Towards an Automated Model of Comprehension (AMoC)

Mihai Dascalu[1,2,3](✉), Ionut Cristian Paraschiv[1,3], Danielle S. McNamara[4], and Stefan Trausan-Matu[1,2,3]

[1] Department of Computer Science, University Politehnica of Bucharest,
060042 Bucharest, Romania
{mihai.dascalu,ionut.paraschiv,stefan.trausan}@cs.pub.ro

[2] Academy of Romanian Scientists, Splaiul Independenţei 54,
050094 Bucharest, Romania

[3] Research Technology S.R.L., Sos. Virtutii, nr. 19D, Bucharest, Romania

[4] Institute for the Science of Teaching and Learning, Arizona State University,
Tempe, AZ 85287-2111, USA
dsmcnama@asu.edu

Abstract. Reading is a complex cognitive process wherein learners acquire new information and consolidate their knowledge. Readers create a mental representation for a given text by processing relevant words that, along with prior inferred concepts, become activated and establish meaningful associations. Our automated model of comprehension (AMoC) uses an automated approach for simulating the ways in which learners read and conceptualize by considering both text-based information consisting of syntactic dependencies, as well as inferred concepts from semantic models. AMoC makes use of cutting edge Natural Language Processing techniques, transcends beyond existing models, and represents a novel alternative for modeling how learners potentially conceptualize read information. This study presents side-by-side comparisons of the results generated by our model versus the ones generated by the Landscape model.

Keywords: Comprehension modeling · Semantic models
Natural Language Processing · Landscape Model

1 Introduction

Reading is a complex cognitive process, which has been subject to many studies throughout the years. It is one of the most common means that learners use to acquire new information and consolidate existing knowledge. Moreover, text resources represent one of the primary sources for learning. Readers create mental representations, which includes previous knowledge, enabling them to comprehend the text. However, text materials are not customized depending on the individual reader, and they are usually addressed to specific categories of readers. As such, computational models that simulate the reading process can serve as important tools for creating personalized learning applications that support the educational process by presenting adequate materials to learners.

V. Pammer-Schindler et al. (Eds.): EC-TEL 2018, LNCS 11082, pp. 427–436, 2018.
https://doi.org/10.1007/978-3-319-98572-5_33

The Construction-Integration model [1] represents a semi-automated approach to simulating the comprehension process, extracting the information from a text and combining it with the reader's personal experience. The model is based on a cyclical process using sentence units and requires manually setting the words' initial activation scores that appear within the text as well as the connections between the words (or nodes). The CI model's construction process has two phases, each responsible for generating concepts and propositions using a different input set. The first phase, known as text-based construction, represents the initial activation of elements from the linguistic, semantic, and situation levels. During the second phase, the knowledge-based constructions are integrated using vector multiplication along with constraint satisfaction, wherein the various propositional nodes' activation levels and links to other nodes are modified depending on their relations in the network.

The aim of this paper is to introduce a novel state-of-the-art automated model of comprehension that can be used to simulate text reading for different categories of learners by employing different parameters and semantic models. In the next section we present two similar models, namely the CI and Landscape models, which are two of the most frequently employed models of comprehension. In the third section we introduce our automated model based on advanced Natural Language Processing (NLP) techniques, alongside a detailed comparison of the results obtained using the Landscape model. The last section concludes the paper and presents future experiments and improvements for our model.

2 Similar Models

2.1 The Construction Integration Model

The Construction-Integration (CI) model [1] represents a semi-automated approach that extracts the information from a text and combines it with the reader's personal experience. The CI model describes a framework used for studying memory in the form of a semi-automated computational model inspired from the way humans read and understand texts. The model is based on a cyclical process using sentence units and requires setting manually the words' activation scores that appear inside the text. The CI theory uses a bottom up approach that combines features from a symbolic system and from a connectionist system. On the one hand, the symbolic system consists of a rule-based system used to construct a network representation of the text and the activated words. On the other hand, the connectionist system uses a constraint satisfaction mechanism to generate a stabilized (or coherent) interpretation of the to-be-comprehended text.

The CI model's construction process has two phases, each one responsible for generating concepts and propositions using a different input set. The first phase, also known as text-based construction, combines elements from the linguistic, semantic, and situation levels. Linguistic elements are equivalent to syntactic links and have been neglected in many studies as they only reflect a surface level of comprehension. The semantic level uses rules to generate text propositions, which represent concepts regardless of form (i.e., including images). The situation level is topic specific and relies on domain and general knowledge to make inferences to generate links among concepts in the text. The second phase uses knowledge-based constructions, where various propositional nodes are added and which can vary in strength, depending on their relations. Each propositional node within the knowledge construction phase has an explicit similarity relatedness with a text-based node.

