Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 1414–1422
Marseille, 11–16 May 2020

When Shallow is Good Enough: Automatic Assessment of Conceptual Text Complexity using Shallow Semantic Features

Sanja Štajner[1] and Ioana Hulpuş[2]
[1]ReadableAI, Cologne, Germany
[2]Data and Web Science Group, University of Mannheim, Germany
sanja.stajner@readableai.com, ioana@informatik.uni-mannheim.de

Abstract

According to psycholinguistic studies, the complexity of concepts used in a text and the relations between mentioned concepts play the most important role in text understanding and maintaining reader's interest. However, the classical approaches to automatic assessment of text complexity, and their commercial applications, take into consideration mainly syntactic and lexical complexity. Recently, we introduced the task of automatic assessment of conceptual text complexity, proposing a set of graph-based deep semantic features using DBpedia as a proxy to human knowledge. Given that such graphs can be noisy, incomplete, and computationally expensive to deal with, in this paper, we propose the use of textual features and shallow semantic features that only require entity linking. We compare the results obtained with new features with those of the state-of-the-art deep semantic features on two tasks: (1) pairwise comparison of two versions of the same text; and (2) five-level classification of texts. We find that the shallow features achieve state-of-the-art results on both tasks, significantly outperforming performances of the deep semantic features on the five-level classification task. Interestingly, the combination of the shallow and deep semantic features lead to a significant improvement of the performances on that task.

Keywords: conceptual complexity, semantics, text classification

1. Introduction

In this digitalized era, readers are overwhelmed with the amount of informational texts freely available online. Although seemingly so easy to reach, many of them are too complex for an average reader, and even more for struggling readers, e.g. non-native speakers, people with low literacy, or people with any kind of reading or cognitive impairments (such as dyslexia, aphasia, autism, Down's syndrome, etc.). To mitigate this problem, many guidelines have been proposed and made available advising on how to make texts easier to understand for everyone, regarding both their presentation (in terms of font styles, sizes, and colors, background colors, etc.) and their textual characteristics (vocabulary, syntactic structures, and overall text structures), e.g. Web Content Accessibility Guidelines (WCAG)[1], 'Make it Simple' guidelines (Freyhoff et al., 1998), Federal Plain Language Guidelines[2]. At the same time, many readability metrics have been proposed to manually or automatically assess the reading level necessary to understand a text, mostly focusing on superficial lexical and syntactic features, and rarely on more semantic-based features (DuBay, 2004). Most of the proposed readability metrics do not assess deeper levels of text processing, such as inference making and the use of world knowledge and discourse structure necessary to comprehend the text (Arfé et al., 2017).

According to the model of reading comprehension proposed by Kintsch and van Dijk (1978), the reader needs to understand both individual propositions and concepts in the text, as well as their relations, in order to make a coherent story and fully understand the text. Text difficulty could thus be seen as the amount of gaps in the text coherence, and the effort required by the reader to repair them by inference making (Arfé et al., 2017). This effort for mentally organizing the content of a text is especially challenging for struggling readers (Lovett et al., 1996; Meyer et al., 1980), and those readers that have problems with memory load (Meyer, 2003). Even in the case of non-struggling readers, to actively engage them, texts need to be on the right level not only by their lexical and syntactic choices, but also in the way the concepts are mutually connected, and by the amount of background knowledge necessary to make inferences which strengthen the representation of the text meaning (McNamara et al., 2006).

Building up on those psycholinguistic models of text comprehension, we earlier proposed the task of automatically assessing conceptual complexity of texts and explored the possibility of using DBpedia (DBpedia, 2014) to construct knowledge graphs with the aim of measuring the amount of background knowledge necessary to understand how are the mentioned concepts interconnected (Štajner and Hulpuş, 2018).

In this work, we propose two sets of features to automatically measure conceptual text complexity, which are computationally less expensive, but seem to have similar or even better predictive power than previously proposed graph-based deep semantic features (Štajner and Hulpuş, 2018). We propose several surface text-based features and shallow semantic features, both types only requiring an entity linker, and some basic entity mention counting and offset calculations. Furthermore, we set a strong baseline by using the four widely-known psycholinguistic features, which have not been used in automatic assessment of text complexity so far. We compare the performance of the newly proposed features with those of the state-of-the-art deep (graph-based) semantic features (Štajner and Hulpuş, 2018), and the strong baseline features, on two tasks: (1) bi-

[1] `https://www.w3.org/TR/WCAG21/`
[2] `https://www.plainlanguage.gov/guidelines/`

nary text classification, and (2) five-level text classification. The newly proposed features outperform both the strong baseline features, and the state-of-the-art features, on both tasks. Interestingly, the combination of shallow and deep semantic features seem to lead to best performances on the five-level classification task.

2. Related Work

Manual assessment of textual complexity has long history in education and psycholinguistic research. Only in the second half of the twentieth century, over 200 readability formulae have been proposed with the aim of assessing the reading level necessary to understand given texts (DuBay, 2004). The majority of them only captures lexical and syntactic properties of a text. Those that go beyond those two levels, and attempt at capturing semantic complexity of a text, cannot be computed automatically with sufficient precision.

