Project Title:
INFRASTRUCTURE AND INTEGRATED TOOLS FOR PERSONALIZED LEARNING OF READING SKILL

Project Acronym:

Grant Agreement number:
731724 — iRead H2020-ICT-2016-2017/H2020-ICT-2016-1

Subject:
D5.2 Content Classifiers

Dissemination Level:
PUBLIC

Lead Beneficiary:
Knowble

Project Coordinator:
UCL

Contributors:
All Partners

Revision	Preparation date	Period covered	Project start date	Project duration
V1	December 2018	Month 1-18	01/01/2017	48 Months

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Grant Agreement No 731724

Table of content

1. EXECUTIVE SUMMARY

This deliverable contains a description of the Content Classifier which is used for an automated classification of texts based on pre-selected metrics using the latest AI technologies. The Content Classifier was developed using Application Programming Interfaces (API's) of two parties, DFKI's Munderline(D5.3) and Knowble's 360AI - both of which have been trained and tested on a large amount of data.

As part of the deliverable, Knowble had to come up with the simple solution for assembling the whole system. To begin with, Knowble integrated the output of the DFKI's Munderline output that includes the user-model driven content metrics (D5.1) into the Content Classifier. In addition to that, Knowble had to build machine learning algorithm that is capable of identifying the appropriate reading age (between 4 and 11 years old) of children's texts.

The document also provides a description of how the Content Classifier will be utilized within the iRead software. In a nutshell, using current architecture, iRead final product should be able to match age-appropriate content with the learner's age and learner needs that are expressed in a form of syntactical features.

2. INTRODUCTION

Content classifier is an automated text classification technology that calculates age appropriateness of a text. It uses preselected quantitative, linguistic, syntactic and word difficulty metrics of the text as measures for age appropriateness. To make the process effective, the latest AI technologies to automate these tasks were used.

Namely, the Content Classification API (Application Programming Interfaces) consists of APIs of two partner parties - DFKI's Munderline and Knowble's 360AI - both of which work on machine learning principles.

The Content Classification System applies content classification that utilizes general quantitative metrics (Section 3.1), word difficulty distribution (section 3.2) and syntactic metrics (Section 3.3) as features of the neural network. This neural network is trained on an abundance of age-labelled texts (training dataset). Around 100 texts were used as training data for each difficulty level. The resulting AI model allows to automatically classify any texts for age appropriateness.

2.1 Content Classification within the iRead Product

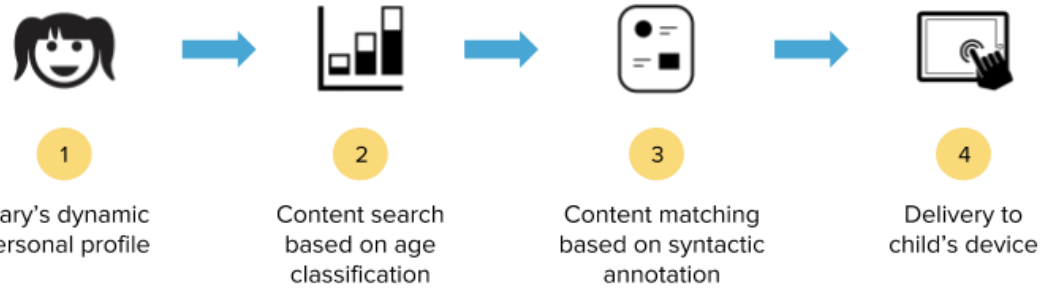

Image 1: The iRead content classification in the iRead product

This is an overview of how the Content Classifier will be used in the iRead product:

1. The learner's dynamic profile is used to inform personalized content selection.
2. Based on the learner's age, the age-appropriate content is filtered from the entire content library. For instance, content in the range of 8-9 years appropriateness is pre-selected for a learner of 9 years of age.
3. From the subselection of age-appropriate content, the one with the best match of linguistic and syntactic features based on the user-driven metrics of D5.1 is selected. In other words, content that matches the learner needs as expressed in syntactical features is used.
4. The selected text is then delivered in the iRead reader app to the learner.

3. CONTENT CLASSIFIER ARCHITECTURE

Content Classifier consists of two Application Programming Interfaces (APIs), DFKI's Munderline (see also http://iread.dfki.de/) and Knowble's 360AI (see also https://iread.360ai.nl). The technical architecture of DFKI's Munderline is in detailed described in the deliverable document of the iRead project - *D5.3 Syntax Analysers*.

