PROJECT MUSE®

Unsupervised Short Answer Grading Using Spreading Activation
over an Associative Network of Concepts / La notation sans
surveillance des réponses courtes en utilisant la diffusion
d'activation dans un réseau associatif de concepts

Amir H. Jadidinejad, Fariborz Mahmoudi

Canadian Journal of Information and Library Science, Volume 38,
Number 4, December / décembre 2014, pp. 287-303 (Article)

Published by University of Toronto Press
DOI: <https://doi.org/10.1353/ils.2014.0018>

➔ *For additional information about this article*

<https://muse.jhu.edu/article/563462>

Unsupervised Short Answer Grading Using Spreading Activation over an Associative Network of Concepts

La notation sans surveillance des réponses courtes en utilisant la diffusion d'activation dans un réseau associatif de concepts

Amir H. Jadidinejad

Department of Computer Engineering, Science and Research Branch, Islamic Azad University, Tehran, Iran

a.jadidi@srbiau.ac.ir

Fariborz Mahmoudi

Computer and IT Engineering Faculty, Islamic Azad University, Qazvin Branch, Qazvin, Iran

mahmoudi@qiau.ac.ir

Abstract: In this article we address the problem of automatic short answer grading from an unsupervised knowledge-based text-relatedness perspective. A comprehensive knowledge base is automatically constructed with the aid of Wikipedia, which has been modelled as an associative network of concepts. Semantic marker gets a short answer, extracts the most important concepts, and initiates a spreading activation algorithm to measure the degree of relatedness between the student's answer and the corresponding ideal answer. The proposed approach significantly and consistently outperforms previous unsupervised methods for short answer grading.

Keywords: semantic relatedness, automatic short answer grading, spreading activation, Wikipedia-mining, concept network

Résumé : Dans cet article, nous abordons le problème de la notation automatique des réponses courtes, en nous plaçant dans la perspective d'une relation textuelle non surveillée et fondée sur le savoir. Une base de connaissances complète est construite automatiquement à l'aide de Wikipédia, modélisée en réseau associatif de concepts. Un marqueur sémantique obtient une réponse courte, extrait les concepts les plus importants et initie un algorithme de diffusion de l'activation pour mesurer le degré de parenté entre les réponses des étudiants et la réponse idéale correspondante. L'approche proposée surpasse de façon décisive et constante les méthodes non supervisées précédentes pour la notation de réponses courtes.

Mots-clés : relation sémantique, notation automatique de réponses courtes, diffusion de l'activation, exploration de données Wikipédia, réseau de concepts

Introduction

In contrast to multiple-choice questions that do not require sophisticated text understanding, free-text questions allow students to express and support their ideas in response to the question (Pinto, Doucet, and Fernández-Ramos 2010). Unfortunately, the grading of these questions is expensive and time-consuming and adds potential measurement errors resulting from inconsistencies in the grading process. In addition, there are certain scenarios in which an instructor is not available and yet students need an assessment of their knowledge (Griff and Matter 2013). In these instances, we often turn to automatic short answer grading (ASAG) systems.

“An automatic short answer grading system is one that automatically assigns a grade to an answer provided by a student, usually by comparing it to one or more correct answers” (Mohler and Mihalcea 2009). This problem is distinguished in the literature from paraphrase detection and key phrase extraction (Mohler, Bunescu, and Mihalcea 2011). The goals of ASAG are as follows:

1. *Automating the task of grading*: The task of grading is time-consuming and subjective. It is possible to provide inexpensive, immediate, and fair feedback on the correctness of answers using ASAG software.
2. *Supporting the learner with immediate feedback on their answers (pedagogy)*: In a typical exam, an instructor provides students with feedback on their answers (McLoughlin 2002). ASAG is a computer system that emulates the grading ability of a teacher, so it is important to explain to the user how the grade is achieved. It is one of the most important aspects of the ASAG process, which has been neglected in previous research (Jordan and Mitchell 2009).

In information science a *conceptualization* is an abstract simplified view of some selected part of the world and the relationships between them. An explicit specification of a conceptualization is an ontology with formally defined concepts and relations. Concept-based information retrieval leverages ontologies and goes beyond the word level and works with concepts instead (Stock 2010). Using concepts instead of words is essential for solving knowledge-intensive tasks such as ASAG.

The contributions of this article are twofold. First, we propose an *unsupervised knowledge-based* technique for ASAG using spreading activation over an associative network of concepts. Unlike previous works (Mohler, Bunescu, and Mihalcea 2011; Sukkarieh, Pulman, and Raikes 2004), which required labelled data or manually constructed patterns, we attempt to propose an unsupervised method that requires no human intervention. Second, we operationalize the proposed method by building a knowledge base from Wikipedia, visualizing the student's answer using activated concepts to provide implicit feedback to the learner regarding his or her answer, and finally evaluating the proposed method using a benchmark data set.