Currently, the only automatic complexity assessment tool that goes beyond lexical and syntactic levels is the Coh-Metrix tool (Graesser et al., 2004), which apart from a high number of lexical and syntactic features, also calculates coreference indices and LSA semantic complexity features. It has been shown that Coh-Metrix can successfully capture cohesion level of manually produced texts, but has problems differentiating between texts with different cohesion levels which are closely related by their topic (McNamara et al., 2006). Furthermore, Coh-Metrix does not have any features related to conceptual clarity, which would measure ambiguity, vagueness, and abstractness of a concept, necessary to measure conceptual complexity of a text, as defined in the previous section.

Several works investigated the correlation of a number of syntactic, lexical, conceptual, and discourse features with text readability, e.g. (Pitler and Nenkova, 2008; Štajner et al., 2012). Pitler and Nenkova (2008) explored six groups of features: baseline (e.g. average number of words per sentence or average number of characters per sentence), vocabulary (e.g. article likelihood according to different language models), syntactic (e.g. average parse tree height, average number of subordinate clauses per sentence, etc.), lexical cohesion (e.g. the number of pronouns and the number of definite articles per sentence, word overlap over nouns and pronouns), entity coherence features based on entity grids (Barzilay and Lapata, 2008), and discourse relation features (e.g. implicit comparisons, explicit temporal relations, etc.). Out of those six groups of features, discourse relation features obtained the highest correlation with human ratings of readability on the Penn Discourse Treebank (Marcus et al., 1993). The major obstacle of using those features in a system for automatic assessment of text readability/complexity is that no robust systems for automatic annotation of discourse exist, and those features can thus be extracted only from texts annotated in Penn Discourse Treebank, as the authors themselves pointed out (Pitler and Nenkova, 2008). In the binary text classification experiments ("given two articles, is article 1 more readable than article 2?"), however, the entity grid features performed the best, reaching the accuracy of 0.79 (Pitler and Nenkova, 2008). Štajner et al. (2012) investigated the correlation between several widely-used readability formulae and several features, indicators of structural complexity and ambiguity in meaning (average number of pronouns, definite descriptions, and word senses) across four different corpora (Simple Wikipedia, news articles, fictional stories, and health leaflets). They found a high level of correlation (Pearson's correlation) between those features and the Flesch readability index (Flesch, 1949) across all corpora. They, however, did not attempt at automatically assessing textual complexity using those features.

The coherence assessment, as a measure of text quality for automatic generation of texts has received significant attention in the last ten years in the field of automatic text generation, e.g. (Barzilay and Lapata, 2008; Karamanis et al., 2009; Mesgar and Strube, 2018). While coherence assessment is somehow related with our task of automatic assessment of conceptual text complexity, it still has several major differences. First, the focus of coherence assessment in those works is on local text coherence (only within a sentence, or between two neighbouring sentences), while we focus on overall text coherence, capturing entity-relations throughout the whole text. Second, we focus only on conceptual text complexity and thus make sure that we do not introduce any bias coming from syntactic or lexical complexity (see Section 3 for details on feature extraction). The aforementioned approaches to coherence assessment, in contrast, use syntactic information, and therefore model coherence assessment without controlling for the syntactic complexity. Finally, we focus on conceptual text complexity from the readers perspective and are therefore focused on measuring background knowledge necessary to understand the text. The aforementioned approaches to coherence assessment, instead, focus on quality of generated text and are thus not interested in the level of background knowledge, but rather only the connection between the entity mentions (not their number or familiarity).

To the best of our knowledge, we were the first to define the task of automatic assessment of conceptual complexity as the level of background knowledge necessary to understand the text (Štajner and Hulpuş, 2018). In the same work, we explored the possibility to use a set of graph-based deep semantic features computed over DBpedia as a proxy of the background knowledge, and approached the problem as a supervised binary classification task on the Newsela corpus (see Section 4 for more details), which contains news articles manually simplified not only on lexical and syntactic level, but also on conceptual level under a strict quality control (Newsela, 2016). Later, we proposed an unsupervised approach for the same task, using spreading activation over the subgraphs of DBpedia as the text is sequentially traversed (Hulpuș et al., 2019).

To the best of our knowledge, there are no other datasets that contain same stories manually adapted to different complexity levels (including conceptual complexity), except the aforementioned Newsela corpus. Building such datasets is a challenging task, as asking humans to rate text coherence has been shown to be unreliable because of several possible confounds such as the annotator interest, level of distraction, and familiarity with the topic (Lapata, 2006). To avoid introducing such confounds in our systems, we opt

Type	#	Description/Name
Surface (text-based)	1	Number of mentions divided by the total number of tokens
	2	Average number of mentions divided by the total number of tokens (in a paragraph)
	3	Average number of mentions divided by the total number of tokens (in a sentence)
	4	Average distance (in sentences) between consecutive mentions
	5	Average distance (in paragraphs) between consecutive mentions
Shallow Semantics	6	Number of unique entities
	7	Average number of unique entities in a paragraph
	8	Average number of unique entities in a sentence
	9	Entities to mentions ratio in the document
	10	Average entities to mentions ratio in a paragraph
	11	Average entities to mentions ratio in a sentence
	12	Average distance (in sentences) between consecutive mentions of the same entity
	13	Average distance (in paragraphs) between consecutive mentions of the same entity
	14	Number of cases with the maximal distance (in sentences) between consecutive mentions of the same entity
	15	Number of cases with the maximal distance (in paragraphs) between consecutive mentions of the same entity
	16	Number of cases with the minimal distance (in sentences) between consecutive mentions of the same entity
	17	Number of cases with the minimal distance (in paragraphs) between consecutive mentions of the same entity
	18	Average distance (in sentences) between all pairs of mentions of the same entity
	19	Average distance (in paragraphs) between all pairs of mentions of the same entity

Table 1: Conceptual complexity features

for already existing readability corpora made under highest quality standards, the Newsela corpus. By using the same corpus as in our previous work (Štajner and Hulpuş, 2018), we are able to more fairly compare the performances of three different types of features on the same task, using the same training and test sets.

