e-Rural: a framework to generate hyperdocuments for milk producers with different levels of literacy to promote better quality milking

Vanessa Maia Aguiar de Magalhaes[1,2], Junia Coutinho Anacleto[1], André Bueno[1], Marcos Alexandre Rose Silva[1], Sidney Fels[3], Fernando Cesar Balbino[1]

[1]Universidade Federal de São Carlos (UFSCar) – São Carlos, SP – Brasil,
[2]Embrapa Gado de Leite, [3] University of British Columbia
{vanessa_magalhaes, junia,andre, marcos_silva,balbino}@dc.ufscar.br; ssfels@ece.ubc.ca

Abstract. We created and tested e-Rural, an approach to allow educators to dynamically adjust the target literacy level for their online learning content using a combination of three tools: PACO-T for planning, COGNITOR for editing hyper documents and Simplifica for text simplification. PACO-T and COGNITOR use the Brazilian Open Mind Common Sense knowledgebase (OMCS-Br) to provide access to commonly held understandings and beliefs on a diverse set of topics associated with a large range of Brazilian demographics, including, people with low literacy. We tested our experiment with 13 users that were creating hyperdocument-based learning content to describe important methods to milk production. We chose milk production as this is one of Brazil's primary agricultural products and yet it has been established that there is a wide gap between the content from researchers with methods to greatly enhance the quality and economic power of milk production and the tacit knowledge and procedures of the farmers who actually produce the milk who are often at low literacy levels consistent with Brazil's low literacy levels being around 75% of the population. Our experiments reveal that educators are able to produce milk related learning content geared towards different literacy levels using our tools with a very satisfying efficacy and efficiency levels. Thus, we believe that the use of our approach that introduces demographically sensitive common sense holds promise to bridge the gap between high literacy researchers with evidence-based approach to milk production and tacitly-based, low-literacy milk producers to better develop the milk industry in Brazil.

Keywords: Accessibility, literacy, textual simplification, textual equivalents, W3C Recommendation.

1 Introduction

In Brazil, despite the revealing data about the improvement in the population's educational levels, when compared to the ones of last decade, it is possible to notice that the average education level of the population is still insufficient and not compatible with the economic development of the country [1]. According to the National Indicator of Functional Illiteracy [2], approximately 75% of Brazilian

population between 15 and 64 years of age do not have the complete level of literacy. In other words, 75% of Brazilian do not have complete writing and reading skills to support them understanding information and using it in their work, habit, life, etc.. Within this percentage 7% are considered as absolute illiterate, 21% as literate at a rudimentary level, 47% at a basic level and only 25% have full literacy. This reality, illustrates that Brazilians have difficulties understanding texts and, consequently, have limited access to information, knowledge and cutting edge technology. Social, cultural, educational, perceptual, cognitive, and motor differences existing among people are cited as primary factors contributing to this reality.

A way to deal with the different needs and diversities in Brazil is to consider adapting the content of texts and information according to literacy level through using ICTs (Information and Communication Technologies). ICT have been shown to be efficient and effective as tools for: working, study, entertainment, means of expression and communication among people of different ages, special needs, abilities, capacities and interests. Understanding how the previously cited factors influence the way people read and understand a text, as well as the way they access and use ICT, can be a pathway to friendlier and easier ways to use these technologies to present a particular adapted text, facilitating better understanding and usage by everyone, independent of their socio-economic, cultural, educational, and cognitive conditions status. If successful, this approach improves the democratization of information and promotes autonomy of citizens.

We have investigated creating and making hyperdocuments according to the literacy level of a specific target public including using specific cultural knowledge represents this public's knowledge, values, etc. The main objective of using the contextualized documents is to allow the public to identify and understand what is written by taking into account their educational and cultural level. This information is integrated within a contextualized hyperdocument.

The use of these contextualized hyperdocuments in the context of Embrapa (Brazilian Company of Research on Agriculture and Livestock), specifically Embrapa Cattle and Milk, is a strategy to disseminate information and knowledge respective to the Normative Instruction 51 (NI 51) of the Ministry of Agriculture, Cattle Breeding and Supply (MAPA) of the Federal Government of Brazil. The NI 51 is mandated to be transmitted to every person involved with the milk production in Brazil, for the purpose of improving the quality of the milk produced in Brazil through establishing minimum requirements during the production. The problem lies in that the information in the instructions is full of complex terms that are hard to understand and is written for an audience that is expected be fully literate. As pointed out above, in Brazil, this assumption is not valid, especially for the farmers who produce the milk, leading to the instructions to be ignored by critical people in the milk production workflow. Thus, The efforts to improve the quality of milk production through the dissemination of evidence-based information is thwarted.

