A Spreading Activation Framework for Tracking Conceptual Complexity of Texts

Ioana Hulpuş[1], **Sanja Štajner**[2] and **Heiner Stuckenschmidt**[1]
[1]Data and Web Science Group, University of Mannheim, Germany
[2] Symanto Research, Nürnberg, Germany
{ioana,heiner}@informatik.uni-mannheim.de
sanja.stajner@symanto.net

Abstract

We propose an unsupervised approach for assessing conceptual complexity of texts, based on spreading activation. Using DBpedia knowledge graph as a proxy to long-term memory, mentioned concepts become activated and trigger further activation as the text is sequentially traversed. Drawing inspiration from psycholinguistic theories of reading comprehension, we model memory processes such as semantic priming, sentence wrap-up, and forgetting. We show that our models capture various aspects of conceptual text complexity and significantly outperform current state of the art.

1 Introduction

Reading comprehension has long been linked to processes over semantic memory, such as semantic priming through spreading activation (Anderson, 1981; Collins and Loftus, 1975; Neely, 1991; Gulan and Valerjev, 2010). While psycholinguistic literature abounds in research and demonstration of such processes (Just and Carpenter, 1980; Kutas and Hillyard, 1984; Carroll and Slowiaczek, 1986), there is a gap in understanding if they can be modeled in an automated way for capturing the cognitive load required by texts. At the same time, the recent advances in the publication of encyclopedic knowledge graphs provide an unprecedented opportunity for modeling human knowledge at scale.

We focus on conceptual complexity which, as opposed to lexical and syntactic complexity (Vajjala and Meurers, 2014; Ambati et al., 2016), has received very little attention so far. Conceptual complexity accounts for the background knowledge necessary to understand mentioned concepts as well as the implicit connections that the reader has to access between the mentioned concepts in order to fully understand a text. It plays an important role in making texts accessible to children, non-native speakers, as well as people with low literacy levels or intellectual disabilities (Arfé et al., 2017). Apart from being one of the main factors for understanding the story, conceptual complexity also influences the readers' interest in the text: readers who lack relevant background knowledge have difficulties in understanding conceptually complex texts (Arfé et al., 2017; Benjamin, 2012), while high-knowledge readers need some obstacles (more conceptual complexity) to maintain their interest (Arfé et al., 2017; Benjamin, 2012; Kalyuga et al., 2003). Therefore, correctly estimating conceptual complexity of a text, and offering a reader a text of an appropriate cognitive load, is of utmost importance for: (1) ensuring correct understanding of a text; (2) maintaining the readers' interest; and (3) promoting deeper-level processing and enhancing the readers knowledge.

In this paper, we are building on top of the psycholinguistic findings that words are recognized faster if preceded by words related in meaning (semantic priming) (Gulan and Valerjev, 2010), and we adopt spreading activation theory as one of the main theories that tries to explain how priming occurs. Specifically, we introduce a framework that considers sequential text reading and models two simultaneous processes: (i) a spreading activation process that runs over long-term memory (approximated by the knowledge graph), activates concepts and transfers them to working memory, and (ii) a process that tracks concepts and their activation in working memory and subjects them to forgetting. We use the activation values of concepts in working memory at different points in the text in order to assess the amount of priming triggered by the text. Our hypothesis is that the higher these activation values (more priming), the lower the conceptual complexity.

We validate our framework through extensive experiments, and show that the models we propose on top of it outperform state-of-the-art measures that aim to predict conceptual complexity.

2 Related Work

In spite of its real-world importance, automatic assessment of conceptual complexity of texts has not received much attention so far. A few approaches have been proposed, but most of them are either not freely available, or have not been tested on large corpora (see (Benjamin, 2012) for the extensive list of approaches and their shortcomings). DeLite (vor der Brück et al., 2008) software and Coh-Metrix (Graesser et al., 2004), for example, do not have any features related to conceptual clarity, which would measure ambiguity, vagueness, and abstractness of a concept, or the level of necessary background knowledge. From this perspective, the work of Štajner and Hulpuş (2018) is the only work that attempts to automatically measure conceptual complexity of texts. They propose a supervised method using a set of graph-based features over DBpedia knowledge graph. In our experiments, we use these features as state-of-the-art for comparison with our approach.

In the cognitive science domain, the work most related to ours is in the direction of capturing knowledge in cognitive architectures (Lieto et al., 2018). Salvucci (2014) proposes the use of DBpedia as a source of declarative knowledge to be integrated with the ACT-R cognitive architecture (Anderson and Lebiere, 1998). They implement a very basic spreading activation model for scoring facts in the knowledge base for answering natural language, factual questions such as "What is the population of Philadelphia?". Several other approaches have been proposed for extending ACT-R with knowledge and reasoning (Ball et al., 2004; Oltramari and Lebiere, 2012), but none of them aim to assess the complexity of texts.

With respect to spreading activation, it has long been adopted as a methodology for information retrieval (Crestani, 1997), used for document summarization (Nastase, 2008), document similarity (Syed et al., 2008), as well as cross-domain recommendation (Heitmann and Hayes, 2016), among others. Nevertheless, there is no prior attempt to apply spreading activation to the recently developed encyclopedic knowledge graphs with the purpose of modeling reading comprehension.