3. Entity-based Conceptual Complexity Features

We propose two sets of features: the surface text-based features, and the shallow semantic features. Table 1 contains the list of all 19 features, computed on a document/text level. All features based on counting entities/mentions are computed in three versions: on a document level, and as the average on the paragraph and sentence levels. All features that compute the distance between mentions are calculated in two versions: as an offset in sentences, and as an offset in paragraphs. To avoid the influence of syntactic features (the length of a document, paragraph, or a sentence) which would make all classification tasks trivial (as simplification performed in Newsela, like in any other manually simplified corpora, shortens texts, paragraphs, and sentences in each simplification step), all surface and shallow features that include counting are normalized with the total number of tokens in the corresponding text snippet.

Figure 1 illustrates the type of data used for the computation of the two newly proposed types of features (the surface text-based features and shallow semantic features) as opposed to the graph-based deep semantic features proposed earlier (Štajner and Hulpuş, 2018). The main prerequisite for the extraction of any feature is that the document is linked to the knowledge graph, in this case DBpedia. Given the linked document, the mentions are used for computing text based features. The mentions, as well as the knowledge graph entities they are linked to, are used for computing shallow semantics features.

3.1. Surface Text-based Features

This group of features aims to capture conceptual complexity expressed in the surface structure of the discourse. It assesses the general memory load required to process all mentions of encyclopedic concepts in the text, as well as the density of such mentions in the text. For computing the textual features, the output of the entity linking step is used to select the textual noun phrases that correspond to encyclopedic concepts - the mentions.

Given the example text snippet in Figure 2, the values for the fourth and fifth surface features (average distance between consecutive mentions) are 0.4 for the distance in sentences and 0.2 for the distance in paragraphs.

3.2. Shallow Semantics Features

In this set of features, we focus on shallow semantic-based features capturing the amount of background knowledge necessary to understand the text (measured as the number of unique entities/concepts) and the distance of the mentions of the same entities in the text. The core hypothesis is that the closer the new mention is to the previous mention of the same concept, the less effort is necessary to reactivate the concept in the memory.

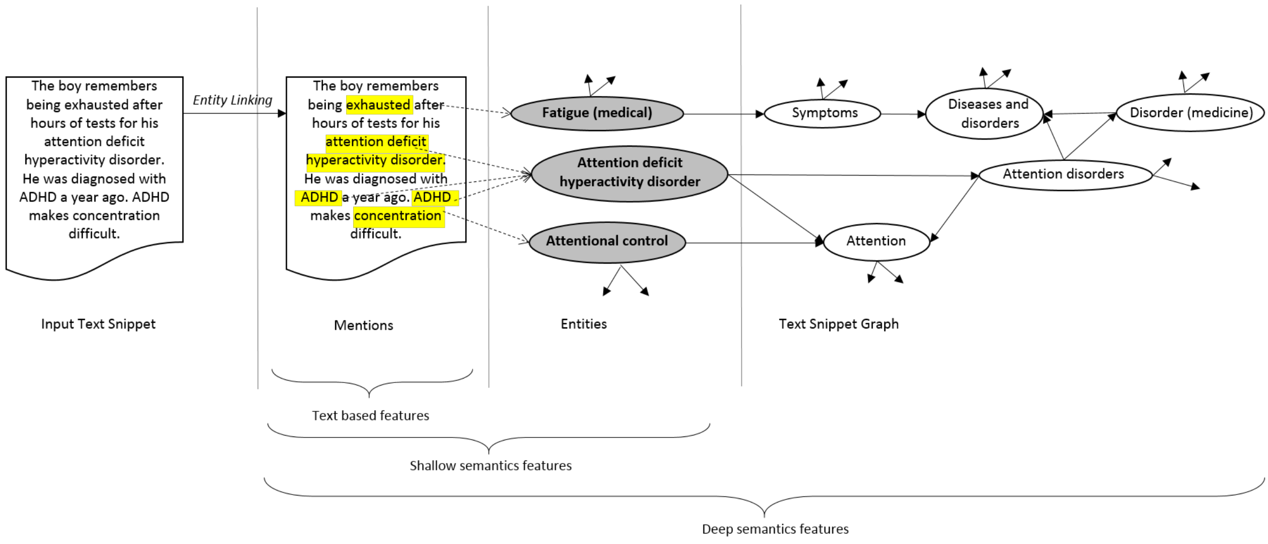
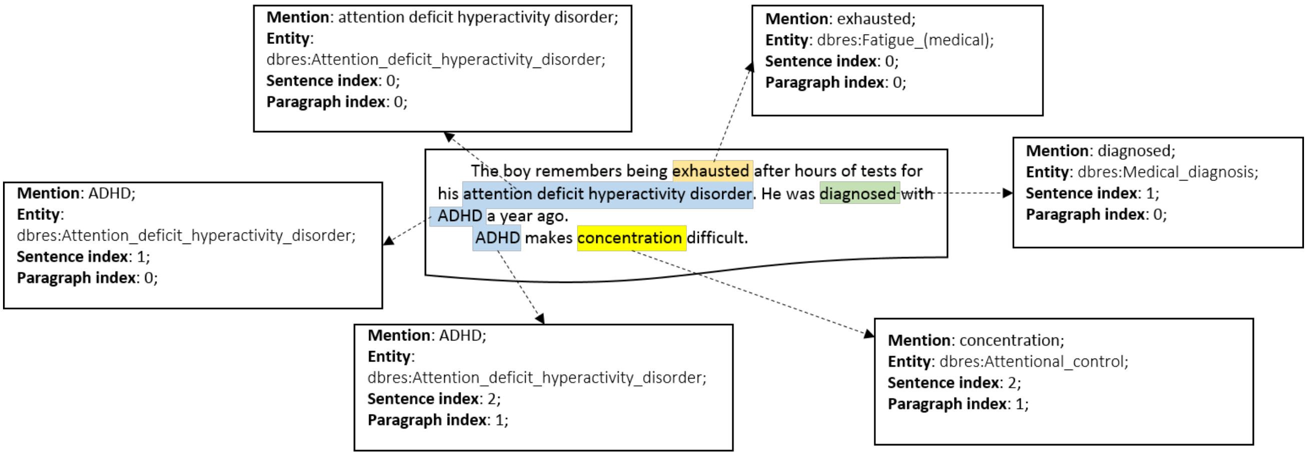

