her new book (a), while John was washing this afternoon's dishes (b).
> Bob was waiting for the bus at the bus stop (c). At the same time, workers were making loud noises on the street (d).

As we can see, in this case reflexivity, symmetry, internal reversal and complete reversal (subsection 3.2) also hold while we cannot say the same thing for central permutation.

The latent relation linking two pairs of sentences can take several forms. Natural candidates include entailment or various discourse relations. In Section 4 of this paper, we use the Penn Discourse TreeBank[2] (PDTB), but other datasets, such as the Stanford Natural Language Inference (SNLI) Corpus[3] could be used as well.

## 4 Experimental context - Evaluation metrics

Let us move to an empirical validation of our approach using word embeddings, and sentence embeddings[4].

[2] https://www.seas.upenn.edu/ pdtb/

[3] https://nlp.stanford.edu/projects/snli/

[4] All our datasets and python code used in this paper are freely available at https://github.com/arxaqapi/analogy-classifier.

### 4.1 Datasets

*Mixing google analogies with template sentences* Given a set of words $W$, the notion of analogical proportion between words is not *formally* defined but we have many examples of well-accepted (cognitively valid) word analogies like $man : king :: woman : queen$. One of the most well-known datasets of word analogies, originally from Google, can be downloaded at http://download.tensorflow.org/data/questions-words.txt

Starting from Google word analogies dataset, we create our own sentences analogies dataset by separating the lexical items therein into different categories. Template sentences are then created with placeholders to be replaced by the correct category of the Google word analogy dataset. With this method sentences $S_A$, $S_B$, $S_C$ and $S_D$ are created so that they meet the requirement $R(S_A, S_B) \wedge R(S_C, S_D)$. Obviously, we expect our classifiers to be successful on this artificial dataset: at this stage, this is just a toy example to check that our initial assumption (weak analogy definition) is not defeated. Below are the categories with examples used

- **Capital Common Cities.** $S_A =$ “She arrived yesterday in *London*” and $S_B =$ “She just landed in *England*” .
- **City in State.** $S_A =$“Citizens in *Chicago* are more likely to vote”, $S_B =$“Citizens in *Illinois* are more likely to vote”.
- **Currencies.** $S_A =$ “What is the currency used in *Japan*?”, $S_B =$ “Which country uses the *yen*?”.
- **Family.** The possessives “his”, “her” and the nouns “he”, “she”, are skipped. For instance in $S_A =$“His *father* could not be present at the annual family gathering”, $S_B =$“His *mother* could not be present at the annual family gathering”.
- **Nationality adjectives.** $S_A =$“The culture in *Chile* is very rich”, $S_B =$“The *Chilean* culture is very rich”.
- **Opposites.** Sentences containing an adjective listed in the Google analogy dataset. The adjective is labelled and replaced with it’s antonym thus creating a pair of sentences. $S_A =$“I was *aware* in which direction he was going”, $S_B =$“I was *unaware* in which direction he was going”.

Once generated, we get a total set of $52,185$ sentence quadruples. Using the extension methods depicted in Subsection 3.2, the dataset is extended by a factor 12, with 4 valid analogies per sentence quadruple, using the *symmetry*, *internal reversal* and *complete reversal* permutations and 8 invalid analogies, using the same permutations applied to the quadruples $b : a :: c : d$ and $c : b :: a : d$. We now have a dataset of size $52,185 * 12 = 626,220$ sentences.

*Penn Discourse TreeBank dataset* The second dataset that we use is the Penn Discourse TreeBank (PDTB) [33].[5] Our goal was to try a preliminary series of experiments to determine whether our approach could identify pairs of sentences that are linked with the same latent relation $R$ where $R$ in this particular case is a discourse relation. PDTB contains more than 36,000 pairs of sentences annotated with discourse relations. Relations can be explicitly expressed via a discourse marker, or implicitly expressed in which

[5] At this stage, we used PDTB version 2.1.

case no such discourse marker exists and the annotators provide one that more closely describes the implicit discourse relation. Relations are organized in a taxonomy which contains 4 classes of relations (Temporal, Contingency, Expansion and Comparison) as well as several types and subtypes of relations (for example Cause, Condition, Contrast, etc). As an example, for the Temporal relation type, we have pairs of sentences like: $S_A =$ “The results were announced”, $S_B =$ “The stock market closed” and $S_C =$ “That happens”, $S_D =$ “Nervous stock investors dump equities and buy treasurys”.

