Answering Confucius: The Reason Why We Complicate

Bernardo Pereira Nunes[1,2,3], Stella Pedrosa[3], Ricardo Kawase[2], Mohammad Alrifai[2], Ivana Marenzi[2], Stefan Dietze[2], and Marco Antonio Casanova[1]

[1] Department of Informatics - PUC-Rio - Rio de Janeiro, RJ - Brazil
{bnunes,casanova}@inf.puc-rio.br
[2] L3S Research Center, Leibniz University Hannover, Germany
{nunes,kawase,alfrai,marenzi,dietze}@l3s.de
[3] Central Coordination for Distance Learning - PUC-Rio - Rio de Janeiro, RJ - Brazil
{bernardo,stella}@ccead.puc-rio.br

Abstract. Learning is a level-progressing process. In any field of study, one must master basic concepts to understand more complex ones. Thus, it is important that during the learning process learners are presented and challenged with knowledge which they are able to comprehend (not a level below, not a level too high). In this work we focus on language learners. By gradually improving (complicating) texts, readers are challenged to learn new vocabulary. To achieve such goals, in this paper we propose and evaluate the 'complicator' that translates given sentences to a chosen level of higher degree of difficulty. The 'complicator' is based on natural language processing and information retrieval approaches that perform lexical replacements. 30 native English speakers participated in a user study evaluating our methods on an expert-tailored dataset of children books. Results show that our tool can be of great utility for language learners who are willing to improve their vocabulary.

Keywords: Technology enhanced learning, language development, learning process.

1 Introduction

Reading is a fundamental activity for all areas of knowledge. The practice of reading, strongly linked to the learning process, starts in the early school years and remains throughout life. Although, it requires mastery of certain techniques, reading is not a technical competence, but a process that begins in the relationship between the reader and text and continues by making sense of the text and promoting the development of new ideas influenced by prior background. For instance, reading helps to develop vocabulary in various forms of written and oral expressions. Thus, reading precedes writing being the main provider of basic elements for the production of texts.

A text ends up from the reading and the meaning that the reader gives to the text according to his understanding and the associations made shaped by the reader's prior knowledge and experiences. In this manner, it is possible to consider reading as a dialogic attitude in which the reader triggers threads of thoughts from a set of relationships with the text.

Trying to reduce the cognitive overload of reading activities, we often seek to develop activities that involve the simplification of texts, i.e., simplification of text preserving its original meaning [4]. Conversely, little has been explored about the possibilities of the *text sophistication* can make towards developing vocabulary. The introduction of words that are unusual for an individual or a group is an opportunity to expand their vocabulary knowledge. This can be achieved by transforming a text with simple vocabulary, read and discussed beforehand, into a more sophisticated terminology focusing on an individual or group.

The language develops in experiences influenced by sociocultural surroundings. If, due to various factors, this environment offers limited opportunities, the vocabulary of this group will be restricted and phrasal structures will be simple. Previous studies show that the language development of children is related to the sociocultural environment and that school interventions in early childhood education can minimize the differences between these children and those included in a privileged sociocultural environment with a greater range of opportunities. Thus, it imposes a natural limit to the expansion of vocabulary. However, reading nurtures new experiences and opportunities that contribute in the process of vocabulary acquisition and language development. This trend, coordinated with other activities, allows the *acquisition*, *expansion*, and *formation* of a more complex vocabulary, which contributes to learning any language, native or not.

In particular, communication is undeniably relevant in social relations amongst the groups that individuals attend. The vocabulary of a group gives the individual a sense of belonging and the sociocultural migration that education can provide is often barred, or at least hampered, by the limitations of the acquisition of new vocabulary. We do not suggest that expressions that are part of the sociocultural environment of origin should be overlooked, but they could be added to allow the expansion of types of communication and other sociocultural contexts. Another reason for vocabulary development is to improve communication. In the classroom, teaching vocabulary is often overlooked, although it is of well-known importance for learners of foreign languages to express their ideas clearly. Thus, new learning strategies are needed for learning vocabulary and development of autonomy.

In this paper, we introduce the 'complicator', a method that construes given sentences into a more sophisticated vocabulary. As hereby mentioned, the rationale behind the method is that one can learn (improve his vocabulary) by reading sentences that contain new and infrequent terms.

2 Method

In this section, we present our method for text sophistication based on lexical replacements as depicted in Figure 1. The method is divided into 4 main steps: (i) part-of-speech (POS) tagging; (ii) synonym probing; (iii) context frequency-based lexical replacement; and (iv) sentence checker.

2.1 Part-of-Speech Tagging

Words cannot be exchanged disregarding the context, otherwise, the sentence may result with a different meaning. For this reason, a first step is to identify the right part-of-speech of each word in a sentence and then look for a suitable synonym.