The boy remembers being exhausted after hours of tests for his attention deficit hyperactivity disorder. He was diagnosed with ADHD a year ago. ADHD makes concentration difficult.
Entity Linking
The boy remembers being exhausted after hours of tests for his attention deficit hyperactivity disorder. He was diagnosed with ADHD a year ago. ADHD makes concentration difficult.
Fatigue (medical)
Attention deficit hyperactivity disorder
Attentional control
Symptoms
Diseases and disorders
Disorder (medicine)
Attention disorders
Attention
Input Text Snippet
Mentions
Entities
Text Snippet Graph
Text based features
Shallow semantics features
Deep semantics features

Figure 1: Document processing and feature extraction pipeline


The boy remembers being exhausted after hours of tests for his attention deficit hyperactivity disorder. He was diagnosed with ADHD a year ago.
ADHD makes concentration difficult.

Mention: attention deficit hyperactivity disorder;
Entity:
dbres:Attention_deficit_hyperactivity_disorder;
Sentence index: 0;
Paragraph index: 0;

Mention: exhausted;
Entity: dbres:Fatigue_(medical);
Sentence index: 0;
Paragraph index: 0;

Mention: ADHD;
Entity:
dbres:Attention_deficit_hyperactivity_disorder;
Sentence index: 1;
Paragraph index: 0;

Mention: diagnosed;
Entity: dbres:Medical_diagnosis;
Sentence index: 1;
Paragraph index: 0;

Mention: ADHD;
Entity:
dbres:Attention_deficit_hyperactivity_disorder;
Sentence index: 2;
Paragraph index: 1;

Mention: concentration;
Entity: dbres:Attentional_control;
Sentence index: 2;
Paragraph index: 1;

Figure 2: An example of a processed text snippet.

The shallow semantic features are computed using the DBpedia concepts linked to the noun phrases from the text. As illustrated in Figure 1, the three mentions *"attention deficit hyperactivity disorder"*, *"ADHD"*, and *"ADHD"* are all linked to the same entity: <dbres:Attention_deficit_hyperactivity_disorder>

To exemplify, the values of the *entities to mentions ratio* feature in the text snippet illustrated in Figure 2 are: 1 in all three sentences and in the second paragraph, and 0.75 in the first paragraph. For the same example, the value of the *average distance between two consecutive mentions of the same entity*, when measured as a distance in sentences, is 1 ($= \frac{1+1}{2}$), and when measured as a distance in paragraphs, is 0.5 ($= \frac{0+1}{2}$).[3] The feature *average distance between all mentions of the same entity* considers the distance between such mentions even if they are not consecutive. In our example (Figure 2), the values for this feature are 1.33 ($= \frac{1+1+2}{3}$) for the distance in sentences and 0.66 ($= \frac{0+1+1}{3}$) for the distance in paragraphs.

[3] The entities that are mentioned only once are ignored in the computation of those features.

4. Experimental Setup

In this section, we give details on the corpora used in all experiments, the entity linking step, and the setup of classification experiments used to test whether surface text-based features and shallow semantic features can lead to similar performances as the previously proposed deep semantic features computed over the DBpedia subgraphs (Štajner and Hulpuş, 2018) listed in Table 2.

4.1. Newsela Corpus

We are interested in finding features which can distinguish between different versions of the same texts. Therefore, we perform our experiments on the English part of Newsela corpus, which contains original news articles, manually simplified at four different complexity levels by trained human editors under a high quality control (Xu et al., 2015). As those are informational texts and simplification is made for language learners and young readers, with the idea of engaging them and maintaining their interest in the texts they read, we can assume that texts are on different conceptual complexity levels, requiring different amounts of background knowledge and inferences in order to comprehend

Type	#	Description/Name
Single node	1	Node degree
	2	Node clustering coefficient
	3	Node PageRank
Pairwise	4	Average shortest path (in the graph) between concepts occurring in the same sentence
	5	Average shortest path (in the graph) between concepts occurring in the same paragraph
	6	Average pairwise semantic relatedness between concepts occurring in the same sentence
	7	Average pairwise semantic relatedness between concepts occurring in the same paragraph
Global	8	The average number of connected components per sentence
	9	Average number of connected components per paragraph
	10	Average local clustering coefficient per sentence
	11	Average local clustering coefficient per paragraph
	12	Average graph density per sentence
	13	Average graph density per paragraph

Table 2: Previously proposed graph-based deep semantic features (Štajner and Hulpuş, 2018).

the texts.