The proposed method is independent from the background knowledge base. Although it is possible to leverage a high-quality, domain-specific, and manually built associative network, we leverage Wikipedia (Medelyan et al. 2009) in our

Figure 1: A taxonomy of automatic short answer grading systems.

experiments because Wikipedia is not only a hi-fi common-sense knowledge base with great coverage but also is publicly available for mining (Medelyan et al. 2009). Even though Wikipedia editors are not required to be professional, the open editing approach yields remarkable quality (Giles 2005).

Literature review

Several different studies are concerned with the automatic assessment of student answers to comprehension questions. This problem is known as “short answer assessment” (Ziai, Ott, and Meurers 2012). Some systems focus on assessing whether or not the student has properly answered the question (Jordan and Mitchell 2009); others aim at giving a grade as accurately as possible, therefore not only assessing meaning, but also performing grading similar to expert teachers. In this section we review different studies in the field of grading. For more information on assessment, please refer to Ziai, Ott, and Meurers (2012).

Generally, the problem of ASAG is divided into supervised and unsupervised approaches. Several state-of-the-art supervised methods require manually constructed patterns (Information Extraction templates), which, if matched, indicate that the question has been answered correctly (Pulman and Sukkarieh 2005). Although some semi-supervised approaches have been proposed in the literature (Sukkarieh, Pulman, and Raikes 2004), constructing an annotated corpus is time-consuming and context-specific. This problem led researchers into unsupervised approaches. Figure 1 shows the taxonomy of related works.

Computing the semantic relatedness (Zhang, Gentile, and Ciravegna 2013) between the student answer and the corresponding ideal answer is a prominent approach in the task of ASAG (Leacock and Chodorow 2003; Mohler, Bunesco,

and Mihalcea 2011; Mohler and Mihalcea 2009). The methods of semantic relatedness are organized as corpus-based (statistical) or knowledge-based approaches.

Statistical approaches build a semantic space of words from the way in which these words are distributed in a corpus of unannotated natural language text (Baroni and Lenci 2010). Semantic space algorithms capture the statistical regularities of words in a text corpus and map each word to a high-dimensional vector that represents the latent concepts. Pérez et al. (2005) presents a combined approach that makes use of Latent Semantic Analysis and n -gram overlap.

Explicit Semantic Analysis (ESA), proposed by Gabrilovich and Markovitch (2009), is the most prominent approach in this field. ESA represents textual information with respect to the external textual article space, indicating how strongly a given word in the input text is associated with a specific article in that external space. In this model, two pieces of text are semantically related in spite of having no word in common.

Although statistically derived concepts can represent the meaning of text better than words, they have some limitations too. These concepts might be difficult to interpret in natural language. Also, polysemy is a serious problem since each occurrence of a word is treated as having the same meaning because the word is represented as a single point in the semantic space without considering its context.

On the opposite side of the statistical (corpus-based) approaches, several knowledge-based methods have been proposed. WordNet (Budanitsky and Hirst 2006) is a well-known resource which encodes different relations between words. It has been leveraged as the background knowledge in the task of semantic relatedness (Budanitsky and Hirst 2006). Criticisms are that manually building and maintaining lexical resources is time-consuming and expensive, while its coverage is limited when dealing with domain-specific technical terms (Zesch and Gurevych 2010). Recent successful works on collaboratively built lexical resources (Hovy, Navigli, and Ponzetto 2013), such as a taxonomy of Wikipedia (Ponzetto and Strube 2011), have revolutionized lexical repositories. They leverage Wikipedia and/or WordNet to build a graph and take advantage of graph theory algorithms for semantic processing. A major difference between these methods is how the background knowledge base is constructed. Ponzetto and Strube (2011) proposed a labelled taxonomy using deep analysis of the category hierarchy of Wikipedia, while Yeh et al. (2009) and Gouws, van Rooyen, and Engelbrecht (2010) proposed a directed graph using internal links. Most prominent corpus-based and knowledge-based approaches are evaluated by Mohler and Mihalcea (2009) for the task of ASAG.

The method proposed in this article does not require labelled (training) data, and it is applicable in real-world problems with fewer limitations. Therefore, it is classified as an unsupervised method. Methods based on computing semantic relatedness are the most effective unsupervised model in this field of study, and these are categorized in two types: knowledge-based and corpus-based. The proposed method is based on processing the knowledge base. Therefore, it is considered a knowledge-based method.

Figure 2: The network structure of our background knowledge base, which combines the category graph and the article graph in Wikipedia. Article–category (art–cat) links are shown as dotted lines.

Methodology

Our proposed method for solving ASAG contains two phases: during an offline phase, an associative network of concepts is generated using Wikipedia; after that, during an online phase, this background knowledge base is used to solve the task of grading with the aid of semantic marker algorithm (Pulman and Sukkarieh 2005). Semantic marker is responsible for extracting seed concepts from the student answer and the corresponding ideal answer, finding the semantic relatedness between them using the association links in the network.

Offline phase: Associative network construction

As mentioned above, the proposed method is independent from the background knowledge base. In this section, we leverage Wikipedia as a high-quality, publicly available source that offers great coverage of common-sense background knowledge (Medelyan et al. 2009) to build an associative network of concepts.

