2012 Seventh International Conference on Knowledge, Information and Creativity Support Systems

Lexical Creativity from Word Associations

Oskar Gross, Hannu Toivonen,
Jukka M Toivanen and Alessandro Valitutti
Department of Computer Science and HIIT
University of Helsinki, Finland
Email: {oskar.gross, hannu.toivonen, jukka.toivanen, alessandro.valitutti}@cs.helsinki.fi

***Abstract*—A fluent ability to associate tasks, concepts, ideas, knowledge and experiences in a relevant way is often considered an important factor of creativity, especially in problem solving. We are interested in providing computational support for discovering such creative associations.**

In this paper we design minimally supervised methods that can perform well in the *remote associates test* (RAT), a well-known psychometric measure of creativity. We show that with a large corpus of text and some relatively simple principles, this can be achieved. We then develop methods for a more general word association model that could be used in lexical creativity support systems, and which also could be a small step towards lexical creativity in computers.

I. Introduction

A fluent ability to associate tasks, concepts, ideas, knowledge and experiences in a relevant way is often considered an important factor of creativity, especially in problem solving. We are interested in providing computational support for discovering such creative associations. As a first step in this direction, we aim to design minimally supervised methods that perform well in the *remote associates test* (RAT) [1], a well-known psychometric measure of creativity.

The remote associates test is based on finding associations between words. In a RAT question, the subject is presented three *cue words*, e.g., 'coin', 'quick', and 'spoon'. Her task is then to find a single *answer word* that is related to all of the cue words. (Try to think of one! The answer word is given at the end of this paper.)

Accordingly our focus in this paper is on lexical creativity. While this may be considered a limited area of associative creativity, it has great potential in those tools for creativity support or problem solving that are based on verbal information, and also in creative language use such as computational poetry [2].

Our aim is to device methods that not only score well on RATs, but also require a minimum amount of explicit knowledge as input. We rely on corpus-based methods that learn word associations from large masses of text with statistical methods. Independence of knowledge bases, lexicons, or grammars also makes the methods easier to be applied to different languages.

In this paper, we first present a simple corpus-based method that has a relatively good performance (approximately 70%) on a standard RAT. RAT questions are well suited for corpus-based computational methods, and 2-gram models are largely sufficient to model and discover associations in them.

Next, inspired by the RAT setting, we propose a more general framework where more liberal, semantic associations between words can be discovered and used to support creativity, instead of the tightly bound, even idiomatic words of the RAT. To this end, we use word *co-occurrence networks*. Co-occurrence statistics of words are again computed from a document corpus, but in this case the words do not need to occur next to each other. The co-occurrence network can then be used as a simple model for creative inference, or as a component of a creativity support tool.

In the next section, we give a brief overview of the remote associates test of creativity. The contributions of this paper are then in the subsequent sections:

- We give a novel method that scores well on RAT questions of creativity using only frequencies of word collocations as its data (Section III).
- We generalize the RAT setting to more abstract relations between words and describe word co-occurrence networks for this purpose (Section IV)
- We propose a method for finding creative associations from word co-occurrence networks and give experimental results (Section V).

We review related work in Section VI, and conclude the paper in Section VII.

II. Background: Remote Associates Test of Creativity

Creativity is usually defined as the ability to find associative solutions that are novel and of high quality. S. A. Mednick [1] defines creativity as "the forming of associative elements into new combinations, which either meet specified requirements or are in some way useful". On the basis of this definition, Mednick developed the remote associates test of creativity.

The RAT measures the ability to discover relationships between concepts that are only remotely associated. It is frequently used by psychologists to measure creativity albeit there is some criticism concerning its validity in measuring creative skills. Each RAT *question* presents a set of three mutually distant words to the subject, and the subject is then asked to find a word (creatively) connecting all these words together [1]. For instance, given the cue words 'lick', 'mine', and 'shaker' the *answer* word is 'salt': 'lick salt', 'salt mine', and 'salt shaker' connect salt with each of the three words. The test is constructed so that the word associations in the

978-0-7695-4861-6/12 $26.00 © 2012 IEEE
DOI 10.1109/KICSS.2012.38

test should be familiar to people brought up in the respective culture (e.g. USA).

Most of the RAT answer words are quite uncommon. Thus, the test subject should propose answer words which are used less frequently in everyday speech to perform well on the test [1], [3]. This supports the idea that creative solutions usually are relevant and novel. The RAT performance has been established to correlate with traditional measures of IQ [4], and there is some evidence that it predicts originality during brainstorming [5]. Additionally, several studies have linked RAT results to more specific creativity-related phenomena, such as intuition and incubation [6], [7], [8]. Thus, the RAT provides arguably a well established method to assess the associative creativity in a psychological context.

III. A Computational Solution to RAT

We will now give a computational method for solving RAT tests with high accuracy, using only frequencies of word pairs in a large corpus. We will walk through the ideas using a number of experiments, so we start by describing the data we have used.