The CI model builds a square term matrix C containing $n + m$ elements, where n represents the number of words, propositions or concepts that appear in the text, and m the knowledge propositions selected from the long-term memory net or in response to specific task demands. The model makes use of subsequent multiplications of the manual input activation row vector A_1 with the term matrix until the change in mean activation value is less than some criterion value ($A_i = A_{i-1} * C$). In the end, the model generates the final activation vector A, which provides the predicted strength for each unit. The CI model also provides a long-term square matrix M with $p + q$ elements, where $M_{ij} = C_{ij} * A_i * A_j$ (See Fig. 1).

$$A = \left[a_i \,..\, a_j \,..\, a_n \,...\, a_{n+m}|\right]$$

$$C = \left[\begin{array}{ccccccccc}
 & a_1 & .. & a_j & .. & a_n & . & . & a_{n+m} \\
a_1 & . & . & . & . & . & . & . & . \\
. & . & . & . & . & . & . & . & . \\
a_i & . & . & c_{ij} & . & . & . & . & . \\
. & . & . & . & . & . & . & . & . \\
a_n & . & . & . & . & . & . & . & . \\
. & . & . & . & . & . & . & . & . \\
. & . & . & . & . & . & . & . & . \\
. & . & . & . & . & . & . & . & . \\
a_{n+m} & . & . & . & . & . & . & . & .
\end{array}\right]$$

$$M = \left[\begin{array}{ccccccccc}
 & a_1 & .. & a_j & .. & a_p & . & . & a_{p+q} \\
a_1 & . & . & . & . & . & . & . & . \\
. & . & . & . & . & . & . & . & . \\
a_i & . & . & c_{ij} * a_i * a_j & . & . & . & . & . \\
. & . & . & . & . & . & . & . & . \\
a_p & . & . & . & . & . & . & . & . \\
. & . & . & . & . & . & . & . & . \\
. & . & . & . & . & . & . & . & . \\
. & . & . & . & . & . & . & . & . \\
a_{p+q} & . & . & . & . & . & . & . & .
\end{array}\right]$$

Fig. 1. Activation vectors, coherence matrix and long-term memory matrix corresponding to the CI model.

Albeit an impactful theoretical model, the CI model lacks some aspects of automation. First, the activation scores from each step must to be added manually before the model is able to distribute them in the current cycle. Second, the knowledge expansion is also performed by hand, thus making hard to generalize the approach. These limitations have put on hold the further development of the model because more advanced validations have been challenging to realize without automation capabilities.

2.2 The Landscape Model

The *Landscape Model* [2] has been designed to simulate the fluctuation of the concepts' activation scores, similarly to the CI Model. The concepts' activation is set manually through strategic assumptions about the source of activation and the amount of activation [3]. Prior knowledge activation is achieved through two different mechanisms: cohort activation and coherence-based retrieval. The first mechanism serves the function of passively mapping related concepts to the reader's mental representation of the text [3]. Concepts are inter-connected forming cohorts or associative memory traces, and the whole group can be activated at once by simply activating one word within the text. The second mechanism, coherence-based retrieval, uses a coherence parameter ranging from 1 to 5 that represents a word's importance with regards to certain relations from the text (causal, temporal, or spatial connections): more superficial reading processes are represented by smaller parameter values.

A visual representation of results from the Landscape Model are presented in Fig. 3, alongside the target text (see Fig. 2), depicting how the words' activation scores theoretically evolved across subsequent sentences [2].

A young knight rode through the forest (1).
The knight was unfamiliar with the country (2).
Suddenly, a dragon appeared (3).
The dragon was kidnapping a beautiful princess (4).
The knight wanted to free her (5).
He wanted to marry her (6).
The knight hurried after the dragon (7).
They fought for life and death (8).
Soon, the knight's armor was completely scorched (9).
At last, the knight killed the dragon (10).
He freed the princess (11).
The princess was very thankful to the knight (12).
She married the knight (13).

Fig. 2. The "Knight" story. Sample text used for visualizing the Landscape Model (http://www.brainandeducationlab.nl/downloads).