Knowble has integrated the output of the Munderline syntactic parsers into the Content Classifier. The parsed content is annotated with the syntactic and linguistic metrics described in D5.1.

This Content Classifier was developed and integrated during the reported period of the iRead project making use of existing technologies and tools previously developed by the Knowble team.

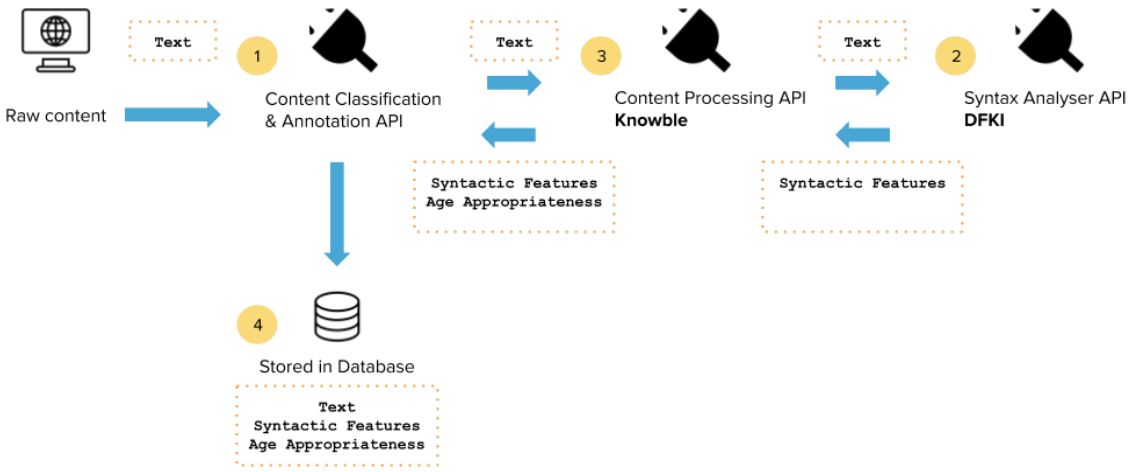

Image 2: Content Classifier Architecture

As part of the delivery of the Content Classifier, Knowble had to build a predictor to be able to determine the appropriate reading age of children's texts. Knowble's approach consists of extracting specific features from texts and training a neural network to predict the appropriate age range given those features.

This is an overview of what the Content Classifier architecture (see Image 2):

1. Content is being sent via a proxy system to reach Knowble's and DFKI's APIs.
2. Content is being analyzed and annotated with all predefined syntactic language features (see Section 3.3) via Munderline API.
3. Content is being classified for age-appropriateness.

4. The results of the content classification and annotation are then stored in the database.

3.1 Quantitative Metrics

The Flesch-Kincaid Ease and Grade level were used to understand the basis of quantitative metrics. Both provide readability scores that are based on the formulas depicted below.

Flesch-Kincaid Ease

$$206.835 - 1.015\left(\frac{\text{total words}}{\text{total sentences}}\right) - 84.6\left(\frac{\text{total syllables}}{\text{total words}}\right)$$

The score is able to indicate how easy it is to read this piece of content and what level of education is required - the higher the score, the easier it is to read this text (Kincaid, J.P., Fishburne, R.P., Rogers, R.L., & Chissom, B.S., 1975).

Flesch-Kincaid Reading Level

$$0.39\left(\frac{\text{total words}}{\text{total sentences}}\right) + 11.8\left(\frac{\text{total syllables}}{\text{total words}}\right) - 15.59$$

The score of this formula helps to identify which grade level the piece of content belongs to and is equivalent to the US grade level of education (Kincaid, J.P., Fishburne, R.P., Rogers, R.L., & Chissom, B.S., 1975).