Each Wikipedia article is seen to represent a concept, and the hyperlink structure relates these concepts to one another. To ensure the topical relatedness of association links, we assume that two articles are associated if they link to each other in Wikipedia (Hu et al. 2009). Other kinds of internal links such as links between articles and parent categories (“art–cat” links) and hierarchical links between categories (“cat–cat” links) have been leveraged without revision. Each article or category is assumed to be a concept. In addition, some articles have equivalent categories, which have been merged together. Figure 2 shows the structure of the article and category graphs and the way we combine them.

Figure 3: An excerpt from the generated knowledge base from Wikipedia.

We used a Wikipedia snapshot as of November 20, 2007, to have comparable experiments to those in previous studies (Mohler, Bunesu, and Mihalcea 2011; Mohler and Mihalcea 2009). After parsing the Wikipedia XML dump using the provided scripts of Wikipedia-Miner (Milne and Witten 2013), about 2 million articles and categories were selected as concepts. Each concept has 7.67 associations on average. To store the repository for faster processing, a graph database was constructed and a visualizer module was developed for better representation and exploration of the underlying graph database. Figure 3 shows a symbolic representation of our associative network.

Online phase: Semantic marker

Semantic marker is responsible for comparing student answers with a manually defined ideal answer to automatically assign a grade. Figure 4 shows different parts of the semantic marker module. First, Wikifier (Milne and Witten 2013) extracts seed concepts from the input student (or ideal) answer. State-of-the-art methods of mapping text to concepts (entity-linking) are reviewed by Wang and

Figure 4: Wikifier (Milne and Witten 2013) extracts seed concepts from the student (or ideal) answer, and semantic marker is responsible for weighting concepts in the associative network according to the seed concepts. The output is an activation vector that assigns each concept $\{c_i\}$ to an activation value $\{a_i\}$.

Han (2014). After that, the semantic marker algorithm (Algorithm 1) sorts all concepts in the associative network according to their association with seed concepts. The output of the semantic marker algorithm is a weighted list of associated concepts, the so-called activation vector. Finally, to grade student answers based on the manually defined ideal answer, the corresponding activation vectors will be compared.

Let a_i denote the total energy for concept c_i , $Q(c_i)$ the set of c_j 's associated concepts, and w_{ij} the weight of association between concepts c_i and c_j . For a concept c_j , we can describe the classic model of spreading activation as follows (Cres-tani 1997):

$$a_j^* = \sum_{c_i \in Q(c_j)} D a_i w_{ij},$$

where D is a global decay factor. This decays activation exponentially in the path length, which penalizes activation transfer over longer paths. Since our background knowledge base is an undirected associative network, we can define the classic spreading activation on it by setting $w_{ij} = 0$ if there is no edge from i to j , and $w_{ij} = 1/d_i$ if there is an edge from i to j , where d_i is the degree of concept c_i .

Algorithm 1: Spreading activation from seed concepts S , up to level L , using global decay factor D and firing threshold F .

Required: L, D, F

function SemanticMarker(G, S)

1. $(C, A) \leftarrow G$ {Split the knowledge base into “Concepts” and “Associations”}
2. $R^t \leftarrow \emptyset$ {The current dynamic activation vector estimate}
3. $R^{t+1} \leftarrow \emptyset$ {The resulting improved activation vector estimate}
4. **for all** $s_i \in S$ **do**
5. $R^{t+1} \leftarrow s_i$
6. **end for**
7. **repeat**
8. $R^t \leftarrow R^{t+1}$ {Update the current activation vector with new estimate}


```

9.   for all  $p \in R^t$  do
10.    if  $\text{path\_length}(p) < L$  or  $R_p^t < F$  then {check for constraints}
11.    continue
12.  end if
13.   $Q_p \leftarrow$  the set of local concepts such that  $(p, q) \in A$  and  $q \in C$ 
14.  for all  $q \in Q_p$  do
15.     $R_q^{t+1} \leftarrow R_q^{t+1} + D \times R_p^t / |Q_p|$ 
16.  end for
17. end for
18. until there are no more concepts to fire

```

The semantic marker algorithm (Algorithm 1) can be briefly described as an iterative process of information percolation from seed concepts via local associative links. Each iteration is followed by another iteration until there are no more concepts to fire. During each iteration, some form of activation decay is applied to the active concepts (Crestani 1997). Iteration after iteration, the activation spreads over the network, reaching concepts that are far from the initially activated ones. So we used a path-length constraint (L) to avoid concept drift in the following experiments.

The algorithm takes an associative network G and a weighted set of seed concepts S as input. An association is represented as a pair (p, q) , where p and q are source and destination concepts. The concepts extracted using Wikifier (Milne and Witten 2013) are used as seed concepts. For example, consider a sample question/answer in Table 1. Seed concepts are shown in italics. These weighted list of concepts have been leveraged as the initial distribution of semantic marker algorithm (S).

The set of seed concepts (S) is a normalized set ($\sum_{i=1}^n S_i = 1$). So if the learner is just putting in all the concepts they can think of, the independent elements of the seed set (S_i) will have tiny weights, and semantic marker will be unable to spread the activation to the other concepts in the network because the firing threshold (F) prevents the spreading of noisy and outlier concepts. Seed concepts

Table 1. A sample question with the ideal answer and two short answers provided by students.