This paper fills in this gap and shows that pairing spreading activation with other working memory processes (such as forgetting) can result in models that accurately assess conceptual complexity of a document.

3 Framework for Unsupervised Assessment of Conceptual Complexity

Our framework tracks the activation of concepts in working memory during reading processes. We consider an encyclopedic knowledge graph, DBpedia[1], as a proxy to long-term memory over which spreading activation processes run and bring concepts into the working memory. Text is processed sequentially, and each mention of a DBpedia concept triggers a tide of spreading activation over the DBpedia knowledge graph. Once brought into working memory, the activated concepts are subject to a forgetting process which decays their activation as the text is being read. At the same time, concepts in working memory accumulate more activation as they are repeated, or as related concepts are mentioned.

We track the cumulative activation (CA) of the mentioned concepts at different points in time: at encounter (AE), at the end of sentences (AEoS) and at the end of paragraphs (AEoP). We use these values to estimate the conceptual complexity of texts, under the overarching hypothesis that a higher activation of text concepts in working memory indicates more accessible texts.

3.1 Spreading Activation over DBpedia

For the spreading activation (SA) process, we exploit the graph structure of DBpedia. Each DBpedia concept is a node in the knowledge graph (KG). Each triple $<s, p, o>$ (short from $<subject, predicate, object>$) whose subject and object are DBpedia concepts, becomes a typed relation (or typed edge), that we denote with $s \xrightarrow{p} o$. This way, the knowledge base is represented as a graph $KG = (V, E, T, \tau)$, where V is the set of concepts, E is the set of directed relations between the concepts and $\tau : E \rightarrow T$ assigns a type in T to each edge in E. We denote by $\rho(x) \subset E$ the set of all relations of node $x \in V$, and by $n_r(x) \in V$ the neighbour of x through relation $r \in E$. We denote by $A^{(p)}(c)$ the amount of activation node c has after pulse p, by $A^{(p)}_{out}(c)$ the amount of activation node c outputs at pulse

[1] http://dbpedia.org

p and $A_{in}^{(p)}(c)$ the amount of activation that flows into node c at pulse p.

The core idea common to all SA models in literature is that concepts become active and *fire*, spreading their activation to their neighbors in KG, who in turn *fire* and activate their neighbors and so on, until preset termination conditions are met. Therefore, the SA process consists of multiple iterations called *pulses*.

In our model, a SA process is triggered whenever a concept is mentioned in the text (the seed concept), by setting its activation to 1.0, and that of all other nodes in V to 0.0. Formally, the initial conditions are $A^{(0)}(seed) = 1.0$ and $A^{(0)}(i) = 0.0, \forall i \in V, i \neq seed$. Then at pulse 1, the $seed$ fires and the SA process starts.

Formally, a SA model must define three functions: the output function, the input function and an activation function (Berthold et al., 2009; Crestani, 1997). In the following, we describe how we define these functions in order to study conceptual complexity of text.

The output function defines how much activation is output by a concept at pulse $p+1$, given its activation at current pulse p. To define this function, we use a *distance decay parameter* α, which decays the activation going out of each node exponentially with respect to p. Furthermore, our output function limits the concepts that fire to those concepts whose activation surpasses a given *firing threshold* β for the first time. Hence, α and β control the number of activated concepts and the intensity of their activation, providing potential for personalization according to memory capacity of the target audience.

$$A_{out}^{(p+1)}(c) = \alpha \cdot f_\beta(A^{(p)}(c)); \tag{1}$$

where $f_\beta(x) = x$ if $x \geq \beta; 0$ otherwise.

The input function aggregates the amount of activation that flows into a node (called target node) given the activations flowing out of their neighbours (called source nodes). Drawing inspiration from spreading activation theory in cognitive science (Collins and Quillian, 1969; Collins and Loftus, 1975), we define *accessibility* of a target concept given a source concept based on how strong is the semantic relation between them, as well as by how familiar the target concept is to the reader. We define the *strength* of the semantic relation between two nodes as its *exclusivity*, introduced by Hulpuş et al (2015) and proven by Zhu and Iglesias (2017) to be particularly effective for computing semantic relatedness. Regarding the user's familiarity with the target concept, in absence of user data we approximate it by the *popularity* of the target concept computed as the normalized node degree as $pop(c) = \frac{log(D(c))}{log(|V|-1)}$, where $D(c)$ denotes the number of neighbors of concept c.

Formally, given the relation $s \xrightarrow{p} o$, the accessibility scores of its endpoints s and o are computed as shown in Formula 2.

$$\begin{aligned} acc(o, s \xrightarrow{p} o) &= excl(s \xrightarrow{p} o) \cdot pop(o) \\ acc(s, s \xrightarrow{p} o) &= excl(s \xrightarrow{p} o) \cdot pop(s) \end{aligned} \tag{2}$$

Consequently, although the edges of the KG are directed, activation can flow in both directions over the same edge. For example, given the relation $Accordion \xrightarrow{isA} Musical\_instrument$, the mention of $Accordion$ will activate the concept $Musical\_instrument$, and vice-versa.

We can therefore generalize our notation, so that given a concept c and one of its relations (incoming or outgoing), r, c's accessibility over the relation r is defined as $acc_r(c) = excl(r) \cdot pop(c)$.