To enable comparison against more advanced features (Vajjala & Meurers, 2012), the following quantitative metrics were employed by the Content Classification system:

- The number of paragraphs in the text
- The number of sentences in the text
- The number of words in the text
- The number of syllables in the text
- The number of phonemes in the text
- The number of letter-graphemes in the text
- The number of graphemes in the text (including punctuation)
- The fraction of unique words with respect to total number of words in the text
- The fraction of punctuation marks with respect to the total number of tokens in the text
- The average sentence length of the text
- The maximum arc length of the dependency tree of the parsed sentences
- The mean arc length of all arcs in the dependency tree of the parsed text

3.2 Word Difficulty Distribution

A different feature was based on a list of words that was labeled for the minimum reading level (Feng, Huenerfauth, Jansche, & Elhadad, 2010). The features consist of a numerical representation of the quantized distribution of 'word difficulty' (i.e scores between 0 and 5) for each word in the text. The feature consists of the positions where quanta of equal size

begin and end. Image 3 below shows a visualization of the method used. The distribution is converted to a cumulative distribution and values are taken at equal intervals. In practice, the feature gives more high values for texts that have a higher number of difficult words, and higher values if those words are more difficult.

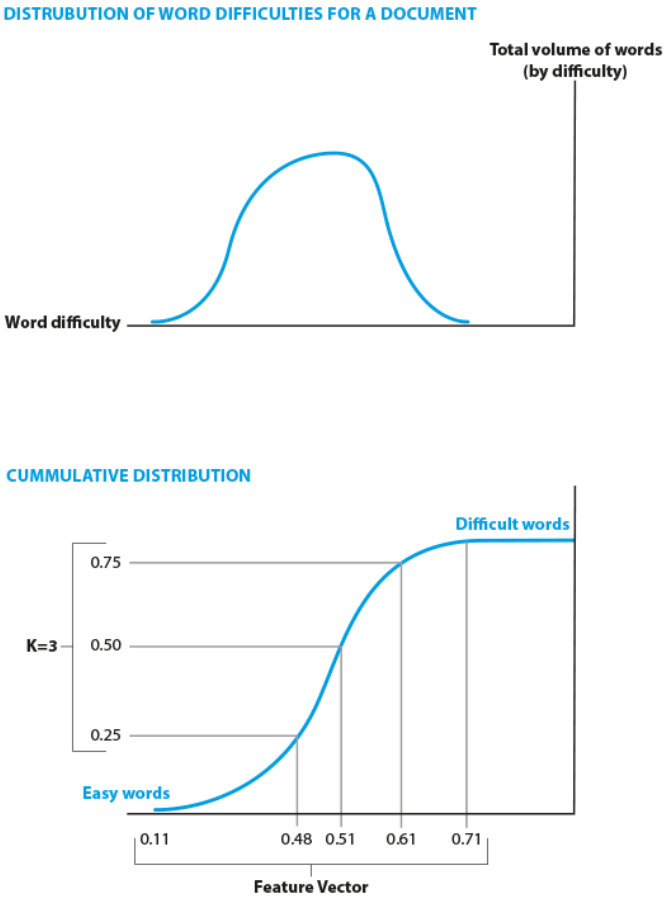

Image 3: Word Difficulties Distribution Method

3.3 Munderline's Syntactic features

As part of the deliverable D5.1, the metrics of syntactic phenomena contained in sentences were outlined and enriched with difficulty weight. A domain model was created by grouping features with similar impact. Thereafter a value on a scale (from 0 to 9) was assigned to each phenomenon, indicating their difficulty based on information available in the literature (Table 1).

The Munderline algorithm was trained to determine the syntactic structure of any new sentence. This syntactic information is used by Knowble's 360AI API for further processing (Neumann et al., 2014).

ID	DM Category	Factor (metric)	weight

183	Binding	Reciprocal pronouns: each other	0,2
184	Binding	Personal (object) pronouns	0,1
185	Binding	Reflexive pronouns	0,1
186	Coordination	or, and, nor, but, or, yet, so, and nor, but nor, or nor, neither, no more	0,3
187	Coordination	either...or, not only...but (also), neither...nor, both...and, whether...or, … etc.	0,5
198	Embedding	Adverbial clauses	0,7
199	Embedding	that-RCs, right-branching	0,8
200	Embedding	that-RCs, centre embedded	0,8
201	Embedding	that-RCs, subject extracted	0,8
202	Embedding	that-RCs, object extracted	0,8
203	Embedding	RCs with a relative pronoun	0,8
205	Embedding	Complement clauses: Complementizers (that compl.)	0,6
209	Negation	do not, don't, am not, is not, did not, have not, haven't, had not, hadn't, etc.	0,4
211	Passive	Short passive	0,9
212	Passive	Long passive	0,9
225	Wh- questions	object extracted 'what' questions	1
226	Wh- questions	Which-NP Questions	1
227	Wh- questions	Subject extracted 'who' questions	1
228	Wh- questions	object extracted 'who' questions	1
229	Wh- questions	adjunct questions	1
230	Wh- questions	subject extracted 'what' questions	1
232	Modals	Predictive: will/would/shall	0,4
233	Modals	Possibility: can/may/might/could	0,4