Sample question, correct answer, and student answers	Avg. grade
Question: What are the main advantages associated with <i>object-oriented programming</i> ?	
Ideal Answer: <i>Abstraction and reusability.</i>	
Student Answer 1: Existing <i>classes</i> can be <i>reused</i> , and <i>program maintenance</i> and <i>verification</i> are easier.	3.5
Student Answer 2: One of the main advantages is you can hide inner details using a technique known as <i>encapsulation</i> . <i>Objects</i> combine the data and operations, but you cannot see how it works. Another advantage is you can <i>reuse</i> classes that have been defined earlier in the <i>program</i> , a method known as <i>inheritance</i> . Finally, another advantage is that objects can determine appropriate operations at execution time, a technique known as <i>polymorphism</i> .	5.0

(S) lead to spreading activation and the firing of more concepts if there is a conceptual correlation between the seed elements (S_i). On the other hand, if the seed concepts are semantically related, then the new activated concepts will be charged by different concepts at the same iteration, and probably these activated concepts can pass the firing threshold constraint.

In the task of ASAG, we are dealing with short texts; therefore, these initial concepts (S) are very limited, and they are not appropriate for computing semantic relatedness. For each concept p in the network having an activation value R_p^t greater than the firing threshold F and no farther than L from seed concepts S , get the associated concepts and update the resulting activation vector. Concepts receiving a new activation value that exceeds the firing threshold F are marked for firing on the next iteration.

Finally, the resulting of spreading activation is a weighted list of semantically related concepts for each answer (R). These new activation vectors provide a better representation of the extracted seed concepts. Let $A_{ideal} = (c_1:a_1, c_2:a_2, \dots, c_N:a_N)$ and $A_{student} = (c_1:\hat{a}_1, c_2:\hat{a}_2, \dots, c_N:\hat{a}_N)$ denote the resulting activation vector for a pair consisting of the student answer and the corresponding ideal answer. We can grade the relatedness between these vectors using a metric motivated by information theory (Lin 1998):

$$grade(A_i, A_s) = \frac{\sum_{c_k \in A_i \cap A_s} a_k + \hat{a}_k}{\sum_{i=1}^N a_i + \sum_{j=1}^N \hat{a}_j}, 0 \leq grade(A_i, A_s) \leq 1$$

Experimental results

To evaluate the proposed method, we leveraged a data set of questions and answers provided by Mohler and Mihalcea (2009). This data set contains three assignments. Each assignment consists of seven short answer questions, and 30 students answered these assignments. Thus, the data set consists of a total of 630 student answers. The answers were independently graded by two human judges, using an integer scale from 0 (completely incorrect) to 5 (perfect answer). As mentioned before, the task of answer grading is highly subjective; in our data set, the two annotators correlated at $r = 0.6443$ (Mohler and Mihalcea 2009). Finally, the effectiveness is measured using Pearson's correlation coefficient (r) against the average of the human-assigned grades:

$$r = \frac{n(\sum x_i y_i) - (\sum x_i)(\sum y_i)}{\sqrt{n(\sum x_i^2)(\sum x_i)^2} \sqrt{n(\sum y_i^2)(\sum y_i)^2}}$$

where n is the total number of instances, while x and y correspond to the relatedness values determined by human and algorithm respectively. The model of spreading activation as introduced in Algorithm 1 relies on three important parameters, namely, maximum path length, L ; global decay, D ; and firing threshold, F . These parameters need to be optimized since they have a large

Table 2. Pearson's correlation of text relatedness grades with human judgements

	r	95% CI	R^2	Type
TF $\times$ IDF (baseline)	0.3647	[0.2926–0.4326]	0.1330	Unsupervised/Corpus-based
LSA Wikipedia	0.4286	[0.3596–0.4930]	0.1836	Unsupervised/Corpus-based
ESA Wikipedia	0.4681	[0.4013–0.5300]	0.2191	Unsupervised/Corpus-based
Shortest path	0.4413	[0.3730–0.5049]	0.1947	Unsupervised/Knowledge-based
Proposed method	0.5612	[0.5008–0.6162]	0.3149	Unsupervised/Knowledge-based

ESA = Explicit Semantic Analysis; LSA = Latent Semantic Analysis; TF $\times$ IDF = Term Frequency $\times$ Inverse Document Frequency.

influence on the accuracy. Following previous studies (Gouws, van Rooyen, and Engelbrecht 2010), we developed an experimental method to estimate these parameters by randomly selecting a sub-set (20%) of the question/answer data set and correlating the algorithm's grades with the human-assigned grades. A grid search was implemented to estimate the appropriate value for each parameter using a repeated holdout approach ($k = 5$) to reduce the possibility of overestimating the performance of our proposed method on the sample sub-set. Finally, we chose $L = 3$, $D = 0.7$, and $F = 0.1$; these parameters have been used in the following experiments.