A. Background

a) RAT tests: We combined RAT tests from two sources [9], [10] and obtained a total of 212 questions. Following good practices of data analysis, this set of tests was then divided into two disjoint sets: a training set of 140 questions and a test set of 72 questions. Method development is carried out using the training set, while the validation set is used to test the performance on the final methods. This procedure avoids overly optimistic results that would be obtained by tuning and testing the methods on the same instances.

b) Corpus: Instead of a full corpus of text, we directly use Google 2-grams [11], a large, publicly available collection of 2-grams (see below).

We next formalize some of the concepts and introduce notation used in the rest of the paper.

c) Notation: n-grams, i.e., frequencies of different sequences of n words, are used widely in language modelling. For solving RATs, we use 2-grams. A 2-gram is a sequence of two words or, more formally, a vector $n = (n_1, n_2)$ of two words n_1 and n_2. The (absolute) frequency of a 2-gram $n = (n_1, n_2)$, denoted by n_c, is the number of times the sequence (n_1, n_2) of words occurred in a given corpus C_G. We denote by N the set of all 2-grams and by N_c the total of their occurrences. Let $N'_c(t)$ denote the sum of frequencies of the 2-grams that contain word t, i.e.,

$$N'_c(t) = \sum_{n \in N : t \in n} n_c.$$

In a similar way,

$$N'_c(t_1, t_2) = \sum_{n \in N : t_1, t_2 \in n} n_c = (t_1, t_2)_c + (t_2, t_1)_c$$

denotes the total of frequencies of 2-grams that contain both t_1 and t_2.

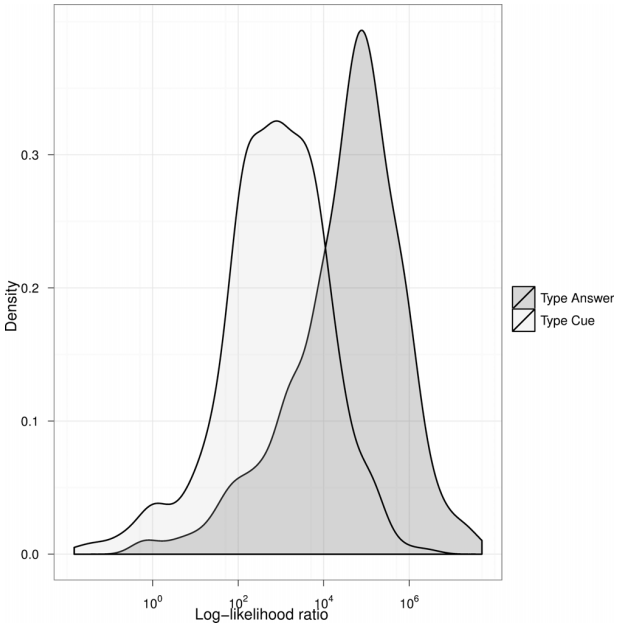

Fig. 1. The log-likelihood distribution of the different types of word pairs

Formally, a RAT is a quadruple $r = (c_1, c_2, c_3, a)$, where c_i is the ith cue word and a is the answer word.

B. Methods

a) Frequencies of RAT word pairs: The way RAT tests are constructed implies that 2-grams (c_i, a) or (a, c_i) consisting of a cue word and the answer word should have relatively high frequencies, and that 2-grams (c_i, c_j) consisting of two cue words should have relatively low frequencies.

Since the individual words in a RAT may have different frequencies, 2-grams also have different expected frequencies. So, rather than directly comparing the frequencies of 2-grams, we estimate how much the observed frequencies differ from the ones expected assuming statistical independence. We measure this deviation by the log-likelihood ratio (LLR) [12]. For this calculation, we estimate the individual frequencies of words by the number of times they occur in 2-grams.

Figure 1 shows the LLR distributions for cue word pairs ('Type cue') and for cue word, answer word pairs ('Type answer'). The cue word, answer word pairs clearly tend to be more closely related than the cue word pairs, but there is also a lot of overlap between the distributions. The difference between the distributions is statistically significant (Wilcoxon rank sum test p-value $< 2 \cdot 10^{16}$).

b) Scoring function: To solve a RAT test we need to find an answer word that is related to all of the cue words. We propose to treat each RAT question r as a probabilistic problem, where we want to find the most likely answer word a, i.e., one that maximizes the conditional probability $P(a|c_1, c_2, c_3)$.

We have

$$\begin{aligned} P(a|c_1,c_2,c_3) &= \frac{P(a,c_1,c_2,c_3)}{P(c_1,c_2,c_3)} \\ &\propto P(a,c_1,c_2,c_3) \quad (1) \\ &= P(c_1,c_2,c_3|a)P(a). \quad (2) \end{aligned}$$

Assuming that the cue words c_1, c_2, c_3 are mutually independent, as they essentially are by construction of RATs, we have

$$P(c_1,c_2,c_3|a)P(a) = P(a)\prod_{i=1}^{3} P(c_i|a). \quad (3)$$

(In machine learning, this is known as the Naïve Bayes model. It often has a good practical predictive performance even if the independence assumption does not hold [13].)