To make sure that the total amount of activation received from a concept by its neighbours, equals the amount of activation it outputs, we normalize the accessibility value as in Formula 3.

$$\overline{acc}_r(c) = \frac{acc_r(c)}{\sum_{r' \in \rho(c)} acc_{r'}(n_{r'} \circ n_r(c))}; \tag{3}$$

Finally, the input function is defined in Formula 4.

$$A_{in}^{(p+1)}(c) = \sum_{r \in \rho(c)} A_{out}^{(p+1)}(n_r(c)) \cdot \overline{acc}_r(c) \tag{4}$$

The activation function is the function that computes the activation of a concept after the pulse $p+1$, given its activation at time p and its incoming activation at $p+1$. Formally,

$$A^{(p+1)}(c) = A^{(p)}(c) + A_{in}^{(p+1)}(c) \tag{5}$$

In order to avoid cycles in which concepts keep activating each other, we constrain the process so that a concept can only fire in the first pulse after its activation overpasses β.

After firing, concepts become *burned* and although during the future pulses they can receive activation, they cannot fire again. When there are no more unburnt concepts with an activation higher than β in the graph, the SA process finishes. The activations resulted after this process are the activations that the nodes have after the last pulse, denoted in the following by $SA(\cdot)$.

3.2 The Working Memory Model

At the beginning of a text, the working memory (WM) is considered empty. As the text is being read, the concepts activated through SA are brought into WM with an activation computed as a function of their SA activation $\phi(SA(c))$.

The WM keeps track of all activated concepts and aggregates the activation that they achieve from different SA processes. Furthermore, a forgetting process takes place on top of WM, that is triggered at every word. We therefore use words as time unit in our WM model. The forgetting process decays the activations of all concepts in WM with a preset decay factor at every encountered word (γ_w), and additionally at every end of sentence (γ_s) and at every end of paragraph (γ_p). Therefore, given the words at indices i and j, ($i < j$), in paragraphs p_i and p_j ($p_i \leq p_j$) and sentences s_i and s_j ($s_i \leq s_j$) respectively, we denote the decay that occurs in the interval of time between the two words as $\gamma_{i,j}$ and compute it as in Equation 6.

$$\gamma_{i,j} = {\gamma_w}^{j-i} \cdot {\gamma_s}^{s_j - s_i} \cdot {\gamma_p}^{p_j - p_i}. \quad (6)$$

We define *cumulative activation*, $CA^{(i)}(c)$ of a concept c as its activation in the WM at time of reading word i. It is defined recursively as it consists of the cumulative activation that the concept has at time $i-1$ and that has been subject to forgetting, together with the activation $\phi(SA^{(i)}(c))$ that it receives as a result from the SA process that takes place at time i (see Equation 7).

$$\begin{aligned} CA^{(i)}(c) &= \gamma_{i-1,i} CA^{(i-1)}(c) + \phi(SA^{(i)}(c)) \\ &= \sum_{k=0}^{i} \gamma_{k,i} \phi(SA^{(k)}(c)) \end{aligned} \quad (7)$$

We illustrate this process with an example (Table 1) which shows, after the given text having been linked to DBpedia, the seed concepts corresponding to each mention and the set of text concepts activated by the seed concepts. Figure 1 shows the evolution of concepts' activation in WM, e.g. the concept db:Shelf_(storage) becomes active when it is mentioned, with an activation of 1.0. We compute the activations in Figure 1 by defining the function ϕ as a constant function in which all concepts that become active in the SA process receive an activation of 1 in WM. We denote this function as ϕ^1. In this example, we use values 0.85 and 0.7 for word (γ_w) and sentence decay (γ_s), respectively. The forgetting process is also illustrated as, unless reactivated, the CA scores decrease with every token, and the decrease is stronger after each sentence. The figure also shows how the concepts' CAs get adjusted every time they are activated by mentioned concepts. For example, at the time "instruments" is mentioned, the concepts db:Musical_instrument, and db:Accordion increase their existing CAs, and db:Band_(rock_and_pop) becomes active in WM.

3.3 Estimating Conceptual Text Complexity

One of the hypotheses that we want to test is that our framework can capture the *forward priming* phenomenon. We therefore hypothesize that in simpler texts, target concepts already exist in WM before they are explicitly mentioned in the text. In other words, the higher $CA(c)$ at the encounter of concept c, the easier it is to comprehend the concept c and connect it to its context. Considering concept c_i is the concept encountered in text at time i, its activation at encounter (AE) is $CA^{(i-1)}(c_i)$, hence its CA at the time of the word that precedes it.

$$AE(c_i) = CA^{(i-1)}(c_i) \quad (8)$$

Furthermore, the psycholinguistic theory of *backward semantic priming* states that concepts can actually receive activation from concepts mentioned afterwards, in a way explaining their previous occurrence. To account for this, concepts keep accumulating CA in WM after they are mentioned. More over, in the psycholinguistic literature the end of sentences have been proven to trigger a *wrapping up* process (Just and Carpenter, 1980), in which the information of the sentence is being reviewed. Based on these insights, we hypothesize that in simpler texts, the concepts exhibit a higher CA at the end of the sentences / paragraphs they occur in, than in more conceptually complex texts. Formally, given a sentence s, and denoting