234	Modals	Necessity: should/must/(ought to/have to)	0,4

Table 1: Syntactic metrics in order of increasing difficulty

The content classification component will make use of information drawn from the individual user profile and each child's performance on the linguistic features included in the domain model of the user-driven metrics. There are three different sets of user-driven metrics (one for each user group - novice readers, dyslexia and EFL)
that were formulated and described along with a justification of their weighting in detail in D5.1.

At this point and time they were not included in the testing because they require user-data which is only available and stored when iRead games are used. Hence, these metrics will be tested in the evaluation phase.

3.4 Neural Network

The model was trained as a neural network with two fully connected layers of 120 and 84 nodes respectively. All activation functions are leaky ReLUs with a random leakiness sampled uniformly between ⅛ and ⅓. The input of the network consists of quantitative metrics extracted from the input texts and described in Section 3.2 and Section 3.3. The output consists of two numerical values, one for minimum and one for maximum age for intended readers of a text (Image 4).

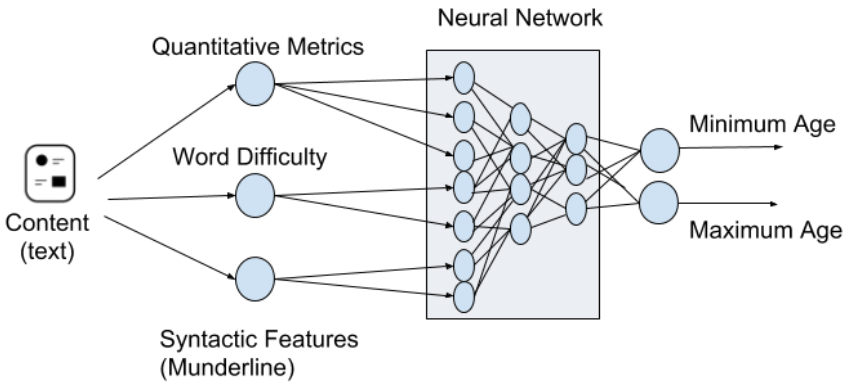

Image 4:Training content using neural network

4. PERFORMANCE OF CONTENT CLASSIFIER

4.1 Content Classifier Training

As a first delivery for the IRead project, Knowble completed a preliminary training of an algorithm to extract first set of given features and predict advised reading age, i.e. min and max, for children on text using regression method.

The dataset used was a combination of texts from the WeeBit[1] dataset and texts provided by the UCL team. Because the WeeBit data is larger in number, and it only covers ages of 7-8, 8-9, and 9-10. Whereas the UCL set contains fewer texts, but from a bigger range of ages (starting from 4), the data set was skewed with more texts intended for older children. WeeBit texts were labeled for more specific ages (e.g. 8-9), whereas UCL texts were mostly labelled for larger ranges, such as 9-11. Preliminary training classified text based on difficulty level and age groups. The difficulty levels used can be found in the Table 2. Once the model was set it was verified manually.

Difficulty levels	Age	Year
0	4-5	Reception
A	5-6	Y1
A	6-7	Y2
B	7-9	Y3-Y4
C	9-11	Y5-Y6

Table 2. Difficulty levels

The total combined dataset contains 2365 texts, of which the minimum and maximum ages already labelled were distributed as follows indicated in Table 3.

Minimum age			Maximum age	
Age	Number of texts		Age	Number of texts

[1] Industry applicable memory platform

4	41		5	41
5	49		6	43
6	60		7	63
7	622		8	608
8	774		9	692
9	818		10	809
10	1		11	10

Table 3: Labeled Data distribution between the minimum and maximum age

4.2 Content Classifier Testing

The testing of the content classifier was focused on validation of the data accuracy. For testing, the data set was split randomly into a test set of 300 texts and a training set of the other 2065. Because the data set was labelled as having an age range for each text, the minimum and maximum age for texts were used for evaluation. This effectively split the task into two separate tasks, although the model used for the two tasks was the same. The metadata was removed from the content and then analyzed by the Knowble's 360AI API. Afterwards, the results were compared with the originally provided metadata.