We are not aware of any other publicly available corpora, large enough to allow for machine learning experiments, with different versions of the same text that are on different conceptual complexity levels. Asking human annotators to annotate conceptual complexity level is a hard task which would very likely lead to low inter-annotator scores, knowing that even assessing simplicity of a sentence on a 1–5 level scale leads to low IAA scores (Štajner, 2018). Furthermore, as mentioned in Section 2, asking human annotators to annotate text coherence is known to bring several possible confounds into play. By using the Newsela corpus, we avoid those drawbacks and rely instead on trained human annotators and proven readers engagement to provide us with the 'gold standard' data for our experiments.

4.2. Entity Linking

The entity linking step links noun phrases of the input document to entities in DBpedia. To link the entities/concepts (common nouns and named entities) to DBpedia, we use KanDis (Hulpuş et al., 2015), which showed comparatively good results at linking both types of entities (recall between 0.59 and 0.69 at a disambiguation accuracy between 0.88 and 0.89 on news items). To minimize the effect of wrong linking, we remove the outliers, i.e. entities that have very weak semantic relatedness (Hulpuş et al., 2015) to other entities in the text. This strategy should eliminate some of the wrongly linked concepts and corner cases. All 19 newly proposed features are computed on the linked documents.

4.3. Tasks

We test our features on two tasks:

- Task 1: Pairwise comparison of conceptual complexity between two versions of the same text;
- Task 2: Five-level conceptual complexity assessment of texts.

For both tasks, we take 200 original English news articles from Newsela (Newsela, 2016) and their four corresponding simpler versions (marked for their level of complexity on a 0–4 scale, where 0 corresponds to the original texts, and 4 to the simplest version).

4.3.1. Task 1 (Binary Classification)

The goal of the first task is to, given the two versions of the same news story, decide on whether the first story is conceptually *simpler* or *more complex* than the second story.

To build the dataset, we take each possible pair of two versions of the same texts, extract all features from each of them, concatenate those two vectors of features, and assign the label *simpler* or *more complex* depending on whether the first text is simpler or more complex than the other, according to their Newsela complexity levels ('gold' label). In each pair, we choose the order of texts randomly, ensuring that we have approximately equal number of instances in both classes. This resulted in having 980 instances with the class 'simpler' and 1020 instances with the class 'more complex'.

We perform all classification experiments in a ten-fold cross-validation setup with ten repetitions in Weka Experimenter (Hall et al., 2009), using four classification algorithms: Logistic (le Cessie and van Houwelingen, 1992), Support Vector Machines with feature normalization (Keerthi et al., 2001), JRip rule-learner (Cohen, 1995), and Random Forest (Breiman, 2001). In all four cases, we use the algorithms with their default parameters as our goal is not to achieve the optimal performances on the task, but rather to assess the quality of our features and their combinations on the task.

4.3.2. Task 2 (Five-Level Classification)

The goal of the second task is to assess the conceptual complexity of a given text (news story) on a 0–4 level scale that corresponds to Newsela levels 0–4.

We use all 1,000 articles (200 titles on five complexity levels) as our dataset, and apply ten-fold cross-validation setup with ten repetitions (in Weka Experimenter) without controlling for which titles and text versions end up in which fold. In other words, we behave as we had 1,000 independent articles on five different complexity levels, but choose to have 200 titles on five complexity levels, in order to:

(1) allow for learning subtle differences between different complexity levels of texts treating the same topic; and (2) avoid bias that might arise from topic differences if we used corpora with five complexity levels, in which each level has different topics (as in the case of typical language learners corpora).

4.4. Baseline

As a strong baseline, we use a set of four well-known psycholinguistic features: familiarity, age of acquisition, concreteness, and imagery (Gilhooly and Logie, 1980). Familiarity of a word can be defined as the frequency with which a word is seen, heard, or used daily. Age of acquisition is the average age at which a word is learned. Concreteness of a word measures how "palpable" the object the words refers to is. Imagery can be defined as the intensity with which a word arouses images (Paetzold and Specia, 2016). According to those definitions, those four features seem relevant for our task of measuring conceptual complexity of text. It has been shown that those objects whose names are learned earlier in life (*age of acquisition*) can be named faster in later stages in life (Carroll and White, 1973). As such, they might be a good indicator of how quickly a specific concept can be retrieved in working memory. The *familiarity* and *concreteness* of the words has been shown to have impact on text comprehension (Paetzold and Specia, 2016). *Imagery* can be seen as a measure of word's abstractness, and as such, be an useful feature in assessing conceptual complexity of texts.