The quantity of the association between the input answer and the extracted seed concepts depends on (1) the coverage of the knowledge base (associative network) and (2) Wikifier (an entity-linking module which is responsible for linking words to the corresponding concepts), which are both independent from the scope of the proposed method. In our experiments using Wikipedia and Mohler's data set (Mohler and Mihalcea 2009) with Milne's Wikifier (Milne and Witten 2013), we have not found any instances of unassociated input material among 630 student answers.

Table 2 shows the effectiveness of the proposed method compared to some other unsupervised corpus-based and knowledge-based methods. Given the state-of-the-art algorithms, hypothesis testing is replaced with CIs to make a comparison based on not only the correlation value but also the sample size. The proposed method is 20% better than the state-of-the-art unsupervised approach (Gabrilovich and Markovitch 2009). Mohler, Bunesco, and Mihalcea (2011) proposed a supervised approach for the task of grading, leveraging a new expanded version of a short answer grading data set (Mohler and Mihalcea 2009). Different configurations were evaluated, and the best correlation is 0.5180, which is comparable to our unsupervised approach.

To assign measures of accuracy to sample estimates, we used a resampling method. First, a sample size ($n < N$) is chosen; then, random samples are iteratively (10,000 times) selected and the correlations calculated. Finally, for each sample size, the following are calculated: mean correlations ($\bar{r}$), Student's t -test statistics (t), and the p -values associated with t -test. Table 3 shows the result of the experiment using different samples. Clearly, when the sample size is increased, the proposed method is significantly better than the state-of-the-art method.

Table 3. Statistical summary for different sample sizes ($\mu=0.4681$)

	20	40	60	80
$\bar{r}$	0.4807	0.5158	0.5260	0.5364
t	4.9139	25.923	38.179	53.621
p -value	9.07E-7	1.87E-143	7.42E-298	0.0

Discussion

The development of new methods of education, especially in the field of online learning, has caused the ASAG system to be considered an essential component in the educational system. For example, the Coursera website (<http://www.coursera.org>) contains a variety of courses from popular universities around the world. In each course, several thousand students from all over the world participate. The education process includes programming projects, quizzes, and homework assignments, which are often characterized as questions with short answers. It is clear that in an online education system with this scale, it is not possible to evaluate students' assignments using classic methods. Therefore, nowadays ASAG systems are one of the main essential elements of large-scale online learning systems (Lintean et al. 2010; Rus et al. 2009). Furthermore, they are one of the hottest research topics in natural language processing, information systems, and learning technology (Rus 2014).

One of the deficiencies in correcting homework manually refers to the dependence of the final score on the context, the grading type, and the background knowledge of trainer. Previous researchers using a set of specific questions have revealed that the scores produced by various trainers were not highly correlated (Mohler and Mihalcea 2009). In the other words, if grading is performed by different trainers (a situation that is not avoidable in large-scale learning systems), it will be a subjective process. This will cause the range of scores in one class to be higher/lower than in another class. Therefore, although using a manual grading system is superior to ASAG systems from some points of view, using ASAG systems in large-scale learning systems not only is unavoidable but also yields fair scores, as the responses presented by several thousand students are corrected by a unique system, not by various trainers with different grading styles.

On the other hand, the results generated by the proposed system can be revised by humans. As described previously, the performance of ASAG systems is influenced by the background knowledge base and its relations. In contrast to previous methods (Gabrilovich and Markovitch 2009; Lintean et al. 2010), the nature of proposed approach makes it able to represent concepts and their relations in the context (referring to Figure 3 from the context presented in Table 1). Therefore, it is possible for the trainer to discover the score of semantic relatedness determined by the system using concepts and relations presented in the visualizer component. Then, by changing the relations in the knowledge base, trainers can operate in a way to improve the performance of the system in

subsequent experiments. Of course, this is a matter for future studies and requires more investigation.

Wikipedia is not a source that universities would use or recommend to their students (Wheeler, Yeomans, and Wheeler 2008). Obviously, using a high-quality, domain-specific knowledge base or entity-linking module could not only improve the accuracy of the proposed method but also prepare it to be embedded in an intelligent tutoring system. It is worth noting that although the performance of the proposed method is influenced by the entity-linking (Wang and Han 2014) component and the method of defining concepts in the knowledge base, the proposed method is completely independent of both. In other words, mapping text to corresponding concepts in the knowledge base (entity-linking) and the structure of defined concepts are leveraged as inputs of the proposed algorithm, and it is independent of these two factors for computing semantic relatedness. This means that the proposed method is able to be applied to a wide range of entity-linking methods and various knowledge bases for computing the semantic similarity between a student's response and the ideal answer. In this article, to demonstrate the applicability of the proposed method in the task of ASAG, Wikipedia (Medelyan et al. 2009) has been used as knowledge base, and Wikifier (Milne and Witten 2013) has been employed as an entity-linking component. In recent years, comparing collaboratively constructed semantic resources (Hovy, Navigli, and Ponzetto 2013) such as Wikipedia (Medelyan et al. 2009) to expert-built semantic resources such as WordNet (Gurevych and Wolf 2010) has been one of the most popular research topics in the field of semantic computing (Gurevych and Zesch 2013; Zesch and Gurevych 2010; Zhang, Gentile, and Ciravegna 2013). The reason is that although expert-built semantic resources are devoid of any noise, not only in the definition of concepts, but also in the determination of relations among them, their domain of applicability is limited, and a small range of specialized concepts are covered by them. In contrast, although collaboratively constructed semantic resources do not offer the accuracy of expert-built semantic resources, thanks to the participation of a million users around the world they are able to cover a wide range of concepts. However, previous studies (Giles 2005) have shown that the accuracy of the Wikipedia knowledge base is comparable to that of the *Encyclopaedia Britannica*, while Wikipedia has covered a very wide range of concepts and is far larger than the *Encyclopaedia Britannica*.