In this series of preliminary results we used primarily the four classes of relations although we also experimented with the implicit/explicit nature of relations. In total we selected 25,000 random quadruplets for a total of 300,000 instances. Each quadruple of sentences was accompanied with a Boolean class indicating whether the quadruple constituted a valid analogy — that is pairs $(a, b)$ and $(c, d)$ are linked with the same relation — or not.

### 4.2 Embedding techniques

A word embedding $\omega$ is an injective function from a set of words $W$ to a real-valued vector space $\mathbb{R}^n$ (usually $n \in \{50, 100, 200, 300\}$ but there is no real limitation). There are well-known word embeddings such as word2vec [25], GloVe [28], BERT [9], fastText [26], etc. It is standard to start from a word embedding to build a sentence embedding. Sentence embedding techniques represent entire sentences and their semantic information as vectors. There are diverse sentence embedding techniques. In this paper, we focus on 2 techniques relying on an initial word embedding.

- The simplest method is to average the word embeddings of all words in a sentence. Although this method ignores both the order of the words and the structure of the sentence, it performs well in many tasks. So the final vector has the dimension of the initial word embedding.
- The other approach, suggested in [1], makes use of the Discrete Cosine Transform (DCT) as a simple and efficient way to model both word order and structure in sentences while maintaining practical efficiency. Using the inverse transformation, the original word sequence can be reconstructed. A parameter $k$ is a small constant that needs to be set. One can choose how many features are being embedded per sentence by adjusting the value of $k$, but undeniably increases the final size of the sentence vector by a factor $k$. If the initial embedding of words is of dimension $n$, the final sentence dimension will be $= n * k$ (see [1] for a complete description). In this paper, we chose $k = 1$ as suggested in [39].

### 4.3 Classifiers

Because we do not rely on any parallelogram-like formula to check whether 4 sentences build an analogy, we move to machine learning to “learn”, in some sense, the formula. In fact, we classify a quadruple of 4 sentences as a valid or non valid analogy. We tried two classical methods which have been successfully used for word analogy classification [20]: Random Forest (RF) and Convolutional Neural Networks (CNN). CNN has been popular for image classification but has also been used for text classification as it could extract and select important ngrams for classification [17]. RF is relatively accurate at classification and is fast [7]. On a 10 core CPU, CNN took about 12 CPU hours (one-hour real-time) to train but RF took only about 18 CPU minutes and real-time to train.

The parameters for RF are 100 trees, no maximum depth, and a minimum split of 2. With CNN, stacking together the 4 vectors , with $n$ components corresponding to a quadruple $a, b, c, d$ of sentences, we get a matrix $n \times 4$ that we are going to process as we would do for an *image*. With filters respecting the boundaries of the 2 pairs, this is the structure of the CNN:

- 1st layer (convolutional): 128 filters of size $height \times width = 1 \times 2$ with strides $(1, 2)$ and Relu activation. This means that we are working component by component (with a vertical stride of 1) and we move from pair $(a, b)$ to pair $(c, d)$ with horizontal stride 2.
- 2nd layer (convolutional): 64 filters of size $(2, 2)$ with strides $(2, 2)$ and Relu activation, reducing the dimension before going to the final dense layer.
- 3rd layer (dense): one output with sigmoid activation.

# 5 Results

Below we present experimental results for two datasets: generated sentences and PDTB.

## 5.1 CNN and RF results for generated sentences

The results in Table 4 are obtained with 10 epochs for the CNN. The Average Accuracy is computed from the average of the 10 folds accuracy for each method. The results are already extremely good (over 98%) even with a low word embedding size for both CNN and RF. The accuracy in CNN increases if we increase the size of the word vectors as more "details" are encoded in the vector so that the CNN captures more of the semantics. For RF, increasing the size of the word vectors provides no increase in the accuracy. Given that the increase in accuracy is at most 2%, we suggest using vectors of size 50 (GloVe) or reduce the fastText vectors from 300 to 50 to reduce the computation in CNN. The authors in [39] also used GloVe average method for sentence embedding and the highest accuracy is 90%, regardless of which parallelogram-inspired formula were used.