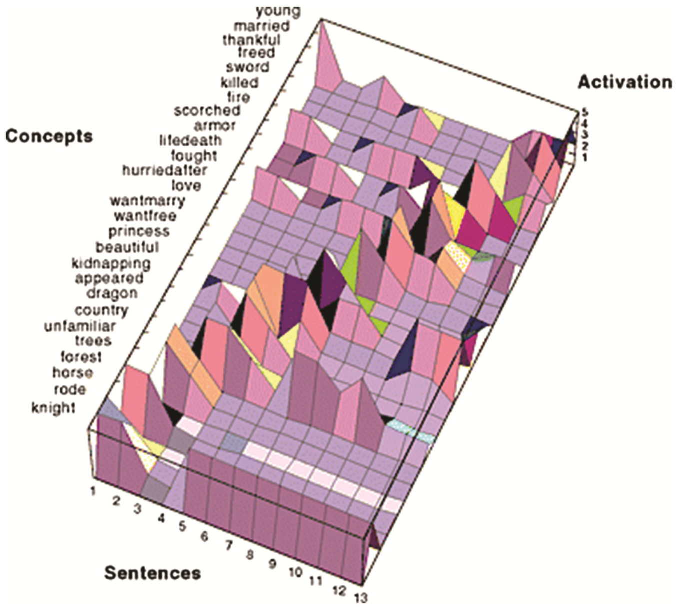

Fig. 3. Visualizing activation scores within the Landscape Model (http://www.brainandeducationlab.nl/downloads).

3 Current Study

Our automated model of comprehension (AMoC) introduces a fully automated method that analyzes the way in which readers potentially assimilate and conceptualize new text information. AMoC was developed on top of the *ReaderBench* framework [4], containing an extensive set of tools and models to analyze unstructured corpora. *ReaderBench* implements Cohesion Network Analysis which provides an in-depth perspective of discourse by relying on cohesive links identified between different text constituents [5]. Moreover, it contains a wide range of textual complexity indices covering syntactic, semantic and discourse structure levels of text analysis [4].

In its current form, AMoC makes use of lexicalized ontologies to determine synonyms that are used for semantic expansion. The system uses WordNet [6], a frequently used ontology in English, containing more than 150.000 concepts. These inferred words are subsequently compared to the rest of the concepts by using semantic models, representing pre-trained models that associate vectors to textual resources so that their semantic distance can be estimated through their cosine similarity. In our current implementation, we opted to rely on two representative and frequently used semantic models. First, Latent Semantic Analysis [7] creates a term-document matrix which counts words' appearances and applies Singular Value Decomposition followed by a dimensionality reduction. Second, the word2vec model, introduced recently in the literature [8, 9], adds support for words with multiple degrees of similarity along with inflections, and makes use of algebraic operations to determine meaningful similarities and links.

AMoC focuses on viewing a dataset from a micro level, in other words analyzes textual resources individually. The focus is on individual paragraphs, which are analyzed automatically through techniques similar to the way people read texts in general. Humans tend to create mental representations for the words encountered in the text, which in return activate other concepts from their memory. This process results in textual annotations that can be used later on to suggest which are the key points in every paragraph driving the evolution of topics within the text. The textual annotations can be the basis of intelligent reading applications used to help learners to understand textual resources better, even without reading them. AMoC, in its current form, analyzes each paragraph separately, only linking activation scores across sentences.

By reutilizing some basic principles from the CI model along with novel activation scores' computing and concepts' inference, AMoC is a novel approach that fully automates the textual annotations with activation scores. The model analyzes the input text in order to determine which are its most important words. It also infers semantically related words from lexicalized dictionaries to simulate the memory and/or knowledge of the reader. The main idea is that some words from the text are able to activate other terms that are not explicit in the text, but are theoretically available in prior knowledge: e.g., if someone reads a text that contains the word "cat", the concept may activate other concepts such as "feline", "lion", or "tiger" based on the semantic context.

Each sentence is comprised of at least one text-based word, further enriched through its dictionary synonyms, so the graph grows proportionally with the number of sentences and content words. Thus, an activation score is imposed that must be exceeded by all the words within the graph to be considered active. The reason behind this implementation choice is that when reading, humans have a short-term memory consisting of a global context with many inactive and only a few active concepts.