The mean squared error (MSE, equation 1) of the minimum and maximum age were used to calculate performance[2]. This punishes outliers while sparing predictions that are only slightly off. This is an appropriate measure for this task, because if a text is labeled as intended for 8-9 year-olds, this does not mean that a child of 7 years and 10 months cannot read it, whereas saying that a person should be at least 9 to read a text that is intended for 5-6 year-olds is inaccurate.

(1) MSE = $\sum_{i=0}^{N} \frac{(x_i - x_{i'})^2}{N}$,

where x_i is the minimum or maximum age indicated by human annotators for text i, and x'_i is the minimum or maximum age predicted by the model.

[2] For example: if the text was intended for 7-8, and the predictor's outputs were 6.91 and 7.82, the errors are 0.09 and 0.17. The error for the minimum (0.09) is squared (0.0081) and then averaged with all minimums of other texts. The same is done for the maximum age.

4.3 Results

The neural network was trained and evaluated once with and once without the Munderline Syntactic features. The initial run with features based on recent literature has reached a standard deviation of 0.6 on both minimum and maximum age, without fine-tuning. This means that on average it is not more than 0.6 years off on both maximum and minimum ages (Table 3).

	Excl. Munderline Syntactic features	Incl. Munderline Syntactic features
Mean Squared Error Minimum age	0.630	**0.570**
Mean Squared Error Maximum age	0.660	**0.556**

Table 3: Training Results showing Mean Squared Error

The data shows that adding syntactic features improved the accuracy of predicting the appropriate age for the texts considerably. The MSE scores correspond approximately to being off by 9 months of age on average, although the error could be lower in most individual cases.

Manual inspection shows that both predictors overestimate the minimum age for the youngest target audiences. It is likely that the skewness of the dataset is the cause of this and that more training examples for the youngest age groups would decrease the error for those groups.

5. OUTPUT OF THE CONTENT CLASSIFIER

The Content Classifier API (D5.2) is completely developed and implemented in Python and can run as a REST API interface.

Same as Munderline, the Content Classifier API runs as a REST service that accepts form-data inside the body of an HTTP POST request that is sent to https://iread.360ai.nl/en_ud. The API currently supports the English language.

The server can be tested using the curl tool. For analyzing some short text with Knowble's API, you can run:

```
curl -d "This is a test" https://iread.tst.360ai.nl/en_ud
```

The response consists of JSON string containing "min_age", "max_age" and "munderline", which contains the analysis from http://iread.dfki.de/munderline/match/en_ud, consisting of parsing data and syntactic features.

6. CONCLUSIONS

This deliverable describes the Content Classifier component that will be incorporated into the iRead product. The resulting Content Classifier API was developed as part of the work package 5 - it utilizes DFKI's Munderline (D5.3), the user-model driven content metrics (D5.1) and Knowble's 360AI technologies.

The content classification machine learning model (D5.2) has been trained and tested for determining the age appropriateness of texts for children between 4 and 11 years of age. We will continue working on this model during the cause of the project to further increase its performance.

The Munderline syntax analyzer (D5.3) has been trained and tested for all relevant languages of the iRead project and shows promising results with respect to performance and speed. We will continue working on Munderline during the cause of the project to further increase its performance and robustness.

7. REFERENCES

Feng, L., Huenerfauth, M., Jansche, M., & Elhadad, N. (2010, August). A Comparison of Features for Automatic Readability Assessment. Coling 2010: Poster Volume, 276-284.

Günter Neumann, Gerhard Paaß, and David van den Akker (2014): Linguistics to Structure Unstructured Information. Towards the Internet of Services: The THESEUS Program, in Wolfgang Wahlster; Hans-Joachim Grallert; Stefan Wess; Hermann Friedrich; Thomas Widenka (eds), Springer International Publishing Switzerland, ISBN 978-3-319-06755-1, pp. 383-392, 2014.

Kincaid, J. P., Fishburn, R. P., Rogers, R. L., Chissom, B.S. (1975). Derivation of new readability formulas for Navy enlisted personnel. Technical Report Research Branch Report, Millington, Tenn, Naval Air Station: 8-75.

Vajjala, S., & Meurers, D. (2012). On Improving the Accuracy of Readability Classification using Insights from Second Language Acquisition. The 7th Workshop on the Innovative Use of NLP for Building Educational Applications, 163-173.