## 5.2 CNN and RF results for PDTB dataset

The accuracies for both CNN and RF are around 60%, much lower for the PDTB dataset (see Table 5) compared to the generated sentences. This is to be expected as the PDTB sentences are much more semantic/pragmatic in nature so it is much more difficult to capture the relationship between the sentences using a simple average embedding technique. The $F1$ values of RF are very low (at most 0.09), suggesting that it has difficulty identifying positive results. CNNs on the other hand perform much better achieving 0.52 $F1$ for valid analogies using fastText, but the overall highest accuracy is only 66.69%.

| ML | Word embedding size | | | | |
|---|---|---|---|---|---|
| method | G50 | G100 | G200 | G300 | F300 |
| CNN-AVG | 98.39% | 99.76% | 99.96% | 99.97% | 99.91% |
| RF-AVG | 99.97% | 99.97% | 99.97% | 99.97% | 99.98% |
| CNN-DCT | 68.23% | 68.30% | 68.31% | 68.31% | 68.24% |
| RF-DCT | 67.14% | 67.14% | 67.14% | 67.14 | 67.10% |

**Table 4.** Average accuracies (10 folds) for CNN (10 epochs) and RF for generated dataset using GloVe vector size 50 to 300 (G50 to G300) and fastText (F300) for average and DCT sentence embedding.

| ML | Word embedding size | | | | |
|---|---|---|---|---|---|
| method | G50 | G100 | G200 | G300 | F300 |
| CNN-AVG | 66.49%(0.42) | 66.64%(0.42) | 66.61%(0.40) | 66.36%(0.41) | 66.69%(0.52) |
| RF-AVG | 62.96%(0.02) | 63.79%(0.02) | 64.08%(0.02) | 64.33%(0.02) | 65.05%(0.01) |
| CNN-DCT | 66.46%(0.12) | 66.49%(0.10) | 65.31%(0.34) | 66.10%(0.15) | 66.05%(0.21) |
| RF-DCT | 58.74%(0.04) | 58.91%(0.04) | 58.92%(0.04) | 58.85%(0.04) | 61.89%(0.09) |

**Table 5.** Average accuracies and $F1$ (10 folds) for CNN 10 epochs and RF for PDTB dataset using GloVe and fastText for average and DCT sentence embedding.

Our results are an improvement with respect to [10, 39], although a strict comparison is difficult because not only do we use different corpora, but also the experimental setup is different. For more semantic sentences, the work reported in [39] achieves an accuracy of 0.43 in the unconstrained scenario, while we have achieved an accuracy of 0.66. In [39], a constrained scenario selected the true answer from six sentences, while the unconstrained scenario selected the true answer from the ”entire corpus”.

## 6 Candidate applications

Analogy-making permeates human cognition and is a mechanism that is very often used in human communication. In terms of computational linguistics the study of analogies has almost exclusively been limited to the detection of word analogies [6, 11, 36, 37, 20, 27, for example]. Nonetheless analogies go beyond the lexical level and can exist between sentences as well as longer discourse units. In the following we describe some areas of computational linguistics that could benefit from analogy-making.

- ***Discourse:*** one of the problems that current chatbots face is that they lack discourse coherence. Disposing of a mechanism that is able to handle sentence analogies, allows us to better select following sentences in a generated text yielding better overall coherence. Consider for example that in a discourse $d = (s_1, \ldots, s_n)$ where $s_i$ represents elementary discourse units, we need to choose the next sentence from a set of candidate sentences $C$ that a chatbot could have generated so as to maximize coherence. In order to do so we can rely upon the formal framework presented in Section 3 in order to choose $s_c \in C$ such that $s_c = \arg min(||(\pi_i - \pi_j) - (s_n - s_c)||)$ where $\pi_i : \pi_j :: s_n : s_c$. In order to bootstrap learning we can use resources such as PDTB. Given that $\pi_i, \pi_j, s_n$ and $s_c$ are learnt representations we can imagine a more complicated scenario in which $\pi_i$ and $s_n$ are substituted for representations of previous discourse. Corpora such as the RST which model larger contexts could be used.