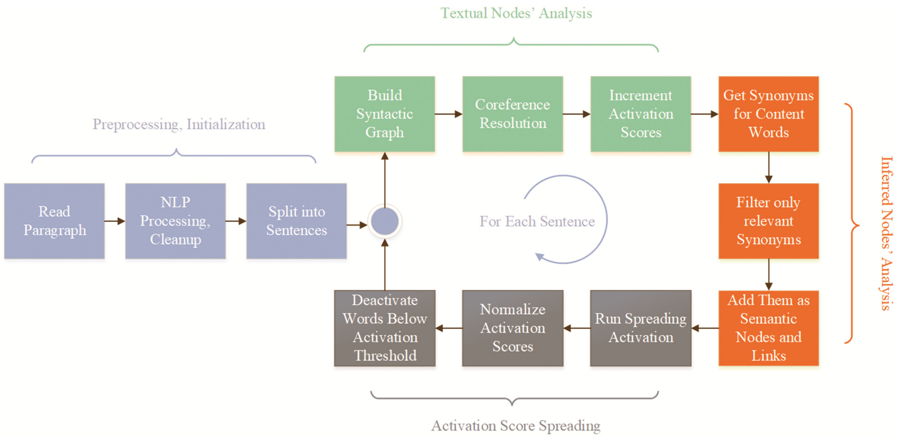
Textual Nodes' Analysis
Preprocessing, Initialization
Read Paragraph
NLP Processing, Cleanup
Split into Sentences
Build Syntactic Graph
Coreference Resolution
Increment Activation Scores
Get Synonyms for Content Words
Filter only relevant Synonyms
Add Them as Semantic Nodes and Links
Inferred Nodes' Analysis
For Each Sentence
Run Spreading Activation
Nomalize Activation Scores
Deactivate Words Below Activation Threshold
Activation Score Spreading

Fig. 4. Automated model of comprehension workflow.

Figure 4 depicts the implemented workflow that uses two types of links: syntactic links reflecting text-based associations between words, and semantic links highlighting semantic relatedness above an imposed threshold in semantic models. In the current analyses, we opted to use word embeddings from word2vec that provided the highest correlations with the activation scores from the Landscape Model, but other semantic models can be easily employed (e.g., Latent Semantic Analysis). The Landscape and CI models were not implemented within the current research and they represent only inspirational models.

During the preprocessing phase, the document undergoes a complete Natural Language Processing (NLP) pipeline which: cleans the input text, splits it into paragraphs and sentences, removes stop words, applies lemmatization, performs part-of-speech tagging and identifies content words (i.e., nouns, verbs, adjectives and adverbs), identifies syntactic dependencies, and replaces pronouns with corresponding nouns using pronominal resolution [10].

Next, the sentence's syntactic graph is extracted and merged within the global network graph depicting the memory's state. The activation scores corresponding to all content words from the sentences are incremented by 1. Afterwards, synonyms are extracted for all the content words using the WordNet ontology [6]. Only the most relevant synonyms (those having the highest semantic correlation with the whole text so far) are retained and merged alongside their corresponding semantic associations within the global network graph. Figure 5 depicts a use case of our model for the fifth sentence in the original text from Fig. 2, in which the semantic and syntactic links are shown, together with the mechanism of co-reference resolution.

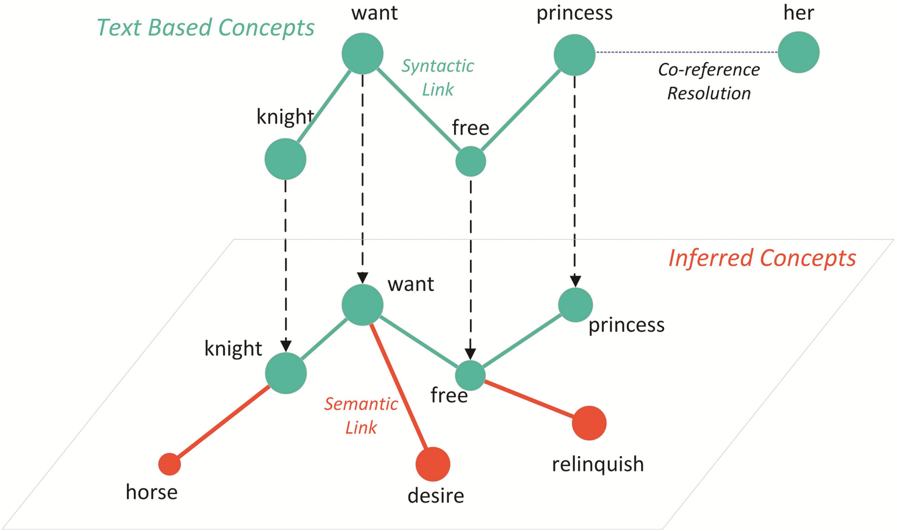

Fig. 5. Use case for "The knight wanted to free her".