We estimate the conditional probabilities from the relative frequencies of the words in the 2-grams,

$$P(a) = \frac{N'_c(a)+1}{N_c+1},\ P(c,a) = \frac{N'_c(c,a)+1}{N_c+1}, \quad (4)$$

giving

$$P(c|a) = \frac{P(c,a)}{P(a)} = \frac{N'_c(c,a)+1}{N'_c(a)+1}. \quad (5)$$

c) Answer word search: Given a RAT test, finding the best scoring answer word a among millions of words is not straightforward. We do this in two steps. In the first step, we extract words that occur at least once with each cue word. Let this set of candidate words be Γ. In the second step, we compute the conditional probabilities of the candidate words and choose the best one, i.e.,

$$\begin{aligned} \arg\max_{a\in\Gamma} P(a)\prod_{i=1}^{3} P(c_i|a) = \\ = \arg\max_{a\in\Gamma} P(a)\prod_{i=1}^{3} \frac{(N'_c(c_i,a)+1)}{(N'_c(a)+1)}. \quad (6) \end{aligned}$$

C. Experiments

We experimented with the RAT solver using the training and test sets with 140 and 72 RATs, respectively.

Already in the first experiment, the method was able to give correct answers to 56% of the RATs in the training set and the accuracy for the test set RATs was 54%. By looking at the results we observed that many false solutions were very frequent words of English (also known as stopwords).

After simple stopword removal (we used the NLTK [14] stopword list) from the candidate set, the accuracy of the system for both sets increased to 66%. Now, many of the seemingly incorrect results were actually solved essentially correctly, but instead of the singular in the correct answer, the plural form of the answer word was proposed by the system. Such minor issues could be easily solved, but since our main interest is more in the principles that may help develop computational creativity, we did not delve into details.

An upper bound for the accuracy of the 2-gram-based technique for the training set is 96% and for the test set it is 99%. This is how often the candidate set included the correct answer word. Many of the remaining failed cases are due to compound words. For instance, for the RAT question with cue words *puss*, *tart* and *spoiled* the answer word *sour* is not detected because in everyday text 'sourpuss' is written together. Again, techniques to take this into account could be developed, but would not probably help finding truly creative associations.

Our results indicate that the method described above solves RAT questions more accurately than an average human. According to Bowden and Jung-Beeman [15], mean human accuracy for their 144 RAT questions is approximately 0.5, whereas the accuracy of our simple method is 0.66.

Overall the results indicate that the computational method based on 2-grams has already captured some principles of creativity, as measured by RATs.

IV. Generalized Approach to Support Creativity

The 2-gram model model used above is severely restricted and essentially only considers idiomatic phrases, such as compound words of exactly two elements. Obviously, many — if not most — relevant and informative associations between terms are manifested by less stringent proximity.

We next propose a more powerful, generalized approach to support creativity based on relations which are semantic in nature [16]. We are motivated by the observation that RATs are relatively easy for computers and that more general notions of relatedness of words or concepts could be used. Since RATs already correlate with creativity, a more general version could likely be used to support more challenging tasks of creativity.

In this section we describe a simple method for creating a network of semantically associated words. We experimentally test and illustrate how connections in this network tend to make sense. We also show how to apply the RAT solving principles to these networks in order to support some sorts of creative inference.

A. Word Co-Occurrence Network Construction

We briefly describe how a word co-occurrence network can be generated using existing text analysis methods. We assume a corpus of unstructured documents, and we treat documents as bags of sentences and sentences as bags of words. Formally, the document corpus C_W is a set of documents $d_i \in C_W$, where each document d_i is a (multi)set of sentences $d_i = \{s_{i1}, \ldots, s_{in}\}$, and each sentence is a set of words $s_{ij} \subset T_W$, where T_W is the set of all words.

We analyse word co-occurrences at the granularity of sentences, since words which are in one sentence have a strong relation to each other [17]. Valid alternative approaches could be based on a sliding window of words or a paragraph, for instance.

Formally, the word co-occurrence network $G = (V, E, W)$ is a weighted, undirected graph with nodes V, edges $E \subset V \times V$, and edge weights $W : V \times V \to \mathbb{R}_+$. For notational convenience, we assume $W(e_1, e_2) = 0$ if there is no edge between e_1 and e_2.

Before constructing the graph we preprocess the documents. First, we extract nouns and named entities from the documents and discard everything else. In addition to simplicity, this choice is motivated by nouns and named entities being conceptually more basic than concepts referred to by verbs or prepositions [18]. Obviously, some information is lost here. We then lower-case and lemmatize all the words. The named entities are concatenated with an underscore.

