ANSWER: An Unsupervised Attractor Network Method for Detecting Salient Words in Text Corpora

Madhavun Candadai, Aashay Vanarase, Mei Mei and Ali A. Minai, *Senior Member, IEEE*

Abstract— **The availability of unstructured text as a source of data has increased by orders of magnitude in the last few years, triggering extensive research in the automated processing and analysis of electronic texts. An especially important and difficult problem is the identification of salient words in a corpus, so that further processing can focus on these words without distraction by uninformative words. Standard lists of *stop words* are used to remove common words such as articles, pronouns and prepositions, but many other words that should be removed are much harder to identify because word salience is highly context-dependent. In this paper, we describe a neurodynamical approach for the context-dependent identification of salient words in large text corpora. The method, termed the *Attractor Network-based Salient Word Extraction Rule* (ANSWER) is modeled as a cognitive mechanism that identifies salient words based on their participation in coherent multi-word ideas. These ideas are, in turn, extracted via attractor dynamics in a recurrent neural network modeling the associative semantic graph of the corpus. The corpus used in this paper comprises the abstracts of all papers published in the proceedings of IJCNN 2009, 2011 and 2013. The list of salient words that the system generates is compared with those generated by other standard metrics, and is found to outperform all of them in almost all cases.**

I. Introduction

Analyzing large corpora of text is important for many applications such as information retrieval, automatic text interpretation, document classification, etc. With the continuing exponential growth of on-line texts, analysis must be automatic and efficient, which requires that algorithms focus on the most relevant information. Thus, one of the first steps in text analysis is often the removal of words that are considered non-salient for the task at hand. These include a standard set of *stop words* – articles, pronouns, prepositions and conjunctions, as well as modal verbs and other very common verbs – but even after their removal, corpora typically contain a mixture of salient and non-salient words. This has motivated the development of several approaches for identifying salient words using features such as word frequency and pattern of word distribution over the corpus [1], [2], [3], [4], [5], [6], [7], [8]. However, all of these methods have limitations, and new approaches are still needed for this important task. In particular, most existing approaches for identifying salient words are focused mainly on discovering keywords or terms for search queries. i.e., extracting small, highly concentrated sets of words, rather than finding all (or a large fraction) of salient words in the corpus. The latter is important for applications such as concept analysis and idea mining in text corpora.

In this paper, we describe an efficient method for extracting salient words from text corpora using a recurrent neural network. The method is completely data-driven, and is based on cognitive and relational arguments rather than purely statistical ones. Very importantly, the method does *not* require that the corpus being analyzed divide neatly into distinct documents, as is the case with the most widely used approach – the *term frequency-inverse document frequency* (TF-IDF) method. We demonstrate the capabilities of the proposed technique using a corpus comprising the abstracts of all papers in the 2009, 2011 and 2013 International Joint Conferences on Neural Networks (IJCNN'2009, IJCNN'2011 and IJCNN'2013). The performance of the method is compared with that of several other commonly used heuristics, and is found to exceed them in almost all cases.

II. Previous Work

The earliest automated approach to identifying salient words in documents and corpora was proposed by Luhn, and was based on word frequency [1]. In this approach, words with very high and very low frequencies in the corpus were considered non-salient and those with frequencies between these cut-offs were designated salient. These salient words were then used to identify significant sentences to automatically assemble an abstract. This approach is useful, but word frequency is, at best, an uncertain indicator of salience. For corpora with well-defined documents, the more focused TF-IDF metric was proposed by Salton and colleagues [2], [3]. It calculates the saliency of word v in a document d as:

$$s(v|d) = TF(v,d) \times IDF(v) = \frac{n(v,d)}{N_d} \times log\frac{M}{m(v)} \quad (1)$$

where $n(v,d)$ is the number of times v occurs in d, N_d is the number of words in d, M is the total number of documents in the corpus, and $m(v)$ is the number of documents that contain the word v. Thus, a word that occurs frequently in d but does not occur in most documents is considered salient for that document. To calculate corpus-wide saliency, one can either average the saliency across all documents, take the maximum saliency, or count the fraction of documents in which the word is sufficiently salient. This approach is used very frequently in information retrieval applications. Other

Madhavun Candadai (Email: candadmv@mail.uc.edu), Aashay Vanarase (Email: vanaraak@mail.uc.edu), Mei Mei (Email: meimi@mail.uc.edu) and Ali Minai (Email: Ali.Minai@uc.edu) are all with the Department of Electrical Engineering and Computing Systems, University of Cincinnati, Cincinnati OH 45221

Acknowledgement: This work was supported in part by a National Science Foundation CreativeIT grant to Ali Minai (IIS-0855714) and a National Science Foundation INSPIRE grant to Ali Minai (BCS-1247971).

978-1-4799-1959-8/15/$31.00 ©2015 IEEE

approaches proposed for identifying salient words include: keyword detection based on analyzing the recurrence interval statistics of words with respect to a stochastic process model [4], [5], [7]; computing the inhomogeneity of a word's local versus global density based on the premise that salient words have a "bursty" distribution, occuring with greater density in local neighbourhoods where they are more relevant rather than uniformly over the corpus [6], [7]; using the statistics of a word's distribution relative to its neighbors in the corpus [9]; and extracting highly connected word communities from the semantic graphs of individual documents [8].

The ANSWER model is based on a network approach, viewing documents and corpora as *semantic graphs* with nodes as words and associations between them as edges. Such models have been used previously in text processing applications such as document summarization [10], [11], visualization [12], question answering [13], content extraction [14], keyword extraction [8], and document clustering and retrieval [15]. However, the meaning of edges has often varied across these applications. In ANSWER, the associations represent co-occurrence within the same semantic unit, i.e., sentence.