Despite their wide use in psycholinguistics, those features have never been used in automatic readability assessment. This is probably due to scarcity of resources and their very poor coverage on regular texts. The original MRC psycholinguistic database[4] (Coltheart, 1981) contains familiarity scores for 9,392 words, age-of-aquisition scores for 3,503 words, concreteness scores for 8,228 words, and imagery scores for 9,240 words (Paetzold and Specia, 2016), and much less words with all four scores. The bootstrapped version of this database (Paetzold and Specia, 2016), in contrast, contains all four scores for 85,942 words. We tested the coverage of both versions of the MRC database, and found that the original database covers less than 5% of words in our texts on average, while the bootstrapped version covers 63.5% words in our texts on average. Therefore, we used the bootstrapped version of the MRC database to extract the set of four aforementioned psycholinguistic features (familiarity, age of acquisition, concreteness, and imagery) and use it as a strong baseline in our classification tasks.

5. Binary Classification Results

The results of the pairwise comparison experiments are presented in Table 3. The shallow semantic features on their own achieve the best results (up to 0.94 weighted F-measure), significantly better than the other two types of features (*surface* and *deep*) regardless of the classification algorithm. The superior performance of shallow features over the surface features is not surprising. The shallow semantic features introduce a semantic level on top of the superficial textual level, which on its own might be too simplistic to capture the overall conceptual complexity. The superior performance of shallow features over the deep semantic features might seem surprising. One explanation could be that the graph-based features, although theoretically being well-suited (Štajner and Hulpuş, 2018), introduce some noise from the DBpedia knowledge graph. The other explanation could be that the graph-based features are not expressive enough to capture the full conceptual complexity, as they do not account for all available positional attributes. The combination of shallow and deep features, and the combination of all three feature sets do not seem to bring any significant improvements over the results achieved by the shallow features on their own.

Features	Logistic	SVM-n	JRIP	RandF
Surface	.80	.80	.72	.80
Shallow	.94	.93	.80	.91
Surface+Shallow	**.95**	.94	.81	.92
Deep	.84	.84	.78	.85
Surface+Deep	.90	.88	.79	.87
Shallow+Deep	**.96**	.95	.82	.92
All	**.96**	**.96**	.82	.93
MRC (strong baseline)	.94	.91	.89	.92

Table 3: The weighted average F-measure (the accuracy results are the same) for the binary classification task (10-fold cross-validation with 10 repetitions). The best results for each algorithm are presented in bold. Standard deviations for the weighted F-measure were in range 0.01–0.03. The majority class baseline has the accuracy of 0.51.

Interestingly, the four psycholinguistic features (familiarity, age of acquisition, concreteness, and imagery) extracted from the bootstrapped MRC database, indeed proved to be a strong baseline for this binary task. At the same time, in the five-level classification task, they performed significantly worse than any other set of features (see Section 6). This can be explained by the fact that in the binary classification task the system needs to decide which of the two versions of the same text is simpler. When constrained by the topic, while manually simplifying texts, human editors naturally try to simplify them lexically, thus choosing more frequent and more familiar words. In the five-level classification task, the system needs to assess the 'absolute' conceptual complexity level of a given text, instead of choosing the simpler of the two versions of the same text. While human editors in each simplification step choose more frequent and more familiar words, the starting point is different for each original article, thus introducing the noise into the five-level classification task and making it more challenging than comparing the two versions of the same article.

[4] `http://websites.psychology.uwa.edu.au/school/MRCDatabase/uwa_mrc.htm`

Algorithm	Logistic		SVM-n		JRip		RandF	
	F	AUC	F	AUC	F	AUC	F	AUC
Surface	.37	.72	.30	.65	.30	.56	.31	.65
Shallow	**.51**	**.83**	.47	.78	.27	.63	.42	.76
Surface+Shallow	**.51**	**.84**	.48	.79	.28	.64	.44	.77
Deep	.34	.80	.33	.67	.33	.58	.31	.66
Surface+Deep	.44	.78	.40	.72	.23	.60	.34	.70
Shallow+Deep	.58	.87	53	.82	.31	.67	.44	.79
All	**.59**	**.88**	.53	.82	.31	.67	.45	.79
MRC (strong baseline)	.31	.68	.27	.63	.25	.57	.30	.64

Table 4: The weighted average F-measure (F) and the weighted average area under curve (AUC) for the five-level classification task (10-fold cross-validation with 10 repetitions). The best results for each algorithm and each evaluation metric are presented in bold. Standard deviations for weighted F-measure were in range 0.04–0.06, and for AUC in range 0.02–0.04. The majority baseline achieves a 0.07 weighted F-measure and a 0.50 AUC.

5.1. Indirect Comparisons to Other Approaches

In this set of experiments we use the same 200 randomly selected Newsela articles (and their corresponding simplified versions) that were used in the experiments with the state-of-the-art unsupervised system for measuring conceptual complexity (Hulpuș et al., 2019). As our current approach is supervised, direct comparison is not possible. Therefore, we compare them indirectly, assuming that the average accuracy over 2000 article pairs from the unsupervised approach roughly corresponds to the average accuracy in a 10-fold cross-validation with 10 repetitions over the same 2000 article pairs. If we assume such approximation, the combination of our surface and shallow features outperforms the unsupervised system (.95 to .91).

Due to different datasets, we cannot directly compare the performances of our features with those proposed by Pitler and Nenkova (Pitler and Nenkova, 2008). However, as we use the same classification algorithm and framework (SVM classifier in Weka), and the accuracy of the majority class baseline is similar in both cases (.50 to .51), we can roughly compare the accuracy of our system (the combination of surface and shallow features) with their best set of features (entity-grid features): .94 to .79. Even their full feature set achieves an accuracy of only .89. However, as mentioned earlier (Section 2), their full feature set cannot be automatically extracted as it relies on having an extensive human annotation such as the one of the Penn Treebank.