We use the log-likelihood ratio (LLR) to measure the strength of an association between two terms [12]. In the word co-occurrence network, lemmatized nouns and named entities are then nodes, and they are connected with an edge whenever the LLR is high enough (see below). The connections are also weighted by the LLRs.

B. Word Co-Occurrence Network of Wikipedia

In order to discover more general connections between words we chose to extract word co-occurrences from a text corpus. Google n-gram data sets are not used here since they only contain information about words which appear very close to each other.

In these experiments we construct the co-occurrence network from the English Wikipedia as of September 2011, consisting of 2,078,604 encyclopedic articles from all areas of life. For preprocessing the data we use Natural Language Processing Toolkit (NLTK) [14].

Without any pruning of edges, the co-occurrence network constructed from Wikipedia would consist of 1,900,846 nodes and 89,076,150 edges. Figure 2 shows the distribution of LLR values, i.e., the weight distribution of all possible edges before any pruning. As is to be expected, a majority of weights are small but there is a long tail to large weights.

Selecting a threshold value for LLR is a complicated task. Our reasoning was, that the minimum log-likelihood ratio value should be at least as high as it is for two terms which co-occur only twice and together. In our case the value $t = 70.44$ was used as the threshold value for the co-occurrence network. This removes approximately 95% of the edges from the network (cf. Figure 2). As a result, the network consists of 595,029 different terms and 4,644,456 edges.

C. Co-occurrence Network vs. WordNet Semantic Relations

To experimentally investigate what kind or semantic relations are discovered by the LLR-based method, we next experiment with WordNet [17]. It is a curated lexical database of English, with a large amount of manually assigned semantic relations of different types between words. WordNet is an accurate and powerful resource but limited in its scope. There are approximately 120,000 nouns in WordNet, when including example sentences and glossaries (see below). The co-occurrence network thus has around 470,000 nodes which do not appear in WordNet at all.

Our goal has not been to reproduce WordNet. Rather, we aim for a coverage much wider than WordNet (our 595k terms vs. WordNet's 120k terms), and also for language-independence so that the methods are applicable also in languages for which WordNet or similar resource do not exist. The sole purpose of these experiments is to shed light on the types of relationships discovered by LLR.

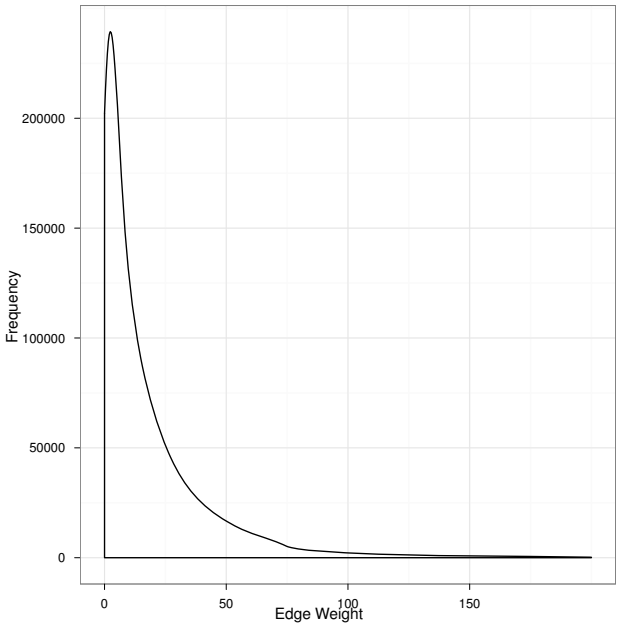

Fig. 2. Weight (LLR) distribution of the co-occurrence network before pruning.

Given two words w_1 and w_2, we consider their following possible relations in WordNet:

- w_1 is a *hypernym* of w_2, or vice versa (e.g. 'vehicle' is a hypernym of 'car').
- w_1 is a *holonym* of w_2, or vice versa (e.g., 'car' is a holonym of 'wheel').
- w_1 is a *holonymic sister* of w_2, i.e., they share a holonym (e.g., 'wheel' and 'door' both are parts of a car).
- w_1 and w_2 are *synonyms* (e.g., 'car' and 'automobile').
- w_1 and w_2 are *coordinate terms*, i.e., they share a hypernym (e.g., 'car' and 'ship' both are vehicles)
- w_1 appears in the *definition* of w_2, or vice versa (e.g., 'motor' appears in the WordNet definition of car: "a motor vehicle with four wheels; usually propelled by an internal combustion engine").
- w_1 appears in the *example sentences* of w_2, or vice versa (e.g., 'work' appears in the WordNet example use of the word car: "he needs a car to get to work").

More distant WordNet similarities could also be considered by transitively applying the above relations (for an overview see, e.g., [19]).

Because of the limited scope of WordNet, for our experiments concerning WordNet relations we randomly picked 5,000,000 pairs of words that do occur in WordNet. We excluded those words in our co-occurrence network that do not appear in WordNet, since obviously WordNet is not able to say anything about their relations.