Broadly, the ANSWER model falls into the category of neural models based on associative learning and recall. It has long been recognized that associations between semantic elements, e.g., concepts, categories, etc., is fundamental to many cognitive processes – especially those involving language and ideas. This has led to extensive studies on word association norms [16], [17], [18], associative recall [19], [20], [21], [22], [23], priming [24], [25], [26], [27], [28], ideation [29], [30], [31], [32], [33], [34], [35], [36], [37] and the associative structure of language [29], [38], [39], [40], [41], [42], [43], [44]. The ANSWER model draws upon this work for its theoretical underpinnings. It is based on a neurodynamical model of thinking called *IDEA* (*itinerant dynamics with emergent attractors*) that we have described previously [45], [46], [47], [48], and our earlier work on computational models of ideation and priming [49], [50], [51], [52].

III. Description of Overall Approach

Several of the existing methods for identifying salient words are based on a broad but fundamental heuristic: The occurrence of salient words in texts or relative to other words is "informative" or "inhomogeneous", while that of non-salient words tends to be more "random". This reflects the idea that salient words occur meaningfully, whereas non-salient ones, such as articles or prepositions, occur for reasons of grammar and syntax rather than semantics. The ANSWER model presented in this paper embodies a more cognitively motivated version of this insight. It is based on three basic postulates:

- Semantic knowledge derived from a text corpus is represented in the mind by an *Associative Semantic Network* (ASN) whose nodes are words (or concepts) and where the edges between nodes represent associations between pairs of words as found in the corpus, e.g., the correlation, point-wise mutual information, or probability of co-occurrence in the same sentence. The construction of such an ASN can be seen as the result of reading the texts in the corpus.
- Meaningful *ideas* induced from the corpus correspond to strongly connected groups of nodes (words/concepts) within the ASN.
- Salient words are over-represented in coherent ideas compared to their representation in the corpus at large, whereas non-salient words are represented at baseline or lower levels, i.e. *ideas concentrate salience.* Thus, the frequency of a word in a large list of coherent ideas derived from a corpus is a better indicator of its salience than its frequency in the entire corpus.

Thus, the approach of the ANSWER model requires three steps: 1) Construction of the ASN from the corpus; 2) Sampling of a large set of coherent ideas from the ASN as strongly connected sub-networks of a few nodes each; and 3) Thresholding words into salient and non-salient groups based on their frequency of occurrence in the sampled ideas. This can be seen as a graph-based analog of the approach in [6], [7], where localized regions of text are identified as being more or less dense in salient words.

The construction of the ASN requires two decisions: Which words from the corpus to include, and how to assign the value of associative weights between words. As described below, we exclude many obviously non-salient words from the ASN using simple pre-processing so that the ASN consists mainly of words whose salience is difficult to ascertain. The associative weights in this work are based on correlated co-occurrence of words in sentences (see below).

The core of the algorithm is the sampling of ideas from the ASN. Unlike methods that identify salient regions in the text [6], [7], ideas sampled from the ASN include *emergent ideas*, i.e., those that never occurred explicitly in the text but are "implied" by its correlational statistics. Elsewhere, we have suggested that these might represent latent ideas that could be the basis of subsequent innovation [53]. In ANSWER, the sampling of ideas from the ASN occurs using the attractor network approach derived from our previously described IDEA model [45], [46], [47], [48]. In that model, the ASN is constructed from the text corpus as a recurrent neural network with neural units representing words and weights encoding associations between them. The weights between pairs of words are computed based on their co-occurrence at the sentence level. In the IDEA model, competitive K-of-n activity in the network generates an itinerant attractor dynamics corresponding to a chain of thoughts. In the ANSWER model, the metastable itinerant dynamics is replaced by repeated random cuing of the network as discussed in the next section. Such cued recall based on word associations has been studied widely in the field of cognitive science [19], [20], [21], [22], [23]. A cue in the form of a (possibly incoherent) group of words triggers activity in the nervous system that eventually converges to a state representing a coherent lexical combination. The ANSWER model captures this dynamics, generating a repertoire of ideas when sampled with random cues. Salient words are then identified from this repertoire.

Key Features: The approach embodied in ANSWER has three features that make it especially useful and broadly applicable:

1) It is an unsupervised method that does not require the existence of a labeled dataset for training. Producing such datasets is extremely laborious and subjective, and the labels often do not generalize across texts from different domains.
2) It does not require that the corpus consist of distinct documents, or that the documents be known a priori. Methods based on TF-IDF do require this, which limits their applicability.
3) It is based on an intuitive model of thinking as an associative neurodynamical cognitive process. As such, results from experimental studies and computational models of several cognitive processes can be applied to improve the method further, and conversely, the method can be used to explore hypotheses about these cognitive processes.

A key challenge in building a model for identifying salient words (or keywords) is the absence of training data. Some models have assumed that hand-labeled data exists for this, but this can be very difficult to generate and may not be valid across multiple corpora. The ANSWER approach is unsupervised and data-driven, using the data in the corpus to set up the ASN as an attractor neural network. However, labeled data is necessary to assess the viability of this approach. To this end, we labeled the set of words used in the ASN as salient or non-salient by inspection, and used this as the basis of analyzing the performance of ANSWER and other heuristics.

The next two sections describe how the corpus is preprocessed to set up the ASN and the attractor network model for sampling ideas.