Mention	Seed Concept	Activated text concepts
shelves	db:Shelf_(storage)	db:Shelf_(storage)
accordions	db:Accordion	db:Accordion, db:Musical_instrument
instruments	db:Musical_instrument	db:Musical_instrument, db:Accordion, db:Band_(rock_and_pop)
pictures	db:Image	db:Image
Irish	db:Irish_people	db:Irish_people, db:The_Pogues
band	db:Band_(rock_and_pop)	db:Band_(rock_and_pop), db:Musical_instrument, db:The_Pogues
The Pogues	db:The_Pogues	db:The_Pogues, db:Accordion, db:Irish_people, db:Musical_instrument, db:Band_(rock_and_pop)
wall	db:Wall	db:Wall

Table 1: Example of text linked to DBpedia, together with the text concepts activated through spreading activation. (Text: *The 2* ***shelves*** *hold a selection of* ***accordions*** *and other* ***instruments*** *for sale.* ***Pictures*** *of the* ***Irish band The Pogues*** *hang on the* ***wall.***). db: stands for the DBpedia namespace http://dbpedia.org/resource/

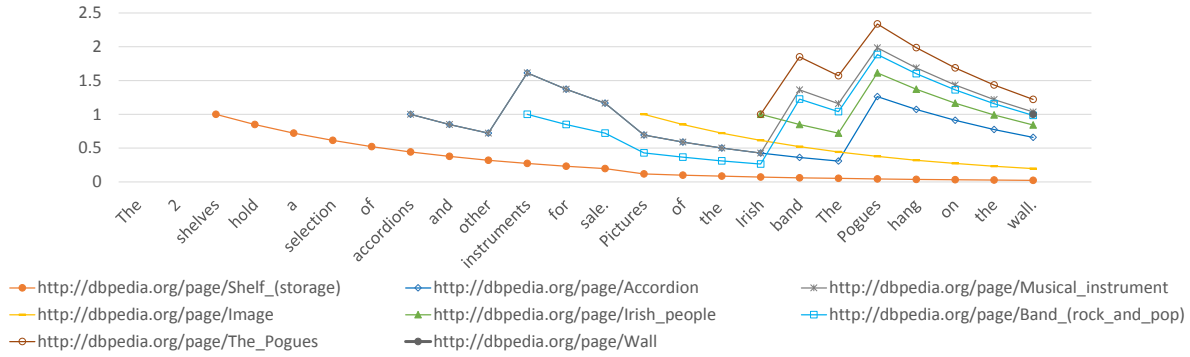

Figure 1: The change of CA in WM for the concepts in Table 1 as the text is sequentially traversed.

the index of its last word as $eos(s)$, we can define the sentence wrap-up activation ($AEoS$) of any concept c that is mentioned in s as in Formula 9. The paragraph wrap-up activation($AEoP$) is defined similarly.

$$\begin{aligned} AEoS_s(c) &= CA^{(eos(s))}(c); \\ AEoP_p(c) &= CA^{(eop(p))}(c); \end{aligned} \tag{9}$$

Therefore, each concept mention in the text produces three CA scores: activation at encounter (AE), activation at the end of the sentence it occurs in ($AEoS$), and activation at the end of the paragraph it occurs in ($AEoP$). Table 2 presents the scores of the defined CAs for the example in Table 1. Scores for AE are seen in Figure 1 on the word just before the target mention. Scores for $AEoS$ are seen on the last word of the corresponding sentence, and scores for $AEoP$ are seen at the end of the text.

For assessing the conceptual complexity of a given document D that has been linked to the knowledge base KG, resulting in m concept mentions, we propose to compute the activations of the mentioned concepts and take the inverse of their average as in Equation 10.

$$con\_comp(D) = \frac{m}{\sum_{i=1}^{m} activation(c_i)} \tag{10}$$

where c_i is the concept that mention i is linked to, and $activation(c_i)$ is a placeholder for any linear combination of the $AE(c_i)$, $AEoS(c_i)$ and $AEoP(c_i)$.

4 Experiments

4.1 Dataset

As ground truth, we use Newsela corpus which provides English news text on five complexity levels, the original story, and four manually simplified versions, gradually simplified by trained human editors under high quality control (Xu et al., 2015). As the target audience are children and second language learners, and texts are intended to maintain readers' interest, texts are not only simplified at a linguistic level but also at a cognitive level.

We report our experiments on 200 randomly sampled original texts from the English Newsela corpus, and for each of them, their four corresponding simplifications resulting in 1000 documents. All texts have been linked to DBpedia using KanDis (Hulpuş et al., 2015).

Mention	shelves	accordions	instruments	pictures	Irish	band	The Pogues	wall
Concept	db:Shelf_(storage)	db:Accordion	db:Musical_instrument	db:Image	db:Irish_people	db:Band_(rock_and_pop)	db:The_Pogues	db:Wall
AE	0	0	0.72	0	0	0.26	1.57	0
AEoS	0.20	1.17	1.17	0.20	0.84	0.98	1.21	1
AEoP	0.02	0.66	1.04	0.20	0.84	0.98	1.21	1

Table 2: AE, AEoS and AEoP scores for the mentions from the example in Table 1.