To address this problem, we conducted an experiment performed using our framework for creating contextualized hyperdocuments that considers rudimentary literacy level and cultural knowledge obtained through a project called Open Mind Common Sense in Brazil (OMCS-Br) described in the next section. This paper is organized as follows: section 2 describes the common sense and OMCS-Br project, in

section 3 covers our framework and its characteristics, section 4 details our experiment and results followed by final considerations in section 5.

2 Common sense and OMCS-Br Project

Common sense is defined as set of known facts by most people who live in a particular culture in certain age, "arraying a wide part of the human experiences, knowledge about the spatial, physical social, temporal and psychological aspects" [3]. In order to collect this kind of common sense knowledge and use it to develop contextualized technological applications, i.e., applications which consider cultural human knowledge in their interface and content, the Advanced Interaction Laboratory (LIA) of the Federal University of São Carlos (UFSCar) in collaboration with MediaLab-MIT developed OMCS-Br Project, which is a project in Brazilian Portuguese language [6].

In this project, there is a website www.sensocomum.ufscar.br where Brazilians can access it to tell what they know, belief, think, etc., in other words, they can tell about their commonsense knowledge. To enter in this website, a previous enrollment is necessary because according it the project can provide filters that allows queries considering a particular profile (age group, geographical localization, gender and level of academic training).

In this website, there are nine distinctive themes and twenty activities aiming to collect and approach types of knowledge that compounds the people's common sense. For example, there are themes to collect what Brazilians know about colors, sex, slang, among others. It was created a theme called "All about Milk" to collect information about milk production for this specific research, described in this paper. Through this theme, it was possible to collect people´s cultural knowledge and vocabulary about cow, milk production utilities, steps for milking and other useful information necessary to create hyperdocuments taking into consideration people´s reality. It is important to say that this research also uses the whole knowledge collected from other themes and activities.

Collecting common sense is done through templates, as shown in Fig. 1, a template from "All about milk" theme. Templates are sentences with a dynamic part (dashed green), fixed part (outlined in yellow) and gaps (the second rectangle), to be filled by people according to what they know or believe to be true. The dynamic part changes for each user interaction harnessing the knowledge already collected in other interactions, thus, the website has a feedback system to use the knowledge from the base to collect new information.

Fig. 1. Example of a template from "All about milk" theme

In the beginning the collected knowledge is stored in the OMCS-Br knowledgebase in a natural language. As the computer does not deal well with the natural language, a processing which generates a dynamic net called ConceptNet was performed, based on the concepts and twenty Marvin Minsky relations. Minsky was a

researcher of the artificial intelligence area who researched about the human knowledge mapping for the computational [4]. ConceptNet communicates with the computer tools, like some tools existents in e-Rural framework explained follow, from a set of functions, an Application Program Interface (API), which was developed for this purpose.

3 e-Rural framework

e-Rural framework was created to support educators to dynamically adjust the target literacy level for their online learning content using a combination of three tools PACO-T, Simplifica and Cognitor. This framework describes step by step what educators need to do and what computer tools they need to use in order to do hyperdocuments according target group´s profile. i.e., target literacy level.

It is important to say that the main characteristic of this framework is to support educators who do not have good technological background to use many different and independent tools to do a hyperdocument. Once this kind of background is not common among Brazilians educators and, in previous experiments using these tools independently it was possible to notice that educators did not know when they could use tools. Then, we integrated these three tools in a single framework. Thus, some adjustments were made in the tools to improve their use and to make them work better together. As one example, we can cite the insertion of the functionality of search on the base of common sense in the Simplifica tool.

In framework (see Fig.2), the first tool used is PACO-T [9], designed to assist teachers in the planning of learning consists of seven steps yet we used this tool until step six because in the seventh step we used other two tools which are Simplifica and Cognitor.