Subsequently, spreading activation derived from the PageRank algorithm [11] is applied to distribute the activation strengths within the network, and only a limited number of words remain active (or words above a normalized activation score); follow-up sentences are treated in a similar manner by the model. The activation scores for each sentence are saved in order to render the three-dimensional visualizations (terrain rendering similar to the Landscape Model, 3D bar-charts and an evolutionary grid) in Figs. 6 and 7. As it can be easily observed, the most active concept is "knight", followed by "princess" and "dragon", central concepts within the presented story. These visualization techniques aim at depicting the evolution of the words' activation scores across subsequent sentences.

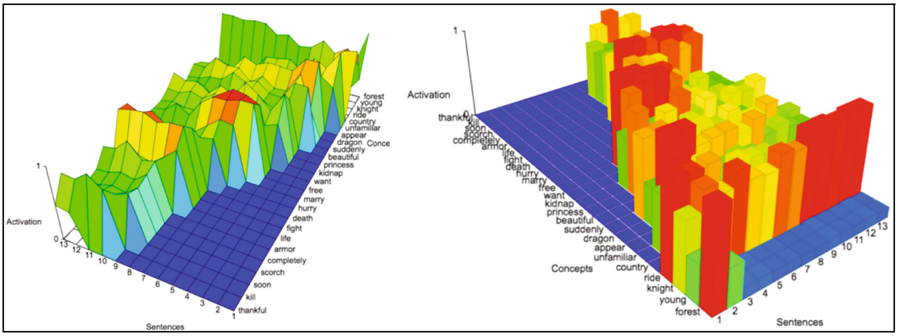

Fig. 6. Visualizing activation scores within AMoC using 3D bar-charts.

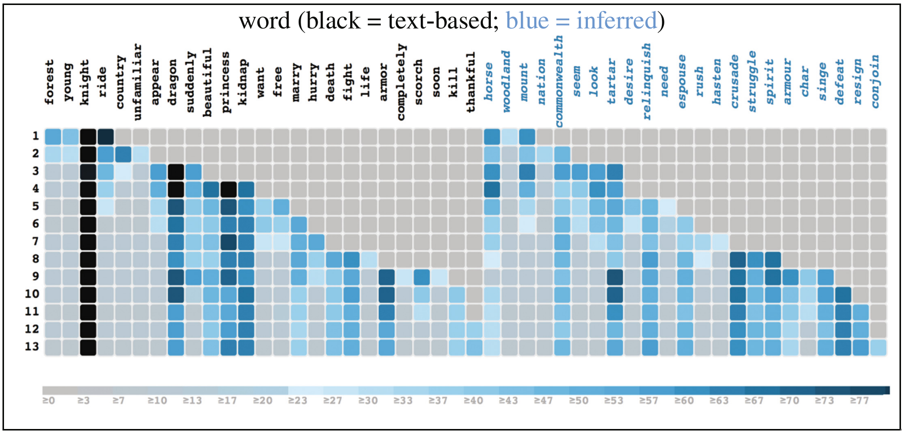

Fig. 7. Visualizing activation scores within AMoC using an evolutionary grid.

Statistical analyses were conducted to assess the extent to which AMoC measures words' activation scores in relation to predictions reported for the Landscape Model. The Landscape Model's activation scores were reported in previous experiments [12], and its values were compared with the ones generated from our model. The results of the experiment yielded high correlations (80%) between the activation scores of our model and the ones from the Landscape Model applied on the text from Fig. 2 (see Table 1). As such, the two models derive similar predictions, though the our model is entirely automated.

Table 1. Correlations between the activation scores from the Landscape Model and AMoC for the sentences from the "Knight" story (**p < .001).

Correlation	Sentence													Avg.
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	
Pearson	.927**	.872**	.899**	.907**	.825**	.530**	.726**	.809**	.823**	.768**	.550**	.879**	.910**	**.802****
Spearman	.898**	.924**	.932**	.810**	.739**	.462**	.547**	.670**	.592**	.554**	.557**	.569**	.737**	**.692****

4 Discussion and Conclusions

In this paper we introduce AMoC, an automated method for modelling human reading. AMoC can be used to simulate comprehension and offers the means to manipulate variables within the model in order to make model-based predictions. Text learning materials can be personalized by employing AMoC and learners may be helped, for example, by having highlights of central ideas. Understanding and simulating the reading process is a central element towards creating more contextualized learning environments that enhance the assimilation of new information.