4.2 SA Model Settings

Settings of the output function. To explore our output function, we study how α (the graph decay) and β (the firing threshold) influence the performance of the models. We implemented models for α taking values from the set $\{0.25, 0.5, 0.75\}$ and β taking values from the set $\{0.0025, 0.005, 0.0075, 0.01\}$, excluding the $\alpha = 0.75$ and $\beta = 0.0025$ combination because the activated sub-graph becomes computationally too expensive.

Settings of the input function. To explore our input function, we implemented four accessibility computation schemes that we name according to the $exclusivity$ and $popularity$ factors (Excl-Pop) being used or not:

(No-No): $acc(o, s \xrightarrow{p} o) = acc(s, s \xrightarrow{p} o) = 1.0$;

(Yes-No): $acc(o, s \xrightarrow{p} o) = acc(s, s \xrightarrow{p} o) = excl(s \xrightarrow{p} o)$;

(No-Yes): $acc(o, s \xrightarrow{p} o) = pop(o)$ and $acc(s, s \xrightarrow{p} o) = pop(s)$;

(Yes-Yes): following Equation 2.

We transmit the intuition behind the input/output-function settings by reporting the average number of activated KG nodes per SA process over a sample of 1579 SA processes (triggered by 100 texts: 20 titles on all 5 levels). The results, shown in Table 3, indicate that exclusivity dramatically reduces the number of activated concepts. This is because exclusivity gives preference to less common relations, directing the activation in the graph to the few concepts that are strongly related to the seed. At the same time, the use of popularity in the absence of exclusivity has the opposite effect because popularity gives preference to the nodes with high degrees. When both exclusivity and popularity are used, only the high degree concepts that have very specific relations to the seed are activated.

With respect to the output function parameters, as expected, more concepts are activated as α decreases and as β increases.

Output function		**Input function settings (Excl-Pop)**			
β	α	**Yes-Yes**	**Yes-No**	**No-Yes**	**No-No**
0.0025	**0.5**	1,245	4,448	135,381	115,935
0.0025	**0.25**	1,106	2,977	95,142	75,491
0.005	**0.75**	1,003	3,428	108,086	90,082
0.005	**0.5**	1,002	2,576	85,895	68,318
0.005	**0.25**	979	2,190	51,190	34,921
0.0075	**0.75**	935	2,424	79,858	65,182
0.0075	**0.5**	917	2,111	60,840	45,175
0.0075	**0.25**	911	1,893	30,561	19,652
0.01	**0.75**	903	2,016	61,280	47,664
0.01	**0.5**	897	1,807	45,065	31,839
0.01	**0.25**	897	1,535	20,864	13,916

Table 3: Number of activated nodes in different SA settings.

4.3 WM Settings

We experimented with multiple definitions for the ϕ function, and report values for two definitions, ϕ^A and ϕ^1 as shown below:

$$\phi^A(SA(c)) = \begin{cases} SA(c) & \text{if} \quad SA(c) < 1.0 \\ pop(c) & \text{if} \quad SA(c) = 1.0 \end{cases}$$

$$\phi^1(SA(c)) = 1 \quad \text{if} \quad SA(c) > 0.0$$

ϕ^A uses the activations computed in the SA process, except for the seed concept where it uses its popularity score. This ensures that concepts mentioned in text become active in WM according to their popularity. ϕ^A is therefore sensitive to the actual SA scores, and to the popularity of mentioned concepts. On the contrary, ϕ^1 is only sensitive to changes in *the set* of activated concepts.

We investigated six parameter combinations for the values of the token, sentence and paragraph decay factors ($< \gamma_w, \gamma_s, \gamma_p >$):

no forgetting: <1, 1, 1>;

no paragraph transfer: <1, 1, 0> - there is no forgetting within a paragraph, but complete forgetting takes place between paragraphs;

no sentence transfer : <1, 0, 0> - there is no forgetting within a sentence, but complete forgetting takes place between sentences;

weak decay: <0.995, 0.9, 0.8> - the CA of a concept drops by one order of magnitude every 6 paragraphs of original texts (assuming

20 words per sentence, and 4 sentences per paragraph) and every 8 paragraphs for simple texts (assuming 12 words per sentence, 4 sentences per paragraph)[2];

medium decay: <0.85, 0.7, 0.5> - the CA decays one order of magnitude every 2 original sentences, and every 3 simple sentences respectively;

strong decay: <0.75, 0.5, 0.25> - the CA decays one order of magnitude with every original sentence, and with every 2 simple sentences.

Given the described SA and WM models, we implemented a total of 264 models in our framework.

4.4 State-of-the-art Measures

We compare our system to the graph-based metrics proposed by Štajner and Hulpuş (2018), as well as the baseline (ent/men) that computes the number of unique concepts per mention. For completeness, we briefly describe these features:

PageRank represents the average of the PageRank scores computed over the knowledge graph for all the mentioned concepts in the text;

PairDistSent and PairDistPar compute the average shortest path distance over the knowledge graph between all pairs of concepts mentioned in a sentence or paragraph respectively, averaged over all sentences or paragraphs respectively;

PairSemRelSent and PairSemRelPar are similar to the previous two measures, but instead of shortest path distances, they compute the exclusivity-based semantic relatedness (Hulpuş et al., 2015);

DensitySent and DensityPar compute the average density of the subgraphs extracted such that they connect all the pairs of concepts mentioned in the sentences or paragraphs respectively, by paths of at most 4 hops;

ConnCompSent and ConnCompPar are computed using the same subgraphs as those extracted in the previous measures, but computing the number of connected components averaged over sentences or paragraphs respectively.