Simplifica [7] aims to lexically and semantically simplify a text for people of rudimentary and basic levels of literacy. The lexical simplification process starts with the identification of difficult words in the text, that is, words not found in the Porsimples dictionary that lists words from the children vocabulary and from daily newspaper. These words are less frequent in the children's daily lives, i.e., agricultural technique words are considered complex. For each complex word have been found we perform a search in other dictionaries and in common sense knowledgebase to find synonyms, if there are any. After lexically simplification it is done strong semantically simplification.

The third tool is Cognitor [8] aiming to support organizing and editing educational content. Cognitor allows teachers to create contextualized contents considering concepts and analogies used and well known by learners'.

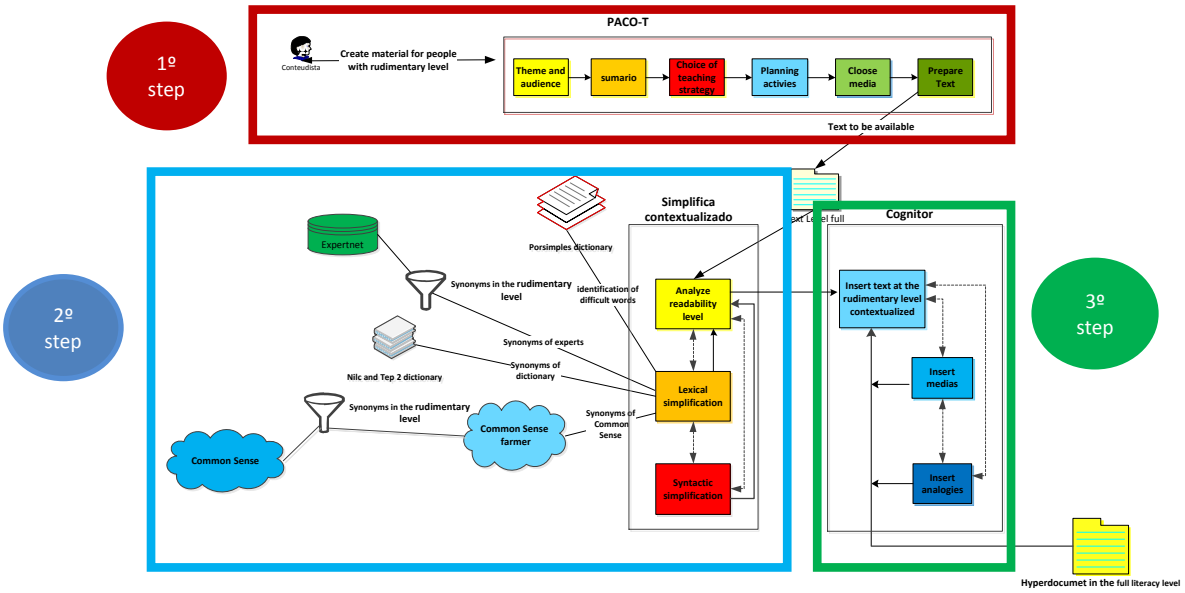

Fig. 2. e-Rural framework

For the context approached in this paper, the rural producer ones, who have low literacy most of the times, the usage of texts with simplified structure and of reduced dimensions does not guarantee total access to the information, being necessary other means to transmit this knowledge. Thus, the development of applications for these individuals requires special care in the elaboration of the content. In order to make this content more accessible, the textual equivalent cultural contextualization is proposed, as well as the linguistic simplification to this public level of literacy.

4 Experiment

The experiment was conducted with 13 people organized into three groups A with computer science researcher's users, B with professionals in computer science and digital content creator users and C with researchers and technicians specialized in milk quality users. All users received a script to be followed and applied at the e-Rural framework. The purpose of this script was to create and to plan a learning content using the PACO-T tool at first. A text set out in high level of literacy, part of the NI 51, chosen by the user should be transformed to the rudimentary level using Simplifica, and finally should be created a hyperdocument with the transformed text using the Cognitor tool. Each user chose a text from the script (there were 6 texts) and then the level of readability of the text chosen was measured, lexically simplified, and when necessary, syntactically simplified. After the textual simplification process was done (with the level of text readability in rudimentary level of literacy), the user created the hyperdocument with the simplified text. At the lexical simplification process the user could look for synonyms in the OMCS-Br cultural knowledge base, use his/her own knowledge or the offered replacement suggestions offered by

Simplifica. In the syntactic simplification process, users could choose the best sentence using the offered suggestions by Simplifica and change them as well as using their own common sense.