IV. Data Extraction and Preparation

A. The Corpus

The text corpus, C, used in this paper comprises all abstracts from the Proceedings of the 2009, 2011 and 2013 IJCNN meetings. In all, 1,410 papers were processed to extract just the content of the abstracts, excluding title, authors, affiliations, etc. This was followed by pre-processing as follows:

1) A Porter stemmer that is available at *http://www.cs.cmu.edu/∼callan/Teaching/porter.c* was used to stem the words. Words in the set that stemmed to the same root were replaced by a single reference word from the set to make all words in the final dataset recognizable (e.g., “use”, “user”, “using”, “uses” were all replaced by “using”).
2) Standard stop words were removed using the list at: *http://norm.al/2009/04/14/list-of-english-stop-words/*.
3) A heuristic algorithm described in [48] was used to remove further non-salient words. A *relative prominence* value, $R(v_i)$, was calculated for each word, v_i, using two quantities: $f_{ELP}(v_i)$, its frequency (in occurrences per million words) in the 40,481-word *English Lexicon Project* (ELP) corpus (elexicon.wustl.edu); and $f_C(v_i)$, its frequency in the IJCNN corpus, C. The relative prominence was given by:

$$R(v_i) = log\frac{f_C(v_i)}{f_{ELP}(v_i)} \qquad (2)$$

All words v_i with $R(v_i) < 0.001$ were removed as non-salient.
4) Every sentence was represented as a set of word tokens without repetition, i.e., each unique word in the sentence received only one token regardless of how many times it occurred in the sentence.

The preprocessing described above resulted in the removal of six abstracts, leaving 1,404 documents for subsequent analysis. Finally, in order to reduce the size of the data and the strong but unwarranted effect of the low-frequency words, we eliminated words that occurred fewer than 4 times (i.e., had fewer than 4 word tokens) in the corpus.

The final processed corpus, C_p, had N_S =12,011 sentences, N_W = 99,169 word tokens, and N_V = 2,309 unique words. The vocabulary of unique words is denoted by $V = \{v_i\}$.

V. The Network Model

A. Model Structure

The neural network is a one-layer recurrent network of $n = N_V =$ 2,309 neural units, and has competitive K-of-n dynamics. Each neural unit in the network represents a word in the vocabulary, and the weights between units are set based on the associations between words computed from the corpus. The corpus C_p is processed to calculate the *occurrence probability*, p_i, of each word, v_i, given by the fraction of sentences that include the word, and the *co-occurrence probabilities*, p_{ij}, for every pair of words, v_i and v_j, given by the fraction of sentences that include both words. Finally, the weight between unit i and unit j is set to the correlation coefficient of words v_i and v_j calculated as:

$$a_{ij} = \frac{p_{ij} - p_i p_j}{\sqrt{p_i(1-p_i)}\sqrt{p_j(1-p_j)}} \qquad (3)$$

Weight values less than 0 are set to 0, as are the self-weights for all units. The resulting weights are thus symmetric and range between 0 and 1. The exclusion of negative weights is justified heuristically by the fact that almost all of these are small and provide little specific information beyond the global lateral inhibition already assumed in the network’s competitive K-of-n dynamics.

B. Model Dynamics

The input received by unit i at time t is given by:

$$x_i(t) = \sum_{j=1}^{n} w_{ij}(t) z_j(t) + \gamma_{noise} \xi_i(t) \tag{4}$$

where z_j is the output of unit j, γ_{noise}, is a gain parameter, and $\xi_i(t)$ is uniform white noise. The state of unit i is updated at time t using:

$$y_i(t) = \alpha y_i(t-1) + (1-\alpha) x_i(t) \tag{5}$$

The value of α is set just below 1 to simulate continuous-time dynamics. At step t, the K most highly activated non-refractory units fire provided they have $y_i(t) > y_{min}$, where $y_{min} > 0$ is a small positive base value. The output of unit i is:

$$z_i(t) = f(y_i(t)) = \begin{cases} 1, & \text{if } i \in \{K \text{ most excited units}\} \\ 0, & \text{otherwise} \end{cases} \tag{6}$$

We have found that using a rigid K-of-n rule can lead to diminished performance, and the system uses a soft version where any unit, j, with $y_j(t)$ within 2% of the nominal K-of-n threshold also fires. Thus, while we choose a particular value of K for all simulations, the actual number of active neurons at any given time may be larger.

Cueing the network with a random initial state triggers activity that settles down to a fixed-point attractor due to the symmetric weights [54], [55]. Though the network has no inhibitory weights, the soft K-of-n effectively induces global lateral inhibition. This ensures that the neurons active in the final attractor represent a mutually coherent group of words, which corresponds to our definition of an *idea.* Repeatedly cuing the network with random initial states allows the generation of a large number of such coherent ideas as attractors latent in the network. It should be noted that, unlike classical, threshold-based models such as Hopfield networks [54], the competitive dynamics in our model ensures the existence of a large number of sparse attractors, which are retrieved via random stimulation. In the current implementation, all attractors are binary vectors.

With a sufficiently large number of random cues, the system yields a representative sample of the ideas implicit in its structure. The frequency with which each word occurs in this set is then calculated, and words occurring with frequencies above a threshold are deemed salient.

VI. Simulations, Results and Discussion

A. Experimental Setup

To study the potential benefits of the ANSWER approach, we tried the system with three values of K, i.e., the nominal desired number of active neurons in each attractor. These values were $K = \{5, 7, 10\}$. The values were chosen partly in view of the fact that the average sentence length in the corpus, after preprocessing, is 8.25 words, so the generated ideas in the ANSWER trials ranged in sizes around this value and can be seen intuitively as representing a sentence (minus stop words, etc.) For each K value, seven independent simulations were run, each with 5,000 random cues defined by initially setting K random neurons to be active before the first step. The "softness" of the threshold was set at 98%, the continuous-time dynamics parameter α at 0.9 and γ_{noise} at 0.1.