All state-of-the-art features were computed over the same knowledge-graph (DBpedia) and using the same entity linker, KanDis (Hulpuş et al., 2015). Therefore, there are no biases stemming from the choice of those two in the comparison of our models with the state of the art.

4.5 Tasks and Evaluation Metrics

For each of our models we calculated 4 scores, by plugging into Equation 10 the values for AE, $AEoS$, $AEoP$, and All (the sum of previous three). Based on these scores, we test our models on two tasks: (i) Ranking five versions of the same news story according to their conceptual complexity; (ii) Identifying the conceptually simpler of two versions of the same news story.

In the ranking task, we compare our models' ranking of the five versions to the ground truth ranking, by computing their Kendall's Tau-b (Kendall, 1948), which calculates the difference between the number of concordant and discordant pairs, out of the number of all pairs, while also handling ties. Generally, Kendall Tau values range between -1 and 1, with 1 being obtained when there is perfect agreement between the compared rankings, -1 when one ranking is the reverse of the other, and 0 when the two compared rankings are independent. Hence for a random ranking we would expect Kendall Tau-b results close to 0.

In the second task, we calculate accuracy as the percentage of text pairs in which the simpler version was predicted as less conceptually complex by our models. In this task, the scores can range from 0 to 1, with a value of 0.5 for random picks.

5 Results and Discussion

We present our results starting with the WM settings because variations in these settings lead to the highest variations in the results.

5.1 Impact of WM Settings

Table 4 presents the average Kendall Tau-b scores for the six WM decay settings, four types of activation scores and two ϕ functions.

The first conclusion that stands out from this table is that there is a certain sweet spot when the WM decay is strong or medium, in which $AEoS$ performs substantially better than all other scores

[2]The numbers of 20 words per normal sentence and 12 words per simple sentence are taken from published statistics on the dataset we use (Xu et al., 2015).

WM decay	ϕ^1				ϕ^A			
	AE	**AEoS**	AEoP	All	AE	**AEoS**	AEoP	All
strong decay	-.15	.74	.34	.58	-.06	.76	.40	.64
medium decay	-.08	**.77**	.36	.60	.03	**.79**	.41	.66
weak decay	-.08	-.11	-.15	-.12	.16	.19	.10	.17
no forgetting	-.08	-.11	-.08	-.09	.16	.15	.19	.17
no paragraph transfer	.01	-.19	-.01	-.06	.19	.12	.20	.17
no sentence transfer	-.44	-.46	NA	NA	-.28	-.05	NA	NA

Table 4: Kendall Tau-b scores averaged over the 200 titles for all models with corresponding reading decay.

ϕ	α	β	exclusivity-popularity			
			y-y	y-n	n-y	n-n
ϕ^A	any	any	.80	.80	.79	.79
ϕ^1	0.50	0.0025	.81	.74	.70	.72
	0.25	0.0025	.81	.75	.72	.74
	0.75	0.005	.82	.76	.72	.73
	0.50	0.005	.82	.76	.74	.75
	0.25	0.005	.82	.76	.75	.76
	0.75	0.0075	.82	.77	.75	.76
	0.50	0.0075	.82	.77	.76	.76
	0.25	0.0075	.82	.78	.76	.76
	0.75	0.01	.82	.77	.76	.76
	0.50	0.01	.82	.78	.77	.76
	0.25	0.01	.82	.79	.77	.77

Table 5: Kendall Tau-b scores of the $AEoS$ measures computed with WM medium decay setting averaged over all 200 titles.

in all other settings. If the WM decay is either too strong (no sentence transfer) or too weak (no forgetting, weak decay and no paragraph transfer), all models perform poorly.

The second finding that is revealed by this table is that AE achieves very poor results across all WM settings. On the one hand, this indicates that our experiments are not able to confirm the forward semantic priming hypothesis. On the other hand, given the good results of $AEoS$, our experiments confirm the backwards priming hypothesis and sentence wrap-up.

5.2 Impact of the SA Settings

Table 5 shows the influence of the graph settings parameters in the ranking task. We focus on the best performing settings from Table 4, which measures $AEoS$ using WM medium decay.

Input function. Among all the SA settings, the definition of accessibility has the most influence. Our results show that the use of both exclusivity and popularity leads to $AEoS$ scores that best correlate with our ground truth complexity levels.

Output function. The choice of α and β parameters makes no noticeable difference for ϕ^A, while it makes a statistically significant difference[3] for ϕ^1. In the latter case, the best results are obtained when $\alpha = 0.25$ and $\beta = 0.01$, which corresponds to the setting which activates the smallest DBpedia subgraph (Table 3).

A somehow unexpected finding that has a great impact on SA parameter selection is that the bigger the activated DBpedia subgraph, the worse the results. This indicates that allowing the activation to spread through more of KG, might result in more noise. Consequently, controlling the flow of activation through relation and concept relevance scoring dramatically reduces the activated network, while improving the results.