Hypotheses were formulated for the study regarding the creating contextualized hyperdocument process.

Null hypothesis (Ho) - Using Simplifica tool is enough resource on text simplification to process text from full literacy level to rudimentary level. In this context, it is not necessary to use common sense because it does not support Simplifica on lexically simplification.

(H_1) - Using common sense is a useful knowledge to create contextualized hyperdocuments with simple texts in rudimentary level.

Hypothesis test. The purpose of this step is to verify with any significance degree (α), if it's possible to reject the null hypothesis (H_0) in favor of some other alternative hypothesis (H_1) based on the data set obtained. If it's not possible to reject the null hypothesis, nothing can be said about the study result.

Table 1 shows the summary of the textual simplification process made by the group. The first column represents the number of words that were automatically identified by Simplifica as complex during the lexical simplification process. The second represents the number of words replaced by their synonyms found in the OMCS-Br cultural knowledge base, available with the contextualized Simplifica tool, that was improved with the functionality to be inserted in the e-Rural framework. The third column represents the number of words replaced by the users' suggestions that we also considered as a result from the e-Rural framework. The fourth column represents the number of complex words that users have replaced using the replacement suggestions from Simplifica (not the contextualized version). And the last column represents the need for syntactic simplification to achieve the level of readability of rudimentary literacy.

Table 1 - Results of the textual simplification process by group A, B and C:

Texts	Nº of Words in the text	OMCS-Br cultural knowledge (Ms)	Replacement suggestions from Simplifica(Md)
Group A	3	26	1
Group B	5	15	0
Group C	3	13	0

At first, when comparing synonyms obtained from the common sense knowledgebase with the synonyms suggested by the Simplifica tool in the Table 1, it is possible to see the first indication that the use of the common sense knowledgebase helped in the contextualized culturally hyperdocuments creation allowing the chance to transform a hyperdocument into its rudimentary level. To see this effect in a statistical way, we applied the paired t-test [6]. This parametric test is used to compare two independent samples and check if their averages are statistically different and thus show that the hypothetical effect was demonstrated.

In our experiment, there are the text data samples contain complex words that were replaced by synonyms found in the common sense knowledgebase or by the user own common sense and by words suggested by the Simplifica tool. The paired t-test is based on the idea that, when you are looking for differences between two samples X and Y, you must judge the difference between their averages considering the dispersion or the variability of the data that compose them. It is because as greater the

data variability from the two samples compared, greater the chances of these data are superimposed on the normal distribution, even if the average doesn't change. The following equations show the method of paired t-test. It is one reason: the numerator is the average difference, and the denominator, also called the standard error of difference, there is a measure of variability or data dispersion.

The T value will be positive if the first sample average is greater than the second, and negative if it's less. After calculating T, the default distribution table of statistic probability t-Student [10] should be checked to discover whether the calculated ratio is sufficiently large so that we can attest that it is unlikely that the sample difference was mere coincidence. It establishes, therefore, a significance level α which represents the "cut point" or "risk level" with which it is allowed to claim that there is difference between the samples and to reject the null hypothesis. Becomes α = **0.025**. This means that we assume a **2.5%** risk of finding a significant difference between the samples averages, even though this gap was by chance (false positive). The confidence level of the test result in this case would be **97.5%**. It is also necessary to establish the freedom degree (fd) for the test, which corresponds to the total number of samples groups subtracted of two (fd = n-1). Thus, taking T, α and fd the t value should be noted in the standard table to determine if the T value is large enough to be significant. The t value checked in the table indicates the probability of having obtained the calculated difference between the samples averages if both factors were equal. For the study, the t value would indicate the probability of having been obtained the difference if the two samples were based only on the conventional process of replacement from the Simplifica tool. If this probability is too small, then can be concluded that the observed study result is statistically significant. In this case, if the t value on the table related to the chosen cut point and the freedom degree is less than the calculated T, the null hypothesis is rejected. Otherwise, the null hypothesis is not rejected and no conclusions can be obtained from the study.