- ***Question-answering:*** A similar approach has already been used in the context of question answering by [39] (see Section 2). Better selection of answers is achieved by hypothesizing that an answer is more plausible for that question if the pair $(q, a)$ is analogical to other "prototypical" pairs of questions and answers. Extensions of this work could include representations of larger context as well as less prototypical pairs.
- ***Computational creativity:*** Being able to identify and propose new analogies can be very useful in understanding and advancing computational creativity. In the context of generating more "creative" text, such as poem generation [38], relaxing to a certain degree the constraint of minimization mentioned above could yield creative analogies.
- ***Argumentation:*** Capturing analogies that go beyond the sentential level could be very useful for argumentation. A conversational agent trying to convince their interlocutor that, for example, using a car is not anodyne since it pollutes the environment and ultimately is a cause of death for many people, could draw the analogy between driving a car and smoking. Smoking was widely accepted but studies have shown that it is detrimental to the health of active and passive smokers. The same has been shown for car pollution, so cars should be restricted or banned altogether under certain circumstances. Achieving such a goal requires representations that are capable to handle larger chunks of text.

## 7 Conclusion

In this paper, we have provided the basis of a weak analogical proportion theory, removing the classical central permutation postulate. Our new postulates better reflect what is generally considered as a natural language analogical proportion between sentences. From a machine learning perspective, the new system is also used to rigorously extend a set of examples. Using standard embedding techniques for sentences, we have tested our approach to classify 4 sentences into valid and invalid proportions. Preliminary experiments using simple architectures show that we can achieve an accuracy of 0.66 (with an $F1$ of 0.52 for valid analogies) for sentential analogies based on latent semantic and pragmatic similarity using the PDTB corpus. In the future we plan to perform further experiments using transformers and BERT embeddings in order to investigate which of the aforementioned postulates (central permutation, internal reversal) transfer to natural language datasets. By crossing powerful real-valued embeddings with a more semantic ontology-based representation of words, the new formal setting paves the way to hybrid approaches where analogical inference could be done on natural language corpora.

# References

1. N. Almarwani, H. Aldarmaki, and M. Diab. Efficient sentence embedding using discrete cosine transform. In *EMNLP*, pages 3663–3669, 2019.
2. N. Barbot, L. Miclet, and H. Prade. Analogy between concepts. *Artif. Intell.*, 275:487–539, 2019.
3. N. Barbot, L. Miclet, H. Prade, and G. Richard. A new perspective on analogical proportions. In*Proc.15th Europ. Conf. Symb. & Quantit. Appr. to Reas. with Uncert. (ECSQARU)*. LNCS 11726, 163-174, Springer, 2019.
4. P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov. Enriching word vectors with subword information. In *Transactions of the Association for Computational Linguistics*, page 135–146, 2017.
5. M. Bounhas, H. Prade, and G. Richard. Analogy-based classifiers for nominal or numerical data. *Int. J. Approx. Reasoning*, 91:36–55, 2017.
6. Z. Bouraoui, S. Jameel, and S. Schockaert. Relation induction in word embeddings revisited. In *COLING*. 1627-1637, Assoc. Computat. Ling., 2018.
7. L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, Oct 2001.
8. M. Couceiro, N. Hug, H. Prade, and G. Richard. Ana-logy-preserving functions: A way to extend Boolean samples. In *Proc. IJCAI*, Melbourne, 1575-1581, 2017.
9. J. Devlin, M.-W. Chang, K. Lee, K. Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018.
10. A. Diallo, M. Zopf, and J. Fürnkranz. Learning analogy-preserving sentence embeddings for answer selection. In *Proc. 23rd Conf. Computational Natural Language Learning*. 910 - 919, Assoc. Computat. Ling., 2019.
11. A. Drozd, A. Gladkova, and S. Matsuoka. Word embeddings, analogies, and machine learning: Beyond king - man + woman = queen. In *COLING*, 3519-3530, 2016.
12. R. Fam and Y. Lepage. Tools for the production of analogical grids and a resource of n-gram analogical grids in 11 languages. In *LREC*, 2018.
13. R. M. French and D. Hofstadter. Tabletop: An emergent, stochastic model of analogy-making. In *Proc. 13th Annual Conf. of the Cognitive Science Society*, pages 175–182. Lawrence Erlbaum, 1991.
14. D. Gentner, K. J. Holyoak, and B. N. Kokinov, editors. *The Analogical Mind: Perspectives from Cognitive Science*. MIT Press, Cambridge, MA, 2001.
15. M. Hesse. *Models and Analogies in Science*. 1st ed. Sheed & Ward, London, 1963; 2nd augmented ed. University of Notre Dame Press, 1966.
16. D. Hofstadter and M. Mitchell. The Copycat project: A model of mental fluidity and analogy-making. In *Fluid Concepts and Creative Analogies: Computer Models of the Fundamental Mechanisms of Thought*, pages 205–267. Basic Books, Inc., 1995.
17. Alon Jacovi, Oren Sar Shalom, and Yoav Goldberg. Understanding convolutional neural networks for text classification. *CoRR*, abs/1809.08037, 2018.
18. Y. Lepage. Analogy and formal languages. *Electr. Notes Theor. Comput. Sci.*, 53, 2001.
19. Y. Lepage and E. Denoual. Purest ever example-based machine translation: Detailed presentation and assessment. *Machine Translation*, 19(3-4):251–282, 2005.
20. S. Lim, H. Prade, and G. Richard. Solving word analogies: A machine learning perspective. In*Proc.15th Europ. Conf. Symb. & Quantit. Appr. to Reas. with Uncert. (ECSQARU)*. LNCS 11726, 238-250, Springer, 2019.
21. H. Lu, Y Wu, and K.H. Holyoak. Emergence of analogy from relation learning. In *Proc. of the National Acad. of Sciences*, volume 116, pages 4176–4181, 2019.
22. L. Miclet, N. Barbot, and B. Jeudy. Analogical proportions in a lattice of sets of alignments built on the common subwords in a finite language. In *Computational Approaches to Analogical Reasoning: Current Trends*, pages 245–260. Springer, 2014.