5.3 Results on Pairwise Text Comparison

The pairwise comparison task provides insight on the models' ability to discriminate between two versions of the same news story. The results of the models with a medium WM decay and with the combination of α and β at the opposite sides of the proposed spectrum are shown in Table 6 for both tasks, together with the results of the state of the art and the baseline (ent/men).

The first observation is that our models distinguish almost perfectly between very complex and very simple versions of the same text (0−4, 1−4, 0−3). Also, generally they significantly outperform the baseline and state-of-the-art measures. However, our models perform close to random on distinguishing between the two most complex versions of the same title (0−1), the only setting in which they are outperformed by some state-of-the-art features and the baseline. Manual inspection indicates that the simplification that takes place between the two levels mostly involves sentence /paragraph splitting (syntactical simplification) which, as a side effect can have the decrease in the number of connected components, favouring ConnCompPar and ConnCompSent measures.

The results of the best model using ϕ^1 surpass the results of the best model using ϕ^A, particularly for the close level pairs (1−2, 2−3 and 3−4), which are generally harder to distinguish (paired t-test at 0.001 level of significance). This indicates that the fact that a concept is activated by SA is more relevant than the actual amount of activation, particularly for capturing subtle differences in texts.

[3] Statistically significant difference refers to a 0.001 level of significance using paired t-test, whenever mentioned.

Model					Level pairs										
Type	α	β	Exc.	Pop.	0-1	0-2	0-3	0-4	1-2	1-3	1-4	2-3	2-4	3-4	Kendall Tau-b
ϕ^A	any	any	yes	yes	.64	.88	.97	**1**	.88	**.97**	**.99**	.87	.95	.80	.80
ϕ^1	0.5	0.0025	no	no	.54	.83	.94	.95	.86	.93	.96	.83	.92	.82	.72
ϕ^1	0.25	0.01	no	no	.58	.84	.95	.99	.85	.94	.98	.86	.95	.86	.77
ϕ^1	0.5	0.0025	no	yes	.52	.83	.94	.95	.86	.92	.96	.83	.91	.82	.70
ϕ^1	0.25	0.01	no	yes	.58	.86	.95	.99	.87	.94	.98	.87	.93	.87	.77
ϕ^1	0.5	0.0025	yes	no	.54	.85	.94	.97	.87	.95	.98	.85	.91	.82	.74
ϕ^1	0.25	0.01	yes	no	.58	.88	.96	**1**	.86	.96	**.99**	.88	.95	.85	.79
ϕ^1	0.5	0.0025	yes	yes	.59	**.89**	.97	**1**	.90	**.97**	**.99**	.89	**.97**	.87	.81
ϕ^1	0.25	0.01	yes	yes	.61	**.89**	**.98**	**1**	**.92**	**.97**	**.99**	**.90**	**.97**	**.88**	**.82**
			ent/men		.67	.76	.83	.82	.71	.80	.79	.71	.72	.54	.47
			PageRank		.50	.53	.57	.57	.62	.58	.55	.53	.55	.57	.12
			PairDistSent		.58	.64	.65	.67	.58	.66	.64	.63	.59	.50	.23
			PairSemRelSent		.56	.62	.68	.77	.56	.69	.75	.63	.71	.55	.24
			DensitySent		.61	.69	.68	.72	.60	.63	.66	.51	.56	.58	.25
(Štajner and Hulpuş, 2018)			ConnCompSent		.67	.71	.83	.83	.60	.72	.74	.68	.73	.56	.41
			PairDistPar		.58	.70	.77	.84	.60	.76	.80	.67	.78	.60	.42
			PairSemRelPar		.60	.74	.87	.88	.70	.83	.88	.77	.83	.71	.56
			DensityPar		.59	.64	.57	.64	.57	.56	.62	.56	.62	.56	.19
			ConnCompPar		**.69**	.64	.74	.76	.61	.66	.62	.65	.62	.52	.22
			SeedDegree		.53	.51	.59	.55	.58	.55	.50	.53	.54	.58	.12

Table 6: Accuracies of the pairwise comparison task, and the Kendall Tau-b correlations for the AEoS scores of our models for medium WM decay, and for the state-of-the-art measures. Level 0 is the original text, while level 4 is the simplest version. *Any* signifies that the reported results were the same for all parameter choices.

6 Conclusion

We introduced a framework for tracking the conceptual complexity of texts during sequential reading, by mimicking human memory processes such as forward and backward semantic priming through spreading activation, sentence wrap-up and forgetting, and implemented a series of unsupervised models within it.

Our results confirmed the hypothesis that texts are simpler when the concepts therein are highly active at the end of their corresponding sentences. From the SA perspective, we showed that measures that account for relevance of relations and nodes make a significant impact, and that targeted search in the close proximity of the seeds performs best. Finally, our models strongly outperform the state-of-the-art measures in automatic assessment of conceptual complexity.

References

Bharat Ram Ambati, Siva Reddy, and Mark Steedman. 2016. Assessing relative sentence complexity using an incremental ccg parser. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1051–1057.

John Robert Anderson and Christian Lebiere. 1998. *The atomic components of thought*. Lawrence Erlbaum Associates.