AMoC represents a textual analysis tool that utilizes various Natural Language Processing techniques along with CI methods. The model annotates unstructured text with various computational methods that can be used for many purposes. First, the graphs' nodes are textual units which can be linked with various ontological facts, thereby simulating the activation of the semantic meaning by the reader. Secondly, understanding the most important words across sentences can be utilized for computing the overall textual complexity and to link readers to more detailed explanations.

In addition, understanding the reading process represents a key point in creating more contextualized learning environments that enhance the assimilation of new information and present learning materials tailored to the student's level. Moreover, this feature can enhance a student's learning experience in the context of cluttered domains with unstructured information. In other words, it can be a very useful tool in any educational context that relies on reading activities. AMoC was shown to have a high correlation (80%) with the results presented in the Landscape Model, an initial validation on top of which other experiments and validations can be built.

AMoC represents a completely autonomous method for simulating the human reading process. Besides validating the model, there are some parts which can be improved. Verbs tend to be more generic than other words, thus their semantic expansion is usually larger than nouns or adjectives. Furthermore, the static activation threshold seems to not be sufficient for filtering a small number of active concepts, so it should be replaced with a dynamic function. The current version of AMoC analyzes only activation scores within the sentences from a paragraph, thus we need to also account for activation scores between different paragraphs, which are already defined in the literature. Nonetheless, it represents a solid step towards a completely automated model of comprehension.

However, this is clearly only a first, albeit significant step. For example, one limitation is that the model currently focuses solely on a local analysis of texts, addressing only the short-term memory cycle (i.e., AMoC analyzes each paragraph separately, only linking activation scores across sentences). Additionally, various parameters need to be further tuned and the model needs to be subjected to extensive validations. Our current work is focusing on further extensions of the model and assessing the model's validity by comparing its predictions against prior research findings in the discourse literature.

Acknowledgment. The work presented in this paper was funded by the European Funds of Regional Development with the Operation Productivity Program 2014–2020 Priority Axe 1, Action 1.2.1 D-2015, "Innovative Technology Hub based on Semantic Models and High Performance Computing" Contract no. 6/1 09/2016.

References

1. Kintsch, W., Welsch, D.M.: The Construction-Integration Model: A Framework for Studying Memory for Text, p. 21. Institute of Cognitive Science, Boulder (1991)
2. van den Broek, P., Young, M., Tzeng, Y., Linderholm, T.: The landscape model of reading. In: van Oostendorp, H., Goldman, S.R. (eds.) The Construction of Mental Representations During Reading, pp. 71–98. Erlbaum, Mahwah (1999)
3. McNamara, D.S., Magliano, J.: Toward a comprehensive model of comprehension. Psychol. Learn. Motiv. **51**, 297–384 (2009)
4. Dascalu, M., Dessus, P., Bianco, M., Trausan-Matu, S., Nardy, A.: Mining texts, learner productions and strategies with *ReaderBench*. In: Peña-Ayala, A. (ed.) Educational Data Mining. SCI, vol. 524, pp. 345–377. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-02738-8_13
5. Trausan-Matu, S., Stahl, G., Sarmiento, J.: Polyphonic support for collaborative learning. In: Dimitriadis, Y.A., Zigurs, I., Gómez-Sánchez, E. (eds.) CRIWG 2006. LNCS, vol. 4154, pp. 132–139. Springer, Heidelberg (2006). https://doi.org/10.1007/11853862_11
6. Miller, G.A.: WordNet: a lexical database for English. Commun. ACM **38**(11), 39–41 (1995)
7. Landauer, T.K., Dumais, S.T.: A solution to Plato's problem: the latent semantic analysis theory of acquisition, induction and representation of knowledge. Psychol. Rev. **104**(2), 211–240 (1997)
8. Mikolov, T., Sutskever, I., Chen, K., Corrado, G., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Advances in Neural Information Processing Systems, pp. 3111–3119 (2013)
9. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representation in vector space. In: Workshop at ICLR, Scottsdale (2013)
10. Manning, C.D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S.J., McClosky, D.: The Stanford CoreNLP Natural Language Processing toolkit. In: 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations, pp. 55–60. ACL, Baltimore (2014)
11. Page, L., Brin, S., Motwani, R., Winograd, T.: The PageRank citation ranking: bringing order to the web. Stanford InfoLab (1999)
12. Britton, B.K., Graesser, A.C.: Models of understanding text. In: Britton, B.K., Graesser, A.C. (eds.) Models of Understanding Text. Psychology Press, New York (1995)