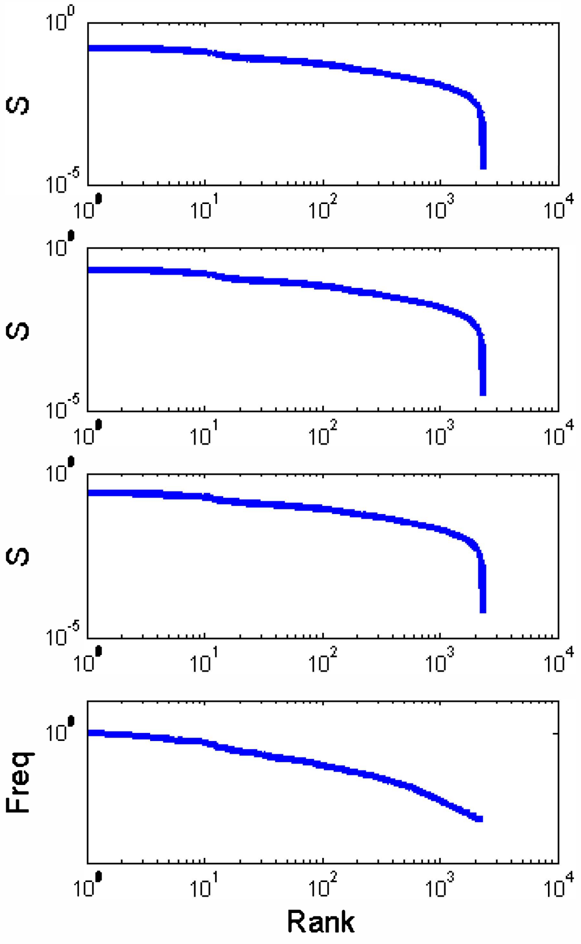

Fig. 1. The top three graphs show log-log rank-value plots of the saliency scores for the $K = 5, 7$ and 10 cases averaged over 7 trials (top to bottom). The bottom plot is the log-log rank-value plot of the normalized per-sentence word frequencies for the corpus.

A *saliency score* (S) was computed for every word as

$$S(v_i) = \frac{\sum_{k=1}^{N_c} z_i^k}{N_c} \tag{7}$$

where N_c is the total number of cues and z_i^k denotes the i^{th} bit of the binary attractor vector generated by the k^{th} cue. Thus, $S(v_i)$ indicates the fraction of the N_c attractors in which word v_i makes an appearance, i.e., its relative frequency over the attractors. ANSWER's list of salient words is generated by thresholding the words by the saliency

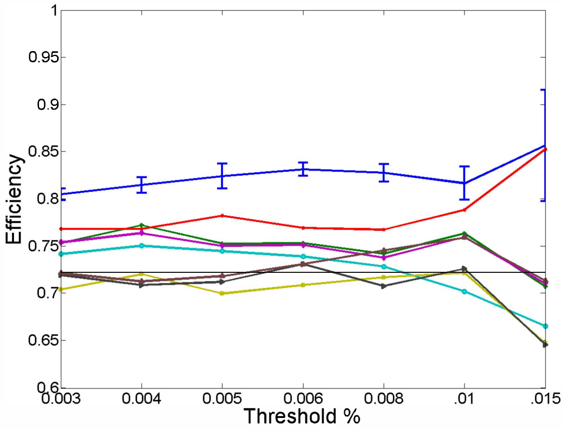

Fig. 2. The efficiency of ANSWER and other heuristics for the $K = 5$ case at different thresholds. The threshold is expressed as a fraction of $5,000$, which is the total number of attractors sampled. Thus, the words identified as salient at a threshold of $x\%$ occurred in at least $5000 \times x/100$ of the attractors. The curves are: Blue with error bars: ANSWER; Green with vertical line marker: Frequency; Magenta with diamond marker: Mean TF-IDF; Yellow with star marker: Max TF-IDF; Red with no marker: Degree; Cyan with circle marker: Weights; Brown with m̂arker: Betweeness Centrality; Black with ¿ marker: Eigenvector Centrality. The straight horizontal line at 0.72 is the performance of a uniform random selection from the word list.

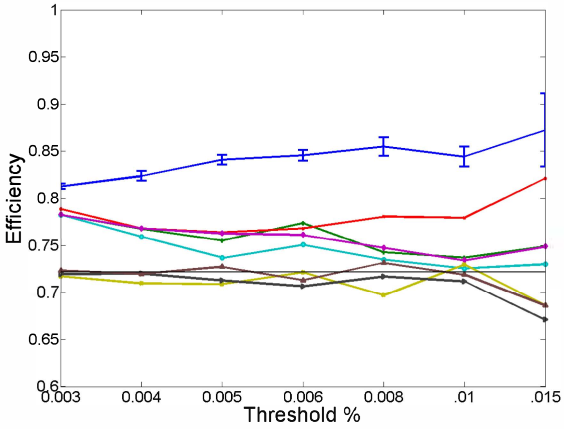

Fig. 3. The efficiency of ANSWER and other heuristics for the $K = 7$ case at different thresholds. All other parameters and plotting conventions are as in Figure 2

scores and regarding words with saliency scores greater than or equal to the threshold as salient. Thus, choosing a higher threshold produces a smaller list of salient words.

Figure 1 shows the rank-value plot of the saliency scores of all the words for the $K = 5$, 7 and 10 cases, and the rank-value plot for the normalized corpus frequency of words. The log-log plots for saliency scores (top three plots) show three regions: a) A few words (3 or 4) with almost equal saliency scores just below 1; b) A broad range of about 1,600 words with saliency scores declining as a power law of rank; and c) A sharp cutoff for words rarely or never seen in the attractors. By comparison, the log-log plot of the normalized corpus frequency shows a steady power law decrease with rank, with a gentle cutoff or tapering for rare words. We expect that using an N_c much greater than 5,000 will reduce the abruptness in the saliency score cutoff, but the turning point gives some indication of where a good threshold may lie. However, in this paper we consider several possible thresholds to study whether the saliency decisions made by ANSWER are fairly robust.