Jonathan Anderson. 1981. Analysing the readability of English and non-English texts in the classroom with Lix. In *Proceedings of the Annual Meeting of the Australian Reading Association*.

Barbara Arfé, Lucia Mason, and Inmaculada Fajardo. 2017. Simplifying informational text structure for struggling readers. *Reading and Writing*.

Jerry Ball, Stuart Rodgers, and Kevin Gluck. 2004. Integrating act-r and cyc in a large-scale model of language comprehension for use in intelligent agents. In *AAAI Workshop*, pages 19–25.

Rebekah George Benjamin. 2012. Reconstructing Readability: Recent Developments and Recommendations in the Analysis of Text Difficulty. *Educational Psychology Review*, 24(1):63–88.

Michael R Berthold, Ulrik Brandes, Tobias Kötter, Martin Mader, Uwe Nagel, and Kilian Thiel. 2009. Pure spreading activation is pointless. In *The 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems-GIS'09*, pages 1915–1918.

Tim vor der Brück, Sven Hartrumpf, and Hermann Helbig. 2008. A readability checker with supervised learning using deep indicators. *Informatica*, 32(4):429–435.

Patrick Carroll and Maria L. Slowiaczek. 1986. Constraints on semantic priming in reading: A fixation time analysis. *Memory & Cognition*, 14(6):509–522.

Allan M. Collins and Elizabeth F. Loftus. 1975. A spreading activation theory of semantic processing. *Psychological Review*, 82:407–428.

Allan M. Collins and M. Ross Quillian. 1969. Retrieval time from semantic memory. *Journal of Verbal Learning and Verbal Behavior*, 8(2):240 – 247.

Fabio Crestani. 1997. Application of Spreading Activation Techniques in Information Retrieval. *Artificial Intelligence Review*, pages 453–482.

Arthur C. Graesser, Danielle S. McNamara, Max M. Louwerse, and Zhiqiang Cai. 2004. Coh-Metrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments, & Computers*, 36(2):193–202.

Tanja Gulan and Pavle Valerjev. 2010. Semantic and related types of priming as a context in word recognition. *Review of psychology*, 17(1):53–58.

Benjamin Heitmann and Conor Hayes. 2016. Semstim: Exploiting knowledge graphs for cross-domain recommendation. In *2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW)*, pages 999–1006.

Ioana Hulpuş, Narumol Prangnawarat, and Conor Hayes. 2015. Path-based semantic relatedness on linked data and its use to word and entity disambiguation. In *The Semantic Web - ISWC 2015*, pages 442–457, Cham. Springer International Publishing.

Marcel Adam Just and Patricia A. Carpenter. 1980. A theory of reading:from eye fixations to comprehension. *Psychological Review*, 87(4).

Slava Kalyuga, Paul Ayres, Paul Chandler, and John Sweller. 2003. The expertise reversal effect. *Journal of Educational Psychology*, 38:23–31.

Maurice G. Kendall. 1948. *Rank correlation methods*. Griffin, London.

Marta Kutas and Steven A. Hillyard. 1984. Brain potentials during reading reflect word expectancy and semantic association. *Nature*, 307(161).

Antonio Lieto, Christian Lebiere, and Alessandro Oltramari. 2018. The knowledge level in cognitive architectures: Current limitations and possible developments. *Cognitive Systems Research*, 48:39 – 55. Cognitive Architectures for Artificial Minds.

Vivi Nastase. 2008. Topic-driven multi-document summarization with encyclopedic knowledge and spreading activation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '08, pages 763–772, Stroudsburg, PA, USA. Association for Computational Linguistics.

James H. Neely. 1991. Semantic priming effects in visual word recognition: A selective review of current findings and theories. In D. Besner and G. W. Humphreys, editors, *Basic processes in reading: Visual word recognition*, pages 265–335. Lawrence Erlbaum Associates, Hillsdale.

Alessandro Oltramari and Christian Lebiere. 2012. Pursuing artificial general intelligence by leveraging the knowledge capabilities of act-r. In *Artificial General Intelligence*, pages 199–208, Berlin, Heidelberg. Springer Berlin Heidelberg.

Dario D. Salvucci. 2014. Endowing a cognitive architecture with world knowledge. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 36.

Sanja Štajner and Ioana Hulpuş. 2018. Automatic assessment of conceptual text complexity using knowledge graphs. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 318–330. Association for Computational Linguistics.

Zareen Syed, Tim Finin, and Anupam Joshi. 2008. Wikipedia as an ontology for describing documents. In *Proceedings of the Second International Conference on Weblogs and Social Media*. AAAI Press.

Sowmya Vajjala and Detmar Meurers. 2014. Assessing the relative reading level of sentence pairs for text simplification. In *Proceedings of the EACL 2014*, pages 288–297.

Wei Xu, Chris Callison-Burch, and Courtney Napoles. 2015. Problems in Current Text Simplification Research: New Data Can Help. *Transactions of the Associaton for Computational Linguistics*, 3:283–297.

Ganggao Zhu and Carlos. A. Iglesias. 2017. Computing semantic similarity of concepts in knowledge graphs. *IEEE Transactions on Knowledge and Data Engineering*, 29(1):72–85.