Conclusion

This article proposed an unsupervised knowledge-based ASAG system. It leverages an associative network of concepts as background knowledge. The semantic marker algorithm embedded in the proposed model is responsible for ranking the whole network according to the initial seed concepts using spreading activation. Finally, the resulting activation vectors are leveraged to grade student answers in comparison with the corresponding ideal answer.

To operationalize the proposed method, we leveraged Wikipedia (Medelyan et al. 2009) and Wikifier (Milne and Witten 2013) as a valuable source of

common-sense knowledge and an outstanding entity-linking module, respectively (Wang and Han 2014). The experimental results revealed that our system leads to concepts which cannot be deduced from the input answer alone and consequently outperforms the conventional corpus-based and knowledge-based methods for ASAG (Mohler and Mihalcea 2009). For example, as illustrated in Figure 3, the semantic marker algorithm has achieved emergent concepts which are not visible in the student's response text (specifically "abstraction" and "reusability"), by receiving the initial concepts that appeared in the student's response and developing these concepts in the context of the knowledge base. These concepts provide a connection between the concepts presented by the student and the concepts presented by the trainer. Moreover, they play a key role in computing semantic relatedness. In other word, semantic relatedness is not based only on the overlap of terms/concepts among the student's response and the ideal answer, but a series of inferred concepts also plays a bridge role between them, such that semantic relatedness is computed more accurately and efficiently.

According to the definition, semantic relatedness between two concepts includes any type of relation, while semantic similarity is defined based on finite relations (Rusu, Fortuna, and Mladenicić 2014). The proposed semantic marker algorithm aims to compute semantic relatedness (not semantic similarity); it desires to discover any type of relations among input seed concepts by navigating the network structure and exploring the paths with a length of more than 1. Therefore, in contrast to expert-built semantic networks (such as WordNet), which include a variety of different relations among concepts, the concept network employed here is organized only based on associative relations. On the other hand, if the semantic marker algorithm performs based on different types of relations, it must also be able to infer types of concepts and use relations in different situations depending on the input seed concepts in the student's response or the ideal answer. Moreover, it must have customized navigation based on the input seed concepts. It is a remarkable improvement that has been established for future work.

Using concepts and associations not only is advantageous for comparing student answers with the corresponding correct answer but can also be useful in supporting the learner with feedback on his or her answers. The visualizer module (as shown in Figure 3) represents the relation between the input student answer (empty circles) and the ideal answer (solid circles) in a conceptual manner. At the same time, according to the Wikipedia writing rules (Medelyan et al. 2009), the first paragraph of each article concisely describes the content. The learner can leverage this potential to revise the answer and submit it again. For example, consider the first answer in Table 1; after posting the initial answer, the student would get a sub-graph of the associative network which is concentrated around "OOP," "Reusability," "Inheritance," and "Polymorphism" (as shown in Figure 3). Actually, semantic marker provides the learner with some associated concepts which were not mentioned in the initial answer. Also, the definition of each concept is available. Now, the student can check the structure and the corresponding definitions, revise the answer, and monitor both the new

associated concepts and the grade achieved. In contrast, corpus-based approaches (Baroni and Lenci 2010), such as LSA and ESA (Gabrilovich and Markovitch 2009), do not have the potential of structural concepts. So the output of corpus-based approaches is just a semantic relatedness score. Obviously, the potential benefits of the structural concepts used in the proposed method need elaborate details, which is beyond the scope of this paper and must be left for future work.

Although the performance of the proposed method is affected by both entity-linking and knowledge base components, the proposed algorithm is independent of both in computing the semantic relatedness between the student's response and the ideal answer. It can be extended to multilingual environments when the question or the provided answer could be in different languages. As previously mentioned, the Wikifier (Milne and Witten 2013) algorithm is leveraged for mapping text to concepts in the knowledge base. This component is able to train on documents in languages other than English. At the same time, the important thing is that existing concepts in Wikipedia are language independent. For example, although the "Natural Language Processing" article in Wikipedia has been translated into more than 38 languages, it is represented as a unique node in our associative network. Consequently, by developing a multilingual entity-linking component for supporting various languages, the proposed system would be able to operate in multilingual environments.