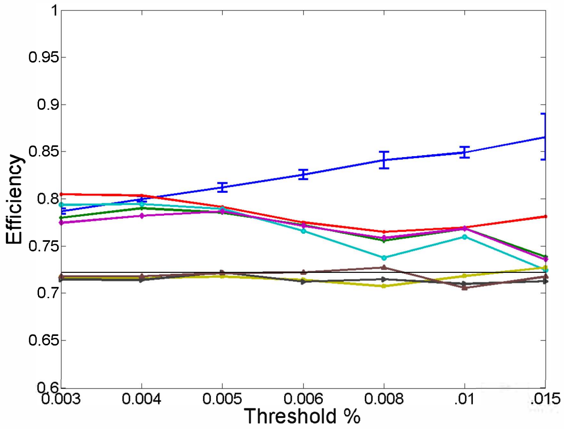

Fig. 4. The efficiency of ANSWER and other heuristics for the $K = 10$ case at different thresholds. All other parameters and plotting conventions are as in Figure 2

B. Comparisons

Since the main purpose of this study was to evaluate whether the ANSWER approach provides any significant benefit, it was important to compare its performance with other heuristics that have been suggested for identifying salient words in texts. Here, we compared ANSWER with seven other heuristics described below. The first three are based mainly on the frequency with which words occur in the corpus and in individual documents, while the other four are network-based significance metrics for words in the ASN.

1) *Frequency*: The words in the corpus are ranked by their frequency of occurrence in the corpus, with the most frequent regarded as salient. This works only if common words such as articles, pronouns, etc., have already been removed from the text.
2) *Mean TF-IDF*: The TF-IDF value for each word is calculated over all 1,404 documents, and the mean of this value is used for ranking. Words with higher mean TF-IDF are regarded as more salient. Unlike the previous three heuristics (and ANSWER), this method requires that the corpus be divided into documents and that these be known a priori.
3) *Max TF-IDF*: In this case, words are ranked by the maximum TF-IDF value over the 1,404 documents, so a word that is especially concentrated in even one document get a high score. This too requires that the corpus consist of known documents.
4) *Node Degree*: Words are ranked based on the number of edges incident on them, i.e., the number of words

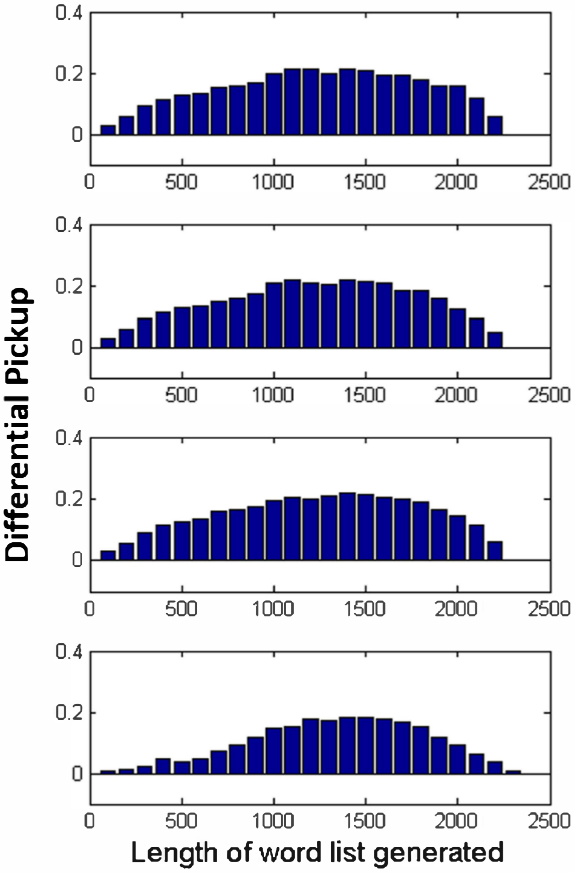

Fig. 5. The top three graphs show the differential in the fraction of salient versus the fraction of non-salient words picked up while setting the threshold to produce word lists of specific lengths. The graphs are plotted for values of $K = 5, 7$ and 10 (top to bottom). The bottom figure shows the same metric for picking words based on their per-sentence frequency of occurrence in the corpus.

with which they are associated in the corpus. Higher degree nodes are regarded as salient. Unlike frequency, this measure assigns salience to words based on their relationship with other words, but as with frequency, degree can only be used after common words have been removed.

5) *Node Weight*: Words are ranked based on the summed association weights of their incident edges, and higher weight nodes are designated as salient. In contrast to degree, this gives preference to nodes with strong associations rather than nodes with more numerous associations.
6) *Betweenness Centrality*: The betweenness centrality [56] of a node is defined as the fraction of shortest paths between all node pairs that pass through it. It is widely used in network analysis as a measure of node significance. Here, the betweenness centrality is used to rank words, with higher value indicating greater salience.
7) *Eigenvector Centrality*: The eigenvector centrality [56] is a recursively defined network metric quantifying the significance of a network node in proportion to the strength of its connections to other significant nodes. It can be shown [56] that, if A is the connectivity matrix of the network, the eigenvector centrality, c_e^k, of node k is proportional to u^k, the kth component of the eigenvector of A with the largest eigenvalue. The PageRank metric used in Google's search engine is a variant of eigenvector centrality. Here, the centrality is used to rank words, and words with higher values are regarded as more salient.

Each heuristic is calculated for all the words, and then thresholded to give a salient word list of a specific size.