6. Five-Level Classification Results

On the five-level classification task, the logistic regression significantly outperformed the other three classification algorithms (paired t-test; $p < 0.05$). On this task, not only the shallow semantic features significantly outperform the deep semantic features, but even the surface features do so. In contrast to the case of the binary classification task (Section 5), here the combination of shallow and deep semantic features, and the combination of all three feature sets, achieve significantly better results than the shallow features, or the combination of surface and shallow semantic features. The combination of all three feature sets reaches an average weighted F-measure of 0.59 and AUC of 0.88 (Table 4). More importantly, the confusion is much more frequent between neighboring classes (Newsela levels) than between distant classes (Table 5). In most of the runs, there was not a single misclassification between the Newsela levels 0 and 4.

Correct level	Predicted level				
	0	1	2	3	4
0	157	27	13	3	0
1	33	110	54	2	1
2	14	52	85	43	6
3	4	4	46	99	47
4	0	1	2	56	141

Table 5: The confusion matrix for one of the 100 runs of logistic function on the full feature set (the confusion matrices in other runs follow a similar pattern).

Regardless of the classification algorithm, performances on different feature sets follow the same pattern, which we already saw in the pairwise text comparison task. Out of the three sets of features (text, shallow, and deep), the shallow features significantly (paired t-test, $p < 0.05$) outperform the other two feature sets. Combining shallow and deep semantic features lead to significantly better results, while adding textual features (either to the shallow features only, or to the combination of shallow and deep features) does not improve the results.

To the best of our knowledge, there were no other studies that attempted at five-level text classification in the scope of conceptual text complexity. Although we used the same Newsela dataset in our unsupervised approach for automatic assessment of conceptual text complexity (Hulpuș et al., 2019), due to the different nature of the approaches (unsupervised vs. supervised), the five Newsela levels were used in a ranking task to evaluate the unsupervised approach, while here they are used in a five-level classification task.

Task	Features	with	without
Binary	Surface+Shallow	**.95**	.90
	Deep	.84	.80
	All	**.96**	.94
Five-level	Surface+Shallow	**.51**	.42
	Deep	.34	.30
	All	**.59**	.53

Table 6: The weighted average F-measure for both classification tasks (10-fold cross-validation with 10 repetitions) with Logistic classifier, in two setups: with and without the features that require paragraph information for their computation.

7. Influence of Paragraph Information

To better assess the universality of our systems, knowing that in real-world scenario some text will not have paragraph division, we wanted to explore how much paragraph organization influences the results of our systems on both classification tasks. Therefore, we compared classification performances on three different feature sets: surface + shallow, deep, and combination of all three types of features, in two scenarios: using all features of the corresponding type, and excluding the features that require paragraph information, e.g. features #2 and #5 for the surface text-based features (Table 1).

As can be seen (Table 6), the exclusion of features that use paragraph information significantly decreases (paired t-test, $p < 0.05$) classification performances on both tasks. The drop in performances is higher for the combination of surface and shallow features (-.09) than for the graph-based deep semantic features (-.04) on the five-level classification task. On the binary task, in contrast, the drop is similar for both feature sets (-.05 vs. -.04). Interestingly, on the binary classification task, the combination of all three types of features (surface + shallow + deep), performs extremely well even when the features which require paragraph information are excluded (.94 weighted F-measure and accuracy).

8. Summary and Outlook

Automatic assessment of text complexity usually takes into account just syntactic and lexical features. The task of automatically assessing conceptual complexity has only been proposed recently. Both proposed approaches, supervised and unsupervised, leverage the information from the DBpedia knowledge graph.

In this paper, we proposed two computationally less expensive and less noisy sets of features, the surface text-based features and the shallow semantic features. The computation for both of them only requires an entity linker and basic text processing (measuring offsets and counting occurrences), without need for accessing or processing any knowledge graphs.

We compared the performances of our two newly proposed sets of features with the state-of-the-art deep semantic features (computed over DBpedia knowledge graph) on two tasks: a binary classification task, and a five-level classification task. The results indicated that our shallow semantic features (alone, and in combination with the surface text-based features) perform significantly better than the previously proposed deep semantic features on both tasks. In the case of five-level classification task, the combination of all three feature sets performed significantly better than any of the three feature sets on their own did.

We also proposed a strong baseline for automatic assessment of conceptual text complexity, using four widely-known psycholinguistic features (familiarity, age of acquisition, concreteness, and imagery). They proved to be a strong baseline in the binary classification task, outperforming the surface text-based features and the graph-based deep semantic features. However, on the five-level classification task, they were outperformed by all other feature sets.

For future work, it would be interesting to explore how the choice of entity linker and mention identification influence the results.

9. Bibliographical References

Arfé, B., Mason, L., and Fajardo, I. (2017). Simplifying informational text structure for struggling readers. *Reading and Writing*, Oct.

Barzilay, R. and Lapata, M. (2008). Modeling local coherence: An entity-based approach. *Computational Linguistics*, 34(1):1–34.

Breiman, L. (2001). Random Forests. *Machine Learning*, 45(1):5–32.