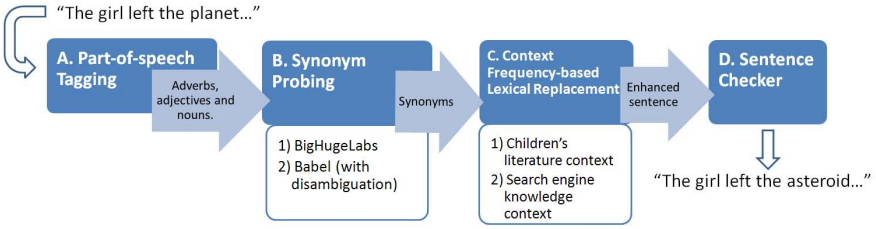

Fig. 1. Complicator workflow

In this step we used the Stanford Log-linear Part-Of-Speech (POS) Tagger [5] to annotate each word in a given sentence. We focused on 3 POS groups: *adjectives*, *nouns* and *adverbs*. As depicted in Figure 1, the first step outputs the original sentence with the POS tags assigned to each word.

2.2 Synonym Probing

In this step, we identify synonyms of given nouns, adverbs and adjectives of a sentence. As a word can have multiple synonyms, we divide the synonyms into categories and filter out the synonyms that express different meanings.

To identify the sense of a word in a given context, we used the Babelnet API[1] developed by Navigli and Ponzetto [3]. Babelnet is built on the top of Wordnet[2], which is a lexical database in English that groups words into synsets, i.e. words that denote the same concept.

We also used a thesaurus database[3] of the Big Huge labs. However, this thesaurus does not provide any information regarding the sense of each word. Hence, we only filter out the synonyms that belong to a different POS category and without filtering by sense.

Thus, this step is responsible for finding a set of synonyms for a particular word in the original sentence and for outputting a filtered set of synonyms according to a specific context.

2.3 Context Frequency-Based Lexical Replacement

As we already have the set of synonyms filtered by sense from the previous steps, this step aims at identifying a synonym for a word that best fits in a determined context. For this, we rely on how frequent a word is found in a controlled vocabulary. The synonym that will replace the original word is the next most frequent word in the vocabulary, but less frequent than the original one.

We can focus on a specific domain to sophisticate a sentence according to a target audience. Given a controlled vocabulary, our method is able to select the most suitable words that match a context level (see Section 3).

[1] `http://lcl.uniroma1.it/babelnet/`
[2] `http://wordnet.princeton.edu/`
[3] `http://words.bighugelabs.com`

2.4 Sentence Checker

To validate the lexical replacements made in the previous step, we are aided by the inherent knowledge of search engines. As search engines crawl content created by humans on the Web, the co-occurrence of words in the same sentence implicitly represent the common sense. Thus, the validation process of the new sentence generated begins by splitting the sentence into chunks of a predetermined size (i.e. windows size of a sentence) and querying on search engines to check if these chunks occur on the Web in high scale. If they occur, then the new sentence is validated. Otherwise, no changes are made on the sentence.

3 Evaluation

Our evaluation aims at validating the methods with respect to preservation of the original meaning and its grammatical correctness. Basing on native English speakers, our main goal is to validate our complicating process regarding potential errors introduced by our method and to check if the texts preserve their original meaning. Thus, in this evaluation, we present to the participant a text retrieved from our dataset as well as its complicated form. The questionnaire for the native English speakers is composed by the following questions: (1) Do the texts above have the same meaning? (yes/no); (2) Is the text free from grammar errors? (yes/no).

As for the dataset, we used in total 1325 sentences pairs extracted from the Terence corpus [1]. For each book in the Terence corpus, we tokenized the sentences using the Stanford NLP tool to keep the sentence structure.

The complicator tool contains many parameters for each of which the settings must be specified. Here, we describe the parameters for setting the synonym source, the controlled vocabulary and the windows size of the sentence checker. Our goal is to provide a tool that can be adapted to a specific context.

Synonym source: This parameter is used to control the synonyms suggested for a given word. In our experiments we used WordNet and BigHugeLabs (described in 2.2).

Controlled vocabulary: This parameter is used to customize the simplification to a target audience. Although the list of synonyms provides words with the same sense, a specific word might not be used by a target audience, thus the controlled vocabulary will assist in picking up the right synonym in a given context. We used four vocabularies, (i) Age 7-9 Level, (ii) Age 9-11 Level, (iii) Age 9-11 Level and (iv) Search Engine.

Window sizes: This parameter defines the boundaries of a sentence. The set of words will be checked regarding its popularity, i.e., to prevent obscure and rare sentence formulations. We set the window size between 1 and 3.

4 Results

The questionnaire was answered by 30 native English speakers and covered all sentences in the dataset (original and complicated sentences).

Table 1 presents the results of the evaluations with native English speakers. The column ‘Complicated sentences’ shows the percentage of sentences that were, to some

extent, modified by the methods. The column 'Precision (same meaning)' shows the agreement of the evaluators regarding the sense similarity between the original and the simplified sentence; the column 'Precision (grammatically correct)' shows the rate of the sentences that were simplified and were free from grammatical errors.

The results are also discriminated regarding their different configuration settings for which we vary the window size, the controlled vocabulary and the synonyms source.