In recent years, entity-linking has been one of the major topics in the field of natural language processing, and various methods have been proposed by researchers (Wang and Han 2014). It is worth noting that linking multilingual texts to a knowledge base is a key principle in this field (Nothman et al. 2013). By leveraging modern entity-linking components (Wang and Han 2014), the proposed system can be extended to multilingual environments. This feature has great significance in large-scale learning systems including millions of students from all around the world.

References

- Baroni, M., and A. Lenci. 2010. "Distributional Memory: A General Framework for Corpus-Based Semantics." *Computational Linguistics* 36 (4): 673–721. http://dx.doi.org/10.1162/coli_a_00016.
- Budanitsky, A., and G. Hirst. 2006. "Evaluating WordNet-Based Measures of Lexical Semantic Relatedness." *Computational Linguistics* 32 (1): 13–47. <http://dx.doi.org/10.1162/coli.2006.32.1.13>.
- Crestani, F. 1997. "Application of Spreading Activation Techniques in Information Retrieval." *Artificial Intelligence Review* 11 (6): 453–82. <http://dx.doi.org/10.1023/A:1006569829653>.
- Gabrilovich, E., and S. Markovitch. 2009. "Wikipedia-Based Semantic Interpretation for Natural Language Processing." *Journal of Artificial Intelligence Research* 34:443–98.
- Giles, J. 2005. "Internet Encyclopaedias Go Head to Head." *Nature* 438 (7070): 900–901. <http://dx.doi.org/10.1038/438900a>. Medline:16355180
- Gouws, S., G. J. van Rooyen, and H. A. Engelbrecht. 2010. "Measuring Conceptual Similarity by Spreading Activation over Wikipedia's Hyperlink Structure." Paper

- presented at the Proceedings of the 2nd Workshop on the People's Web Meets NLP: Collaboratively Constructed Semantic Resources, Beijing, August 28.
- Griff, E. R., and S. F. Matter. 2013. "Evaluation of an Adaptive Online Learning System." *British Journal of Educational Technology* 44 (1): 170–76. <http://dx.doi.org/10.1111/j.1467-8535.2012.01300.x>.
- Gurevych, I., and E. Wolf. 2010. "Expert-Built and Collaboratively Constructed Lexical Semantic Resources." *Language and Linguistics Compass* 4 (11): 1074–90. <http://dx.doi.org/10.1111/j.1749-818X.2010.00251.x>.
- Gurevych, I., and T. Zesch. 2013. "Collective Intelligence and Language Resources: Introduction to the Special Issue on Collaboratively Constructed Language Resources." *Language Resources and Evaluation* 47 (1): 1–7. <http://dx.doi.org/10.1007/s10579-012-9178-z>.
- Hovy, E., R. Navigli, and S. P. Ponzetto. 2013. "Collaboratively Built Semi-structured Content and Artificial Intelligence: The Story So Far." *Artificial Intelligence* 194: 2–27. <http://dx.doi.org/10.1016/j.artint.2012.10.002>.
- Hu, J., G. Wang, F. Lochovsky, J.-T. Sun, and Z. Chen. 2009. "Understanding User's Query Intent with Wikipedia." Paper presented at 18th International Conference on World Wide Web, New York, April 20–24. <http://dx.doi.org/10.1145/1526709.1526773>.
- Jordan, S., and T. Mitchell. 2009. "E-Assessment for Learning? The Potential of Short-Answer Free-Text Questions with Tailored Feedback." *British Journal of Educational Technology* 40 (2): 371–85. <http://dx.doi.org/10.1111/j.1467-8535.2008.00928.x>.
- Leacock, C., and M. Chodorow. 2003. "C-rater: Automated Scoring of Short-Answer Questions." *Computers and the Humanities* 37 (4): 389–405. <http://dx.doi.org/10.1023/A:1025779619903>.
- Lin, D. 1998. "Automatic Retrieval and Clustering of Similar Words." In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics* Vol. 2, edited by Christian Boitet and Pete Whitelock, 768–74. Stroudsburg, PA: Association for Computational Linguistics. <http://dx.doi.org/10.3115/980691.980696>.
- Lintean, M. C., C. Moldovan, V. Rus, and D. S. McNamara. 2010. "The Role of Local and Global Weighting in Assessing the Semantic Similarity of Texts Using Latent Semantic Analysis." Paper presented at the 23rd International FLAIRS Conference, Daytona Beach, FL, May 19–21.
- McLoughlin, C. 2002. "Learner Support in Distance and Networked Learning Environments: Ten Dimensions for Successful Design." *Distance Education* 23 (2): 149–62. <http://dx.doi.org/10.1080/0158791022000009178>.
- Medelyan, O., D. Milne, C. Legg, and I. H. Witten. 2009. "Mining Meaning from Wikipedia." *International Journal of Human-Computer Studies* 67 (9): 716–54. <http://dx.doi.org/10.1016/j.ijhcs.2009.05.004>.
- Milne, D., and I. H. Witten. 2013. "An Open-Source Toolkit for Mining Wikipedia." *Artificial Intelligence* 194: 222–39. <http://dx.doi.org/10.1016/j.artint.2012.06.007>.
- Mohler, M., R. Bunescu, and R. Mihalcea. 2011. "Learning to Grade Short Answer Questions Using Semantic Similarity Measures and Dependency Graph Alignments." In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, Vol. 1, edited by Dekang Lin, 752–62. Portland: Association for Computational Linguistics.
- Mohler, M., and R. Mihalcea. 2009. "Text-to-Text Semantic Similarity for Automatic Short Answer Grading." Paper presented at the 12th Conference of the European Chapter