Relation Type in WordNet	Number of Examples
Hypernym Relations	117
Holonym Relations	49
Holonymic Sister Relations	6
Synonym Relations	33
Coordinate Relation	2,729
Definition Relation	948
Example Relation	70
No Relation	4,996,048
Total Sample	5,000,000

TABLE I
DISTRIBUTION OF DIFFERENT WORDNET SEMANTIC RELATION TYPES IN A RANDOM DATASET.

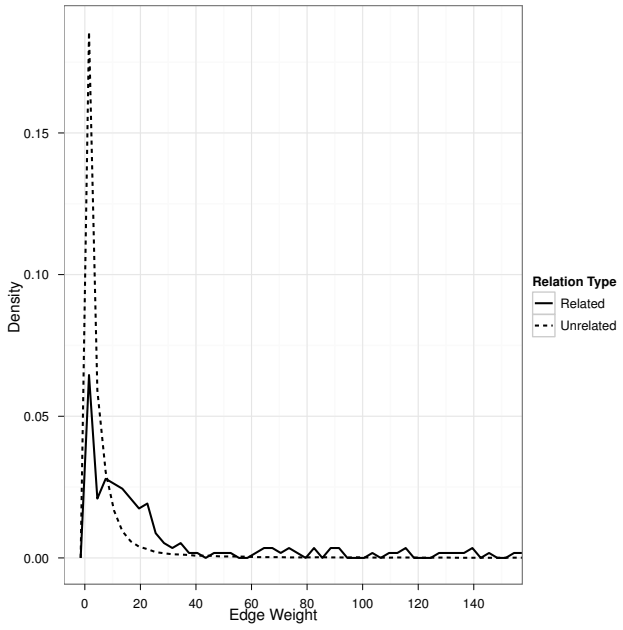

Fig. 3. Edge weight (LLR) distributions of edges which either are or are not related in WordNet.

The distribution of WordNet association types in the random sample of 5,000,000 pairs is shown in Table I. The number of words which are related in WordNet form a very small fraction of the dataset. Also, most term pairs in this random sample have low LLRs, essentially following the distribution of Figure 2.

Correlation between WordNet and LLRs is illustrated in Figure 3, where the edge weight distributions are drawn separately for those pairs that are related in WordNet and those that are not. Visually, the difference is clear: approximately already from edge weight 15 on, related word pairs have a higher density than unrelated pairs.

Since so few pairs are related in WordNet, we also look at the data using ROC (Receiver Operating Characteristic) curve which is suited for unbalanced class distributions. The curve can be seen in Figure 4 (zoomed in to the lower left corner).

The true positive rates grow in the beginning very fast (note

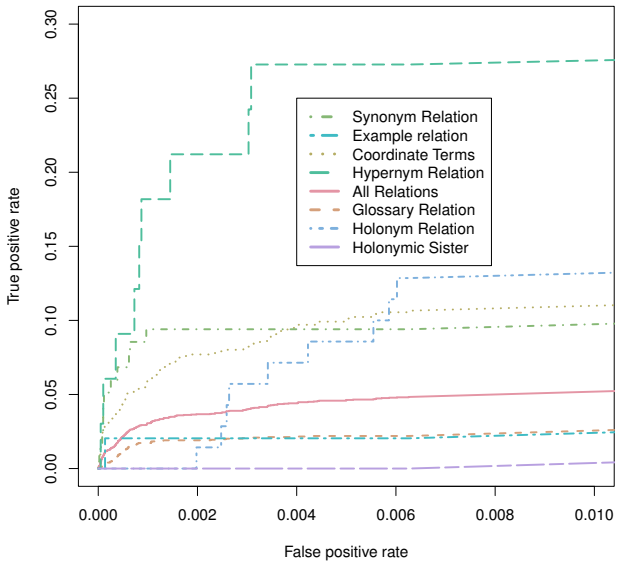

Fig. 4. Zoom-in to the lower left corner of the ROC plot.

the difference in x and y scales in the figure), but then they level off to a straight line towards point $(1, 1)$. This indicates that the top ranking term pairs are typically WordNet related, as suggested also by Figure 3, but after that there is no visible difference.

These experiments show that the relations discovered by LLR tend to make sense semantically. The sheer numbers additionally show that the co-occurrence method has a much higher coverage than WordNet (but obviously WordNet has strengths, such as semantic categories of relationships and manually curated contents). We believe that word co-occurrence based models on which we can build creativity support methods could be much more interesting than the 2-gram models for solving RATs.

V. CREATIVE ASSOCIATION DISCOVERY

We now proposed initial methods for finding more general creative associations. First we will propose a generalized version of the method proposed for RATs in Section III. Note, however, that now the goal is not to solve RATs, they are just used to ensure that the responses of the proposed algorithm are sound.

In the final subsection we will actually propose a method for generating generalized RATs, and we will show that the generation method is quite stable. We will also provide examples of the creative inference to the reader.