Table 1. Results of the complicator method for different strategies (parameter settings) from the evaluation with native English speakers

Strategy ID	Window's size	Vocabulary source	Synonym source	Complicated Sentences (%)	Precision (meaning) (%)	Precision (grammar) (%)
S_1	1	Age 7-9 Level 4	WordNet	19.92	68.32	79.21
S_2	1	Age 7-9 Level 4	BigHugeLabs	75.2	67.28	59.42
S_3	2	Age 7-9 Level 4	WordNet	3.32	81.25	81.25
S_4	3	Age 7-9 Level 4	BigHugeLabs	38.48	67.86	67.35
S_5	1	Age 9-11 Level 1	WordNet	5.27	69.23	69.23
S_6	1	Age 9-11 Level 1	BigHugeLabs	59.38	62.05	55.45
S_7	2	Age 9-11 Level 1	WordNet	1.37	83.33	66.67
S_8	3	Age 9-11 Level 1	BigHugeLabs	15.82	65.00	68.75
S_9	1	Age 9-11 Level 4	WordNet	10.74	64.81	72.22
S_{10}	1	Age 9-11 Level 4	BigHugeLabs	0.2	0	0
S_{11}	2	Age 9-11 Level 4	WordNet	2.54	91.67	91.67
S_{12}	3	Age 9-11 Level 4	WordNet	2.34	72.73	90.91
S_{13}	2	Search Engine	WordNet	6.45	65.63	75.00
S_{14}	3	Search Engine	WordNet	7.23	58.33	80.56
S_{15}	2	Search Engine	BigHugeLabs	42.77	68.52	63.43
S_{16}	3	Search Engine	BigHugeLabs	7.62	60.53	55.26
S_{17}	2	Age 9-11 Level 1	BigHugeLabs	0.2	0	0
S_{18}	3	Age 9-11 Level 1	WordNet	0.59	100.00	100.00
S_{19}	2	Age 9-11 Level 4	BigHugeLabs	13.48	69.12	61.76
S_{20}	3	Age 9-11 Level 4	BigHugeLabs	5.86	58.62	68.97
S_{21}	1	Search Engine	WordNet	58.4	77.78	80.56
S_{22}	1	Search Engine	BigHugeLabs	0.2	0	0
S_{23}	2	Age 9-11 Level 4	BigHugeLabs	0.2	0	0
S_{24}	3	Age 9-11 Level 4	WordNet	3.32	75.00	87.50

5 Discussions and Conclusions

The results show that strategies S_2 (Age 7-9 Level 4), S_6 (Age 9-11 Level 1) and S_{21}(Search Engine) achieve the highest degree of lexical replacements. These are strongly related to the size of contextualized dictionary built for each level. The most important is the result in terms of precision, regarding meaning and grammatical correctness. For this case, we see that most of the values are above 60.0%. In fact, the overall precision of the complicator aggregating the variables (window size, vocabulary and

synonym) is 66.75% for meaning and 64.76% for grammar. This rather high precision numbers support the utility and applicability of our proposed method. Additionally, we believe that the complicator can significantly improve if it is used in combination with better synonyms sources. The freely online available sources used in these experiments are overwhelmed with out-of-context synonyms.

As aforementioned, the 'complicator' supports strategies for expanding vocabulary necessary to convey ideas in a different language, social contexts or environments that might require different language skills. However, in some cases, the synonymy presented may not be suitable and, therefore, every word replaced must be evaluated by a user or group of users that will use the tool.

The dynamics generated by this substitution of words resemble the use of the dictionary and, it helps to expand the vocabulary and learn the different meanings of words and expressions. As studied by Krieger [2], the use of the dictionary can be used for development at different levels of reading and textual production, therefore it plays a key role as a didactic method for expansion and improvement of knowledge of a lexicon language.

Therefore, we believe that similarly to the dictionary the 'complicator' is able to contribute as a didactic resource in the development of skills related to the domain of a language. However, it is worth noting that to achieve the stated objectives, it is essential that the use of the 'complicator' is accompanied by a teacher or someone who has prior knowledge of this tool and that recognizes its didactic potential.

Acknowledgment. This work was supported by the TERENCE project, funded by the EC through FP7 for RTD, ICT-2009.4.2.

References

1. Arfé, B., Jane, O., Pianta, E., Alrifai, M.: Story simplification: User guide. Technical report d2.2, terence project, 2011, Technical report (2011), `http://terenceproject.eu`
2. da Graa Krieger, M.: Dicionários para o ensino de lngua materna: princpios e critérios de escolha. Revista Língua e Literatura Frederico Westphalen 6 e 7, 101–112 (2004/2005)
3. Navigli, R., Ponzetto, S.P.: Babelnet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. Artificial Intelligence 193, 217–250 (2012)
4. Nunes, B.P., Kawase, R., Siehndel, P., Casanova, M.A., Dietze, S.: As simple as it gets - a sentence simplifier for different learning levels and contexts. In: ICALT (to appear, 2013)
5. Toutanova, K., Klein, D., Manning, C.D., Singer, Y.: Feature-rich part-of-speech tagging with a cyclic dependency network. In: Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology, vol. 1, pp. 173–180. Association for Computational Linguistics, Stroudsburg (2003)