Carroll, J. B. and White, M. N. (1973). Word frequency and age of acquisition as determiners of picture-naming latency. *Quarterly Journal of Experimental Psychology*, 25(1):85–95.

Cohen, W. W. (1995). Fast Effective Rule Induction. In *Proceedings of the Twelfth International Conference on Machine Learning*, pages 115–123.

Coltheart, M. (1981). The mrc psycholinguistic database. *The Quarterly Journal of Experimental Psychology Section A*, 33(4):497–505.

DuBay, W. H. (2004). The Principles of Readability. *Impact Information.*

Flesch, R. (1949). *The art of readable writing*. Harper, New York.

Freyhoff, G., Hess, G., Kerr, L., Tronbacke, B., and Van Der Veken, K., (1998). *Make it Simple, European Guidelines for the Production of Easy-toRead Information for People with Learning Disability.* ILSMH European Association, Brussels.

Gilhooly, K. J. and Logie, R. H. (1980). Age-of-acquisition, imagery, concreteness, familiarity, and ambiguity measures for 1,944 words. *Behavior Research Methods & Instrumentation*, 12(4):395–427, Jul.

Graesser, A., McNamara, D., Louwerse, M., and Cai, Z. (2004). Coh-metrix: Analysis of text on cohesion and language. *Behavior Research Methods*, 36:193–202. 10.3758/BF03195564.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H. (2009). The weka data mining software: an update. *SIGKDD Explor. Newsl.*, 11:10–18.

Hulpuş, I., Prangnawarat, N., and Hayes, C. (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.

Hulpuș, I., Štajner, S., and Stuckenschmidt, H. (2019). A spreading activation framework for tracking conceptual complexity of texts. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3878–3887, Florence, Italy, July. Association for Computational Linguistics.

Karamanis, N., Mellish, C., Poesio, M., and Oberlander, J. (2009). Evaluating centering for information ordering using corpora. *Computational Linguistics*, 35(1):29–46.

Keerthi, S. S., Shevade, S. K., Bhattacharyya, C., and Murthy, K. R. K. (2001). Improvements to Platt's SMO Algorithm for SVM Classifier Design. *Neural Computation*, 13(3):637–649.

Kintsch, W. and van Dijk, T. A. (1978). Towards a model of text comprehension and production. *Psychological Review*, 85:363–394.

Lapata, M. (2006). Automatic evaluation of information ordering: Kendall's tau. *Computational Linguistics*, 32(4):471–484.

le Cessie, S. and van Houwelingen, J. C. (1992). Ridge Estimators in Logistic Regression. *Applied Statistics*, 41(1):191–201.

Lovett, M. W., Borden, S. L., WarrenChaplin, P. M., Lacerenza, L., DeLuca, T., and Giovinazzo, R. (1996). Text comprehension training for disabled readers: An evaluation of reciprocal teaching and text analysis training programs. *Brain and Language*, 54(3):447–480.

Marcus, M. P., Santorini, B., and Marcinkiewicz, M. A. (1993). Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.

McNamara, D. S., Ozuru, Y., Graesser, A., and Louwerse, M. (2006). Validating coh-metrix. In *Proceedings of the Conference of the Cognitive Science Society*, pages 573–578.

Mesgar, M. and Strube, M. (2018). A neural local coherence model for text quality assessment. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4328–4339, Brussels, Belgium, October-November. Association for Computational Linguistics.

Meyer, B. J. F., Brandt, D. M., and Bluth, G. J. (1980). Use of the top-level structure in text: Key for reading comprehension of ninth-grade students. *Reading Research Quarterly*, 16:72â–103.

Meyer, B. J. F. (2003). Text coherence and readability. *Topics in Language Disorders*, 23(3):204â–224.

Newsela. (2016). Newsela article corpus. `https://newsela.com/data`. Version: 2016-01-29.

Paetzold, G. H. and Specia, L. (2016). Inferring psycholinguistic properties of words. In *Proceedings of the 2016 NAACL*.

Pitler, E. and Nenkova, A. (2008). Revisiting readability: A unified framework for predicting text quality. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 186–195, Honolulu, Hawaii, October. Association for Computational Linguistics.

Štajner, S. and Hulpuş, I. (2018). Automatic assessment of conceptual text complexity using knowledge graphs. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 318–330, Santa Fe, New Mexico, USA, August. Association for Computational Linguistics.

Štajner, S., Evans, R., Orasan, C., and Mitkov, R. (2012). What Can Readability Measures Really Tell Us About Text Complexity? In *Proceedings of the LREC'12 Workshop: Natural Language Processing for Improving Textual Accessibility (NLP4ITA)*, Istanbul, Turkey.

Štajner, S. (2018). How to make troubleshooting simpler? assessing differences in perceived sentence simplicity by native and non-native speakers. In *In Proceedings of the second LREC workshop on Improving Social Inclusion: Tools, Methods and Resources (ISI-NLP 2), Miyazaki, Japan*.

Xu, W., Callison-Burch, C., and Napoles, C. (2015). Problems in Current Text Simplification Research: New Data Can Help. *Transactions of the Association for Computational Linguistics (TACL)*, 3:283–297.

10. Language Resource References

DBpedia. (2014). *DBpedia dump 2014*. Freely available at: `https://wiki.dbpedia.org/`.

Newsela. (2016). *Newsela Corpus*. Freely available for research purposes upon request at: `https://newsela.com/data`, Version: 2016-01-29.