A. Generalization of RAT-Related Methods

a) Candidate word selection: The generalization of the candidate method from the previously presented method is quite straightforward. In the method which used 2-grams as the

model of co-occurrences the words which co-occur with every cue word were used as candidate answer words. Choosing the candidate set can be done in a similar way for the co-occurrence network by choosing the joint neighbourhood of all the cue words.

More formally, let us consider a set $T = \{t_1, \ldots, t_n\}$ of words which we treat as cue words. We will define the joint neighbourhood as the intersection of all the neighbours of the cue words:

$$\mathcal{N}(T) = \{u \mid \{t_i, u\} \in E \text{ for all } t_i \in T\}. \quad (7)$$

b) Scoring: For ranking the candidates, consider first a single candidate word $a \in \mathcal{N}(T)$. We propose using a score which depends on two aspects of the candidate word a. First, a good answer word a should be strongly related to all of the cue words t_i. Second, a good answer word is specific to the cue words, i.e., does not associate strongly with too many other words. The second criterion also relates to the fact that high-frequency candidates are not considered as creative [3].

We define the scoring function as

$$score(a, T) = \alpha(a, T) \cdot \beta(a), \quad (8)$$

where $\alpha(a, T)$ is the association weight-induced component of the score and $\beta(a)$ is the candidate frequency-induced component of the score.

Some reasonable scores which could be calculated as the α component are the following:

1) The minimum weight (MINW) between the answer word and the cue words, i.e., "the weakest link":
$$\alpha(a, T) = \min_{t_i \in T}(W(a, t_i)).$$
2) The average edge weight (AVGW) between the cue words and the answer word
$$\alpha(a, T) = \frac{1}{|T|} \sum_{t_i \in T} W(a, t_i).$$
3) As the edge weights are ratios, it is also reasonable to consider the harmonic mean (HARM)
$$\alpha(a, T) = \frac{|T|}{\sum_{t_i \in T} \frac{1}{\max(W(a, t_i), 1)}}.$$

Analogously there are different ways to penalize the answer word frequency. In this paper we consider the two most obvious approaches related to the degree of the candidate node a. The first approach penalizes a score by dividing it by the candidate node degree (DEG), i.e.,

$$\beta(a) = \frac{1}{\deg(a)}.$$

A logarithmic smoothing of the degree penalty (DEGL) component might give more stable results:

$$\beta(a) = \frac{1}{\log(\deg(a))}.$$

B. Generalized RAT Creation

In standard RAT questions the goal is to provide an answer word given the cue words. While this measures creative abilities, often the opposite task has more practical value: we have a concept (the answer word, e.g., the topic of a problem we want to solve), and we want to have it associated creatively with other concepts. For instance, let's assume we are interested in the word 'riding' and, to support our creativity, would like to see it associated with different things. The method that we will give below recommends these words: 'election', 'horseback' and 'accident'.

In this task, given an answer word, our goal is to select words that are strongly related to the answer word and at the same time are not related to each other. We propose this simple algorithm for selecting such words given the answer word a: First, choose the node with the strongest connection to a and add it to the (so far empty) cue word set R. Then, consider other nodes in a decreasing order of their association with the answer word a. Add a node to the cue word set R if and only if it is not connected to any member of R. Iterate until the desired number of cue words has been chosen or all neighbours of the answer word have been considered.

C. Experiments

Our first experimental goal is to test how well different scoring functions work on RAT questions. We will conduct these experiments on the training set. Once we have chosen the best method we will validate it using the separate test set.

Recall that the documents were preprocessed to support discovery of non-trivial associations between concepts. This preprocessing, i.e. including only named entities and nouns in the network, actually hinders solving the RATs. Therefore, we compare different scoring functions using those RAT questions where the candidate answer set (the joint neighbourhood of the cue words) contains the correct answer word. 21% of the test cases fell in this category. The relatively low score is explained by preprocessing aspect which we described earlier (i.e. many common entities are treated as one, e.g. 'political', 'party' is treated as 'political_party' in the co-occurrence network).

Results are shown in Table II (for acronyms used in the table, see the previous subsection). For $\alpha(a, T)$, the association weight-dependent component of the score, the harmonic mean (HARM) systematically produced best results. For $\beta(a)$, the candidate frequency-dependent component, the best results were obtained when dividing the score by the number of associations, i.e., the degree of node a in the co-occurrence graph (DEG). Overall, their combination also gave the best result.

To test the stability of the score, we then conducted the same experiment on the test data. The test set size shrinks to only 10 questions after taking the joint neighbourhood, so the statistical power is not high. However, the obtained accuracy of 0.8 indicates that there was no serious overfitting to the training set. In the next experiments we will thus use the combination of the harmonic mean and degree penalty.

TABLE II
COMPARISON OF THE ACCURACY OF DIFFERENT COMBINATIONS OF SCORING METHODS FOR CANDIDATE WORDS.