23. L. Miclet, S. Bayoudh, and A. Delhay. Analogical dissimilarity: Definition, algorithms and two experiments in machine learning. *JAIR*, 32:793–824, 2008.
24. L. Miclet, H. Prade. Handling analogical proportions in classical logic and fuzzy logic settings. In*Proc. 10th ECSQARU Conf.*, LNCS 5590, 638-650, Springer, 2009.
25. T. Mikolov, K. Chen, G. S Corrado, and J. Dean. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781, 2013.
26. T. Mikolov, E. Grave, P. Bojanowski, C. Puhrsch, and A. Joulin. Advances in pre-training distributed word representations. In *Proc. of LREC*, 2018.
27. P.-A. Murena, M. Al-Ghossein, J.-L. Dessalles, and A. Cornuéjols. Solving analogies on words based on minimal complexity transformation. In *Proc. 29th Int. Joint Conf. Artif. Intellig.*, 1848-1854, 2020.
28. J. Pennington, R. Socher, and Ch. D. Manning. Glove: Global vectors for word representation. In *EMNLP*, pages 1532–1543, 2014.
29. H. Prade and G. Richard. Analogical proportions and analogical reasoning - An introduction. In D. W. Aha and J. Lieber, editors, *Proc. 25th Int. Conf. on Case-Based Reasoning Research and Development (ICCBR'17), Trondheim, Norway, June 26-28*, volume 10339 of *LNAI*, pages 16–32. Springer, 2017.
30. H. Prade and G. Richard. Analogical proportions: Why they are useful in AI. In *Proc. 30th Int. Joint Conf. on Artificial Intelligence (IJCAI-21), Montreal, Aug. 21-26*, 2021.
31. H. Prade, G. Richard. Analogical proportions: From equality to inequality. *Int. J. Approx. Reasoning*,101:234–254, 2018.
32. H. Prade, G. Richard, eds. *Computational Approaches to Analogical Reasoning: Current Trends*. Springer, 2014.
33. R. Prasad, N. Dinesh, Alan Lee, E. Miltsakaki, L. Robaldo, A. Joshi, and B. Webber. The Penn Discourse TreeBank 2.0. In *LREC 08*, May 2008.
34. R. Rhouma and Ph. Langlais. Experiments in learning to solve formal analogical equations. In *Proc. 26th Int. Conf. on Case-Based Reasoning ICCBR'18, Stockholm*, pages 612–626. LNCS 11156, Springer, 2018.
35. D. E. Rumelhart and A. A. Abrahamson. A model for analogical reasoning. *Cognitive Psychol.*, 5:1–28, 2005.
36. P. D. Turney. A uniform approach to analogies, synonyms, antonyms, and associations. In *COLING*, pages 905–912, 2008.
37. P. D. Turney. Distributional semantics beyond words: Supervised learning of analogy and paraphrase. *TACL*, 1:353–366, 2013.
38. T. Van de Cruys. Automatic poetry generation from prosaic text. In *Proc. of ACL*, 2020.
39. X. Zhu and G. de Melo. Sentence analogies: Linguistic regularities in sentence embeddings. In *COLING*, 2020.