Once the effect on the dependent variable in the study (common sense effectiveness) involved an aspect, the synonymous substitution, the application of the paired t-test in the sample data set was accomplished in one step. We compared the related samples to the replaced words number by the each group common sense. The null hypothesis test was based on the rejection criterion test, if $T > t\alpha$, H_0 is rejected. Otherwise, H_0 is not rejected.

Considering data from s, it is possible to obtain that T = 4,7595 and the freedom degree is fd = n-1 = 3 -1 = 2, and significance degree α = 0,025. In the distribution table of statistic probability t from Student, t is $t_{0.025,2}$ = 4,303. Taking into consideration $T > t_{0.025,2}$ was possible to reject *Null hypothesis (Ho)* with α = 2,5%, on other words, 2,5% of chance to reject *Null hypothesis (Ho)* as true and 97,5% to confirm that we have a correct test.

5 Conclusions.

Once the null hypothesis was rejected (using only the suggestions provided by the Simplifica tool it is possible to transform a technical paper from the full level of literacy to the rudimentary level of literacy), it is possible to conclude regarding the

independent variables influence on the dependent variables, whereas the experiment is valid and the threats to validity were treated. With the H_0 rejection in the study, we can state that the observed differences in the effectiveness of using synonyms from the common sense knowledgebase to create hyperdocuments for farmers who have rudimentary level of literacy and using the synonymous offered by the substitution suggestions has statistical significance, ie., the treatments applied (the two processes of lexical substitution) were the cause of the efficiency changes and not because of mere chance. As noted on the data presented in Table 2 the average of words replaced by common sense knowledgebase in the culturally contextualized hyperdocument creation process using the e-Rural was higher than the synonyms used and available in the replacement suggestions from Simplifica tool ($\mu s_{Sense} > \mu s_{Tips}$).

Finally, considering the experiment was conducted in vitro under controlled conditions, it is important to emphasize that the conclusions about the observed results in this work are restricted to the scope of researchers in the field of computer science, professors and expert researchers in the agriculture area, in an university environment and a research institution. For reasons of external validity, to extend the generalization of the observed phenomenon to an even broader context of other units of Embrapa interested, it is necessary further studies in other environments, in different contexts in order to obtain a broader validation of the research hypotheses. The tools union and the help of the common sense knowledgebase from a single web computing framework, e-Rural, helps to reduce the effort, avoiding extra work on editing hyperdocuments, including editing technical text in rudimentary literacy level.

REFERENCES

1. IBGE - INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA, 2009. <http://www.ibge.gov.br/home/estatistica/populacao/condicaodevida/indicadoresminimos/sintese indicsociais2009/indic_sociais2009.pdf>.
2. Indicador de Alfabetismo Funcional – *INAF BRASIL 2009*. <http://www.acaoeducativa.org/images/stories/pdfs/inaf2009.pdf>
3. Liu, H.; Singh P., *ConceptNet: A Practical Common Sense Reasoning Toolkit*. BT Technology Journal, v. 22, n. 4, 211-226.
4. Minsky, M.,1986. *The Society of Mind. Simon and Schuster*, New York.
5. Singh, P. *The OpenMind Commonsense project*. KurzweilAI.net. <http://web.media.mit.edu/~push/OMCSProject.pdf>.
6. Anacleto, J. C. et al. "Can common sense uncover cultural differences in computer applications?" In: BRAMER, M. AI in theory and practice. Berlin: Springer-Verlag, 2006. v.217, p1-10.
7. Jr., A. C.,Maziero, E. et al. "Supporting the adaptation of texts for poor literacy readers: a text simplification editor for brazilian portuguese". In Proceedings of Workshop of Innovative Use of NLP for Building Educational Applications at NAACL 2009 (2009).
8. Buzatto, D. ; Anacleto, J. C. ; Dias, A. L. . Providing Culturally Contextualized Metadata to Promote Sharing and Reuse of Learning Objects. In: ACM SIGDOC 2009, Bloomington, Indiana. Proc of the 27th ACM SIGDOC. New York : ACM, 2009. p. 163-170.
9. Carvalho, A. F. ; ANACLETO, J. C ; NERIS, V. P .A. PACO-T: A Computational Framework for Planning Cultural Contextualized Learning Activities by Using Common Sense. In: IFIP World Conference on Computers and Education (WCCE 2009), 2009. In Proc. WCCE 2009.
10. http://en.wikipedia.org/wiki/Student%27s_t-distribution