	$\alpha(a, T)$		
$\beta(a)$	**MINW**	**AVGW**	**HARM**
Constant	0.72	0.72	0.76
DEG	0.86	0.86	**0.90**
DEGL	0.76	0.76	0.83

TABLE III
A SAMPLE OF ARTIFICIALLY GENERATED GENERALIZED RAT QUESTIONS.

Seed Word	**Cue Word 1**	**Cue Word 2**	**Cue Word 3**
imperialism	colonialism	lenin	american
missile	warhead	defense	flight
packaging	product	paper	artwork
slope	steep	ski	western
medley	relay	yankovic	beatles
far	north	greater	moon
kpmg	firm	young	report
concert	band	hall	benefit

We next analyse the generalized RAT creation process, as an approximation of a creative discovery task. To test the sanity of this method we conducted the following experiment. We chose 1000 random words which each had at least 3 mutually unconnected neighbours in the co-occurrence graph. For each such random word we selected 3 cue words by using the RAT creation process described above. We then solved the RAT question given the 3 cue words, and compared if the answer thus obtained was identical to the original seed word. In 97% of the cases the results were same for both methods, indicating consistency of the methodologies.

Finally, a sample of such artificially created generalized RAT questions is shown in Table III. Subjectively judging, they seem to match quite well classical criteria of creativity, such as the Torrance Tests of Creative Thinking [20]. The RAT creation method could be considered to exhibit *fluency* by producing a number of relevant cue words (and more could be easily generated), *flexibility* by discovering cue words that provide complementary contexts or meanings for the seed word, as well as *originality* by providing relatively rare words. Additionally, *elaboration* could potentially be achieved by using the co-occurrence network to describe the contexts for the various associations.

VI. RELATED WORK

A. Measuring Associations Between Terms

The idea of the distributional hypothesis is that words which co-occur in similar contexts tend to have similar meanings [21]. This was nicely put by Firth in 1957: "You shall know a word by the company it keeps" [22]. Followed by these ideas, the semantic similarity between words is calculated by their co-occurrence in documents.

Even if relatively few methods have been proposed for automatic construction of networks of terms, literature on co-occurrence or collocation statistics is abundant. Such measures can be used in an obvious way to build a network of terms. We only review some representative methods here.

Log-likelihood ratio is a non-parametric statistical test for co-occurrence analysis. Using log-likelihood ratio for word co-occurrence analysis was proposed by Dunning [12] who showed, in particular, that log-likelihood ratio does not overestimate the importance of very frequent words like some other measures.

Latent Semantic Analysis [23] aims to find a set of concepts (instead of terms) in a corpus using singular value decomposition. The semantic similarity (relatedness) of two words can then be estimated by comparing them in the concept space. Latent semantic analysis has then evolved to *Probabilistic Latent Semantic Analysis* [24] and later to *Latent Dirichlet Allocation* [25]. Probably any of these methods could be used to derive co-occurrence networks.

B. Creative Association Discovery

Several papers have been published on supporting creativity by discovering links between concepts. In creative biological problem solving, for instance, Mozetic et al. [26] propose a method for finding unexpected links between concepts from different contexts. Examples of methods more directly based on link prediction in heterogeneous networks are given by Eronen and Toivonen [27].

VII. DISCUSSION

Making the 'right' choices is often much easier than making choices which are less rational, but do still make sense. This is what this paper is all about – given constraints, our goal is to propose something as a result which satisfies these constraints, but at the same time is thought-provoking. In creative support systems, one of the purposes is to encourage the user to think more broadly. One way for doing this is by giving answers, which are related to the question, but the relation itself is subtle enough, to induce creative thoughts.

In the paper we briefly described RATs and their underlying mechanisms. We showed that by using 2-grams and a simple probabilistic model it is possible to solve these tests with a good accuracy.

We also described a methodology for creating a network of more general associations than the 2-gram language model could provide. As a ground for the creative inference, we showed that the connections in this network tend to make sense and we can assume that if two words are connected by an edge, they are also semantically related.

Our main contribution is translating the principles which we established in the probabilistic framework for solving RATs to the generalized model with co-occurrence networks. An empirical result was that the associations generated from the network seem to exhibit creativity.

In the future our goal is to validate the methods more objectively, e.g., by some user testing. We plan to test and compare different language models (e.g., LSI, LDA) and provide more in depth analysis for the creative association discovery. Finally, we are planning to use these methods in tasks which relate to lexical creativity (e.g., automatic poetry generation) and in possible lexical creativity support systems (e.g., slogan wizard).

Answer to the RAT Question in the Introduction

The intended answer word related to 'coin', 'quick', and 'spoon' is 'silver'.

Acknowledgements: This work has been supported by the Algorithmic Data Analysis (Algodan) Centre of Excellence of the Academy of Finland.

REFERENCES

[1] S. Mednick, "The associative basis of the creative process." *Psychological review*, vol. 69, no. 3, p. 220, 1962.