C. Performance Evaluation

One key issue in performance evaluation is the fact that, of the list of N_V =2,309 words in the corpus vocabulary, N_{sal} = 1,666, i.e., about 72.15% are labeled as salient by the manual procedure (though the system is never trained on this), and the remaining N_{nonsal} = 643 as non-salient. This fact, which is partly a consequence of pre-processing having removed many non-salient words, means that only an algorithm that can achieve performance somewhat better than 72% is worth using.

Since each of the heuristics had a different range, the thresholds used for each had to be different. In order to make a principled comparison, we used the following approach:

1) ANSWER was run with a specific threshold, yielding a list of n_{trial} words.
2) The threshold for each other heuristic was chosen so that it resulted in extracting n_{trial} words.

Thus, if a threshold in ANSWER led to 900 words being identified as salient, the thresholds for frequency, max-TF-IDF, node degree, etc., were also set to yield 900 words. Of course, as the threshold increased, fewer words were picked up, yielding smaller and smaller lists. Given this situation, the obvious performance metric was to check what fraction of the words labeled salient actually were salient, i.e., true positives. As the threshold increases and lists become shorter, false negatives would grow automatically. However, it is still interesting to check whether the list that is selected is unusually dense in salient words, which would imply that ANSWER (or any of the other heuristics) was picking up some structure in the data.

The measure of performance, termed *efficiency*, was calculated as a percentage measure of true-positives:

$$\textit{Efficiency} = \frac{\textit{number of salient words identified}}{\textit{total number of words identified}} \quad (8)$$

Figures 2 , 3 and 4 show the comparative efficiency of ANSWER for a particular Kat different thresholds. ANSWER performs well in all cases and is better than other techniques in most cases except at very low thresholds for the $K = 10$ case. The number of words identified as salient depends strongly on the value of K. At low K and a very high threshold of around 0.02% (not shown in the figures), there are about 15 words in the list and 99% of them are salient. The general trend that can be observed is an increase in concentration of salient words obtained via ANSWER as we increase the threshold (decrease the length of the word list). This indicates that ANSWER indeed is using attractors as a concentrator of saliency in the corpus. On this dataset, the figures show that ANSWER with all three parameters did considerably better than the other heuristics, though Degree came close to catching up at low thresholds.

While useful, the efficiency metric does not capture the performance of the algorithm completely because of its exclusive focus on true positives. To get a broader view of performance, we looked at the relative level of salient and non-salient words picked up by both ANSWER and simple corpus frequency thresholding as the threshold was decreased from a high value towards zero. This is captured in a metric termed the *differential pickup*, D, defined as:

$$\begin{aligned} D(\phi) &= \textit{pickup of salient words at threshold } \phi \qquad (9) \\ &- \textit{pickup of non-salient words at threshold } \phi \end{aligned}$$

For example, suppose that at a threshold ϕ_1, 1,000 of the 2,309 words in the vocabulary have $S(w) \geq \phi_1$. This is the list of words deemed salient by ANSWER at that threshold. Suppose that, of these, 800 are actually salient and 200 non-salient. The percentage pickup of salient words is $800/N_{sal}$ = 0.4801 (since N_{sal} =1,666), and the pickup of non-salient words is $200/N_{nonsal}$ = 0.3110, since N_{nonsal} = 643. The differential pickup at threshold ϕ_1 is then, $D(\phi_1)$ = 0.4801−0.3110 = 0.1691. The positive value indicates that a greater fraction of salient words is being pickep up here than of no-salient ones. Thus, the more D can stay above 0, the better the algorithm.

Figure 5 shows the metric $D(\phi)$ for the ANSWER algorithm with different values of K (top three plots), and for thresholding based on corpus frequency (bottom plot). To make the plots comparable, we thresholded in each case to obtain list of identical lengths, and used this length as the x-axis in the figure. Thus, for example, the bars at the x-axis value 1,000 correspond to four distinct threholds in the four cases that had exactly 1,000 words greater than or equal to the threshold. This gives a meaningful plot in that one can ask how pure and complete a list of N salient words returned by the algorithms is going to be.

Several things are immediately noticeable from the figure. First, the plots for the three K values are extremely similar, indicating that ANSWER is not very sensitive to the choice of K in this range. Second, the graphs for the three ANSWER applications remain well above 0 over a wide range of word list lengths, whereas the plot for thresholding based on corpus frequency shows a much narrower peak. Third, the differential pickup by all the ANSWER algorithms is higher over a broad range of thresholds than the best pickup obtained via corpus frequency thresholding. Most importantly, the ANSWER algorithms have good performance at large list lengths (close to the actual number of salient words - 1,666). This too shows that using ANSWER is a good way to concentrate salience.

VII. Conclusion and Future Work

In summary, we have applied an unsupervised attractor network-based approach for detection of salient words in text corpora to a labeled corpus derived from abstract for IJCNN 2009, 2011 and 2013. The results reported here show that ANSWER did better than any of the other simple saliency detection metrics that were tried. The proposed algorithm has the advantage of being applicable to undifferentiated corpora, and of requiring specification of relatively few parameters.

Overall, while the results with ANSWER were encouraging, challenges remain and there is considerable room for improvement and extension. Some directions being explored currently include the following:

- Applying ANSWER to multiple corpora of different types, e.g., poetry, fiction, non-fiction texts, etc.
- Exploring other ways to specify association weights in the ASN, e.g., using poistwise mutual information.
- Applying ANSWER to corpora without pre-processing to determine whether it can identify and remove stop words and other obviously non-salient words. This would allow ANSWER to be used in domains where a reference list such as ELP is not available, and perhaps even to texts in other – possibly unknown – languages.
- Applying ANSWER iteratively to distil smaller but purer lists of highly salient words for use as keywords.
- Combining multiple ANSWER and non-ANSWER heuristics in a mixture-of-experts (MOE) setting.
- Applying ANSWER and MOE approaches to related problems such as topic extraction, document classification, keyword identification, search query generation, text summarization, etc.