- of the ACL (EACL 2009), Athens, Greece, March 30–April 3. <http://dx.doi.org/10.3115/1609067.1609130>.
- Nothman, J., N. Ringland, W. Radford, T. Murphy, and J. R. Curran. 2013. "Learning Multilingual Named Entity Recognition from Wikipedia." *Artificial Intelligence* 194: 151–75. <http://dx.doi.org/10.1016/j.artint.2012.03.006>.
- Pérez, D., E. Alfonseca, P. Rodríguez, A. Gliozzo, C. Strapparava, and B. Magnini. 2005. "About the Effects of Combining Latent Semantic Analysis with Natural Language Processing Techniques for Free-Text Assessment." *Revista signos* 38:325–43.
- Pinto, M., A.-V. Doucet, and A. Fernández-Ramos. 2010. "Measuring Students' Information Skills through Concept Mapping." *Journal of Information Science* 36 (4): 464–80. <http://dx.doi.org/10.1177/0165551510369633>.
- Ponzetto, S. P., and M. Strube. 2011. "Taxonomy Induction Based on a Collaboratively Built Knowledge Repository." *Artificial Intelligence* 175 (9–10): 1737–56. <http://dx.doi.org/10.1016/j.artint.2011.01.003>.
- Pulman, S. G., and J. Z. Sukkarieh. 2005. "Automatic Short Answer Marking." In *Proceedings of the Second Workshop on Building Educational Applications Using NLP*, edited by Jill Burstein and Claudia Leacock, 9–16. Stroudsburg, PA: Association for Computational Linguistics. <http://dx.doi.org/10.3115/1609829.1609831>.
- Rus, V. 2014. "Opportunities and Challenges in Semantic Similarity." Paper presented at the 27th International FLAIRS Conference, Pensacola Beach, FL, May 21–23.
- Rus, V., M. Lintean, A. Graesser, and D. McNamara. 2009. "Assessing Student Paraphrases Using Lexical Semantics and Word Weighting." Paper presented at 2009 Conference on Artificial Intelligence in Education: Building Learning Systems That Care: From Knowledge Representation to Affective Modelling, Amsterdam, the Netherlands, 6–10 July.
- Rusu, D., B. Fortuna, and D. Mladenović. 2014. "Measuring Concept Similarity in Ontologies Using Weighted Concept Paths." *Applied Ontology* 9 (1): 65–95.
- Stock, W. G. 2010. "Concepts and Semantic Relations in Information Science." *Journal of the American Society for Information Science and Technology* 61 (10): 1951–69. <http://dx.doi.org/10.1002/asi.21382>.
- Sukkarieh, J. Z., S. G. Pulman, and N. Raikes. 2004. "Auto-Marking 2: An Update on the UCLES-Oxford University Research into Using Computational Linguistics to Score Short, Free Text Responses." Paper presented at the International Association of Educational Assessment, Philadelphia, PA, October.
- Wang, J., and J. Han. 2014. "Entity Linking with a Knowledge Base: Issues, Techniques, and Solutions." *IEEE Transactions on Knowledge and Data Engineering* 99 (PrePrints): 1.
- Wheeler, S., P. Yeomans, and D. Wheeler. 2008. "The Good, the Bad and the Wiki: Evaluating Student-Generated Content for Collaborative Learning." *British Journal of Educational Technology* 39 (6): 987–95. <http://dx.doi.org/10.1111/j.1467-8535.2007.00799.x>.
- Yeh, E., D. Ramage, C. D. Manning, E. Agirre, and A. Soroa. 2009. "WikiWalk: Random Walks on Wikipedia for Semantic Relatedness." Paper presented at the 2009 Workshop on Graph-Based Methods for Natural Language Processing, Stroudsburg, PA, August 7. <http://dx.doi.org/10.3115/1708124.1708133>.
- Zesch, T., and I. Gurevych. 2010. "Wisdom of Crowds versus Wisdom of Linguists—Measuring the Semantic Relatedness of Words." *Natural Language Engineering* 16 (1): 25–59. <http://dx.doi.org/10.1017/S1351324909990167>.

- Zhang, Z., A. L. Gentile, and F. Ciravegna. 2013. "Recent Advances in Methods of Lexical Semantic Relatedness: A Survey." *Natural Language Engineering* 19 (4): 411–79. <http://dx.doi.org/10.1017/S1351324912000125>.
- Ziai, R., N. Ott, and D. Meurers. 2012. "Short Answer Assessment: Establishing Links between Research Strands." Paper presented at the 7th Workshop on Innovative Use of NLP for Building Educational Applications (BEA-7), Montreal, Quebec, June 7. <http://aclweb.org/anthology/W/W12/W12-2022.pdf>.