[2] J. M. Toivanen, H. Toivonen, A. Valitutti, and O. Gross, "Corpus-based generation of content and form in poetry," in *International Conference on Computational Creativity*, Dublin, Ireland, 2012, pp. 175–179.

[3] N. Gupta, Y. Jang, S. Mednick, and D. Huber, "The road not taken creative solutions require avoidance of high-frequency responses," *Psychological Science*, 2012.

[4] M. T. Mednick and F. M. Andrews, "Creative thinking and level of intelligence," *Journal of Creative Behavior*, vol. 1, pp. 428–431, 1967.

[5] G. Forbach and R. Evans, "The remote associates test as a predictor of productivity in brainstorming groups," *Applied Psychological Measurement*, vol. 5, no. 3, pp. 333–339, 1981.

[6] K. S. Bowers, G. Regehr, C. Balthazard, and K. Parker, "Intuition in the context of discovery," *Cognitive Psychology*, vol. 22, pp. 72–110, 1990.

[7] S. Topolinski and F. Strack, "Where theres a willtheres no intuition: The unintentional basis of semantic coherence judgments," *Journal of Memory and Language*, vol. 58, pp. 1032–1048, 2008.

[8] E. Vul and H. Pashler, "Incubation benefits only after people have been misdirected," *Memory & Cognition*, vol. 35, pp. 701–710, 2007.

[9] K. Bowers, G. Regehr, C. Balthazard, and K. Parker, "Intuition in the context of discovery," *Cognitive psychology*, vol. 22, no. 1, pp. 72–110, 1990.

[10] S. Mednick and M. Mednick, *Examiner's Manual, Remote Associates Test: College and Adult Forms 1 and 2*. Houghton Mifflin, 1967.

[11] J.-B. Michel, Y. K. Shen, A. P. Aiden, A. Veres, M. K. Gray, T. G. B. Team, J. P. Pickett, D. Hoiberg, D. Clancy, P. Norvig, J. Orwant, S. Pinker, M. A. Nowak, and E. L. Aiden, "Quantitative analysis of culture using millions of digitized books," *Science*, vol. 331, no. 6014, pp. 176–182, 2011.

[12] T. Dunning, "Accurate methods for the statistics of surprise and coincidence," *Computational linguistics*, vol. 19, no. 1, pp. 61–74, 1993.

[13] P. Domingos and M. Pazzani, "On the optimality of the simple bayesian classifier under zero-one loss," *Machine learning*, vol. 29, no. 2, pp. 103–130, 1997.

[14] S. Bird, E. Klein, and E. Loper, *Natural Language Processing with Python*. O'Reilly Media, 2009.

[15] E. Bowden and M. Jung-Beeman, "Normative data for 144 compound remote associate problems," *Behavior Research Methods*, vol. 35, no. 4, pp. 634–639, 2003.

[16] D. Dailey, "An analysis and evaluation of the internal validity of the remote associates test: What does it measure?" *Educational and Psychological Measurement*, vol. 38, no. 4, pp. 1031–1040, 1978.

[17] G. A. Miller, "Wordnet: a lexical database for english," *Communications of the ACM*, vol. 38, no. 11, pp. 39–41, 1995.

[18] D. Gentner, "Why Nouns Are Learned Before Verbs: Linguistic Relativity Vs. Natural Partitioning," in *Language Development, vol.2: Language, cognition and culture*, S. Kuczaj, Ed. Hillsdale, NJ: Erlbaum, 1982, pp. 301–334.

[19] A. Budanitsky and G. Hirst, "Semantic distance in wordnet: An experimental, application-oriented evaluation of five measures," in *Workshop on WordNet and Other Lexical Resources*, vol. 2, 2001.

[20] E. Torrance, *Torrance Tests of Creative Thinking: Norms-technical Manual. Research Edition. Verbal Tests, Forms A and B. Figural Tests, Forms A and B.* Personnel Press, 1966.

[21] Z. Harris, "Distributional structure," *Word*, vol. 10, no. 23, pp. 146–162, 1954.

[22] J. R. Firth, "A synopsis of linguistic theory 1930-55." vol. 1952-59, pp. 1–32, 1957.

[23] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman, "Indexing by latent semantic analysis," *Journal of the American Society for Information Science*, vol. 41, no. 6, pp. 391–407, 1990.

[24] T. Hofmann, "Probabilistic latent semantic indexing," in *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. New York, NY, USA: ACM, 1999, pp. 50–57.

[25] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.

[26] I. Mozetic, N. Lavrac, V. Podpecan, P. K. Novak, H. Motaln, M. Petek, K. Gruden, H. Toivonen, and K. Kulovesi, "Bisociative knowledge discovery for microarray data analysis," in *The 1st International Conference on Computational Creativity (ICCC-X)*, Lisbon, Portugal, 2010, pp. 190–199.

[27] L. Eronen and H. Toivonen, "Biomine: Predicting links between biological entities using network models of heterogeneous database," *BMC Bioinformatics*, vol. 13, no. 119, 2012.