Results on these will be reported in future papers.

Acknowledgement

The authors would like to thank Amer Ghanem, Sarjoun Doumit, Simona Doboli and Yoed Kenett for useful suggestions related to this work.

References

[1] H. P. Luhn, "The automatic creation of literature abstracts," *IBM Journal of Research and Development*, vol. 2, pp. 159–165, 1958.

[2] G. Salton, E. A. Fox, and H. Wu, "Extended boolean information retrieval," *Communications of the ACM*, vol. 26, pp. 1022–1036, 1983.

[3] S. G. and M. McGill, *Introduction to Modern Information Retrieval*. McGraw-Hill, 1986.

[4] M. Ortuno, P. Carpena, P. Bernola-Galvn, E. Muoz, and A. M. Somoza, "Keyword detection in natural languages and dna," *EPL (Europhysics Letters) 57*, vol. 5, pp. 759–764, 2002.

[5] C. E. Allison, A. G. Pearce and D. Abbott, "Finding keywords amongst noise: Automatic text classification without parsing." *SPIE Fourth International Symposium on Fluctuations and Noise, International Society for Optics and Photonics*, 2007.

[6] H. Zhou and G. W. Slater, "A metric to search for relevant words," *Physica A: Statistical Mechanics and its Applications*, vol. 329, pp. 309–327, 2003.

[7] J. P. Herrera and P. A. Pury, “Statistical keyword detection in literary corpora,” *The European Physical Journal B-Condensed Matter and Complex Systems*, vol. 63, pp. 135–146, 2008.

[8] M. Grineva, M. Grinev and D. Lizorkin, “Extracting key terms from noisy and multitheme documents.” in *Proceedings of the 18th international conference on World wide web, ACM*, 2009, pp. 661–670.

[9] J. Ventura and J. Silva, “Mining concepts from texts,” *Procedia Computer Science*, vol. 9, pp. 27–36, 2012.

[10] J. Leskovec, M. Grobelnik, and N. Milic-Frayling, “Learning substructures of document semantic graphs for document summarization,” in *Proceedings KDD 2004 Workshop on Link Analysis and Group Detection, Seattle, WA*, 2004.

[11] J. Leskovec, N. Milic-Frayling, and M. Grobelnik, “Impact of linguistic analysis on the semantic graph coverage and learning of document extracts,” in *Proceedings of the 12th National Conference On Artificial Intelligence*, 2005, p. 10691074.

[12] D. Rusu, B. Fortuna, D. Mladenic, M. Grobelnik, and R. Sipos, “Document visualization based on semantic graphs,” in *Proceedings of the 13th International Conference Information Visualisation*, 2009, pp. 292–297.

[13] L. Dali, D. Rusu, B. Fortuna, D. Mladenic, and M. Grobelnik, “Question answering based on semantic graphs,” in *Proceedings of WWW’2009, Madrid, Spain*, 2009.

[14] Y. Liang and Y. Liu, “Building a semantic graph based on sequential language model for topic-sensitive content extraction,” in *Workshop on Mining and Learning with Graphs, San Diego, CA*, 2011.

[15] C. Aggarwal and P. Zhao, “Towards graphical models for text processing,” *Knowledge and Information Systems*, vol. 36, pp. 1–21, 2013.

[16] D. L. Nelson, C. L. McEvoy, and T. A. Schreiber, “The university of south florida word association, rhyme, and word fragment norms. http://www. usf. edu/freeassociation/,” 1998.

[17] W. S. Maki, “Some properties of word pairs: Word association, semantic distance, and lexical co-occurrence. retrieved september 14, 2011 from psychonomic society web archive: http://www. psychonomic. org/archive/,” 2003.

[18] ——, “A database of associative strengths from the strength-sampling model: A theory-based supplement to the Nelson, McEvoy, and Schreiber word association norms,” *Behavior Research Methods*, vol. 40, pp. 232–235, 2008.

[19] J. G. W. Raaijmakers and R. Shiffrin, “Search of associative memory,” *Psychological review*, vol. 88, 1981.

[20] G. McKoon and R. Ratcliff, “Spreading activation versus compound cue accounts of priming: Mediated priming revisited,” *Journal of Experimental Psychology: Learning, Memory, and Cognition*, vol. 18, pp. 1155–1172, 1992.

[21] R. Ratcliff and G. McKoon, “Retrieving information from memory: Spreading-activation theories versus compound-cue theories,” *Psychological Review*, vol. 101, pp. 177–184, 1994.

[22] D. L. Nelson, D. J. Bennett, N. R. Gee, T. A. Schreiber, and V. McKinney, “Implicit memory: Effects of network size and interconnectivity on cued recall,” *Journal of Experimental Psychology: Learning, Memory, and Cognition*, vol. 19, pp. 747–764, 1993.

[23] D. L. Nelson and J. Xu, “Effects of implicit memory on explicit recall: Set size and word frequency effects,” *Psychological Research*, vol. 57, pp. 203–214, 1995.

[24] A. Collins and E. Loftus, “A spreading-activation theory of semantic priming,” *Psychological Review*, vol. 82, pp. 407–428, 1975.

[25] H. Moss, M. Hare, P. Day, and L. Tyler, “A distributed memory model of the associative boost in semantic priming,” *Connection Science*, vol. 6, pp. 413–427, 1994.

[26] M. Masson, “A distributed memory model of semantic priming,” *Journal of Experimental Psychology: Learning, Memory, and Cognition*, vol. 21, pp. 3–23, 1995.

[27] D. Plaut, “Semantic and associative priming in a distributed attractor network,” in *Proc. of the 17th Annual Conference of the Cognitive Science Society*. Pittsburgh, PA: Lawrence Erlbaum, Hillsdale, NJ, July 1995, pp. 37–42.

[28] I. Lerner and O. Shriki, “Internally- and externally-driven network transitions as a basis for automatic and strategic processes in semantic priming: theory and experimental validation,” *Frontiers in Psychology*, vol. 5, p. Article 314, 2014.

[29] S. Mednick, “The associative basis of the creative process,” *Psychological Review*, vol. 69(3), pp. 220–232, 1962.

[30] J. P. Guilford, *The Nature of Human Intelligence*. McGraw-Hill, 1967.

[31] T. M. Amabile, *The Social Psychology of Creativity*. Springer-Verlag, 1983.

[32] M. D. Mumford and S. B. Gustafson, “Creativity syndrome: Integration, application, and innovation,” *Psychological Bulletin*, vol. 103, pp. 27–43, 1988.

[33] D. K. Simonton, *Scientific Genius: A Psychology of Science*. Cambridge University Press, 1988.

[34] V. Brown, M. Tumeo, T. Larey, and P. Paulus, “Modeling cognitive interactions during group brainstorming,” *Small Group Research*, vol. 29, pp. 495–526, 1998.

[35] T. B. Ward, “Creative cognition, conceptual combination, and the creative writing of stephen r. donaldson,” *American Psychologist*, vol. 56, pp. 350–354, 2001.

[36] D. K. Simonton, “Scientific creativity as constrained stochastic behavior: the integration of product, person, and process perspectives,” *Psychol. Bull.*, vol. 129, pp. 475–494, 2003.

[37] ——, “Creative thought as blind-variation and selective-retention: Combinatorial models of exceptional creativity,” *Physics of Life Reviews*, vol. 7, pp. 156–179, 2010.

[38] A. E. Motter, A. P. S. de Moura, Y. C. Lai, and P. Dasgupta, “Topology of the conceptual network of language,” *Physical Review E*, vol. 65, p. 065102(R), 2002.

[39] M. Sigman and G. A. Cecchi, “Global organization of the wordnet lexicon,” *PNAS*, vol. 99, pp. 1742–1747, 2002.

[40] M. Steyvers and J. Tenenbaum, “The large scale structure of semantic networks: Statistical analyses and a model of semantic growth,” *Cognitive Science*, vol. 29, pp. 41–78, 2005.

[41] M. E. Bales and S. B. Johnson, “Graph theoretic modeling of large-scale semantic networks,” *Journal of Biomedical Informatics*, vol. 39, pp. 451–464, 2006.

[42] Y. Kenett, D. Kenett, E. Ben-Jacob, and M. Faust, “Global and local features of semantic networks: Evidence from the Hebrew mental lexicon,” *PLoS ONE*, vol. 6, p. e23912, 2011.

[43] A. Morais, H. Olsson, and L. Schooler, “Mapping the structure of semantic memory,” *Cognitive Science*, vol. 2012, pp. 1–21, 2012.

[44] Y. Kenett, D. Anaki, and M. Faust, “Investigating the structure of semantic networks in low and high creative persons,” *Frontiers in Human Intelligence*, vol. 8, p. Article 407, 2014.

[45] N. Marupaka and A. A. Minai, “Connectivity and creativity in semantic neural networks,” in *Proceedings of IJCNN 2011*, 2011, pp. 3127–3133.

[46] N. Marupaka, L. R. Iyer, and A. A. Minai, “Connectivity and thought: The influence of semantic network structure in a neurodynamical model of thinking,” *Neural Networks*, vol. 32, pp. 147–158, 2012.

[47] S. Doumit, N. Marupaka, and A. A. Minai, “Thinking in prose and poetry: A semantic neural model,” in *Proceedings of IJCNN 2013*, 2013.

[48] M. Mei, A. Vanarase, and A. A. Minai, “Chunks of thought: Finding salient semantic structures in texts,” in *Proceedings of IJCNN 2014*, 2014.

[49] A. A. Minai, L. R. Iyer, D. Padur, and S. Doboli, “A dynamic connectionist model of idea generation,” in *Proceedings of IJCNN 2009*, 2009, pp. 2109–2116.

[50] L. R. Iyer, A. A. Minai, S. Doboli, V. R. Brown, and P. B. Paulus, “Effects of relevant and irrelevant primes on idea generation: A computational model,” in *Proceedings of IJCNN 2009*, 2009, pp. 1380–1387.

[51] L. R. Iyer, S. Doboli, A. A. Minai, V. R. Brown, D. S. Levine, and P. B. Paulus, “Neural dynamics of idea generation and the effects of priming,” *Neural Networks*, vol. 22, pp. 674–686, 2009.

[52] L. R. Iyer, V. Venkatesan, and A. A. Minai, “Neurcognitive spotlights:configuring domains for ideation,” in *Proceedings of WCCI 2010*, 2010, pp. 3026–3033.

[53] A. Ghanem and A. A. Minai, “A multi-agent model for the coevolution of ideas and communities,” in *Proceedings of WCCI 2010*, 2010, pp. 388–395.

[54] J. J. Hopfield, “Neural networks and physical systems with emergent collective computational abilities,” *Proceedings of the National Academy of Sciences USA*, vol. 79, pp. 2554–2558, 1982.

[55] D. Amit, *Modelling Brain Function*. Cambridge, UK: Cambridge University Press, 1989.

[56] M. E. J. Newman, *Networks: An Introduction*. Oxford University Press, 2010.