S. L. Wong et al. (Eds.) (2010). Proceedings of the 18th International Conference on Computers in Education. Putrajaya, Malaysia: Asia-Pacific Society for Computers in Education.

Automatically Grading the *Use of Language* in Learner Summaries

Iraide Zipitria, Ana Arruarte, Jon A. Elorriaga
University of the Basque Country UPV/EHU
iraide.zipitria@ehu.es, a.arruarte@ehu.es, jon.elorriaga@ehu.es

Abstract: A summary is a short clear description that gives the main facts or ideas about something. Summaries are widely used in traditional teaching as an educational diagnostic strategy to infer comprehension, or how much information from text is retained in memory. From its early beginning Computer Supported Learning Systems research has faced open-ended learners' response evaluation. Global summary grading includes partial discourse assessment on relevant issues such as coherence, cohesion, comprehension, adequacy and use of language. This paper describes the procedure followed searching for the best available approach to model the *use of language* grading of learner summaries written in Basque language.

Keywords: Automatic summary grading, use of language

Introduction

In educational contexts, a summary contains the main ideas which learners recall from text. Thus, summaries are widely used in traditional teaching as an educational diagnostic strategy to infer comprehension, or how much information from reading text is retained in memory [1-3]. Broadly speaking, there are two ways to approach learning diagnosis: close-ended and open-ended. The close-ended evaluation (multiple choice test, ordering exercises, etc.) has the advantage of being countable, assures the possibility of objective assessment and makes possible accurate computation. Therefore, it has been most widely used in Computer Supported Learning Systems (CSLSs) for learning evaluation. But, the close-ended mode also involves a greater probability of scoring by chance and restricts the freedom and richness of learner responses. Open-ended evaluation assesses free text. This assessment mode, although less accurate, has also been present in Artificial Intelligence and education since the very early works in Socratic dialog systems [4-8]. By using free text learners have the freedom to write anything and can construct their own answer. Therefore, it is more likely to obtain a better approximation to real learners' knowledge in open-ended modes. However, automatic evaluation in open-ended modes is complex and has to face uncertainty. Still, new developments in Natural Language Processing (NLP) allowed a rebirth with a variety of open-ended approaches in various applications: dialogue systems [9-12], feedback in narratives [13], etc. The work presented in this paper has been carried out in the context of an automatic summary grading environment, LEA [14]. Relevant variables identified when producing a summarisation environment are: text related (text type, text present/absent, theme and text length), discourse related (adequacy, coherence, cohesion, use of language and comprehension), learner related (learner level and learner's prior knowledge) and aid tools (dictionaries, spell and grammar check, etc.). The above variables have been identified after a deep analysis of both summary grading and human summary grading performance to model their criteria [15].

In the context of this work, global summary grading decisions are gained by means of a Bayesian Network model based on discourse measures on cohesion, coherence, comprehension, adequacy and *use of language* as independent measures [15].

Comprehension and coherence are modelled based on Latent Semantic Analysis [16] and cohesion, adequacy and *use of language* are computed based on surface measures gathered from tagged text and statistical analysis. The present study describes the procedure followed searching for the best available approach to model *use of language* grading of learner summaries written in Basque language. The paper is organized as follows. Section 1 describes the premises in *use of language* and error detection. Section 2 analyses how automatic measures predict *use of language* by means of a multiple linear regression model. Section 3 presents a validation experiment which takes a corpus based approach to analyse how sensitive the model is discriminating *use of language* in different maturity summaries. Finally, Section 4 presents conclusions and future work.

1. Grading use of language

Grades in *use of language* have been computed based on error diagnosis and error relevance measurement. Cassany [17] states the relevance of the amount of orthographic, syntactic and lexical errors for *use of language* grading purposes. In addition, he proposes a grading scheme based on the amount of errors diagnosed in *use of language*. Cassany weights the presence of those errors considering the next 0 to 5 grading scheme: 0 errors 5 points, 1-3 errors 4 points, 4-6 errors 3 points, 7-10 errors 2 points, 11-15 errors 1 point, more than 16 errors 0 points. This grading scheme will be taken into account in automatic Basque error measures. Lexical error diagnosis has been measured using an orthography check tool for Basque [18] and an error diagnosis tagger available through text parsing [19]: X1 *Basque spell-checker*, X2 *Basque spell-checker + lexicon,* X3 *Basque spell-checker + lexicon + rubric*, X4 *Basque spell-checker+lexicon checker proportion*, X5 *Error tagger*, X6 *ErrorTagg+rubric*, X7, *ErrorTagg proportion*. In addition, *use of language* has also been graded using measures on structure and shallow punctuation error diagnosis: X8 average sentence per paragraph, X9 number of paragraphs, X10 number of sentences, X11 average words per sentence, X12 average comma per sentence, X13 metrics on how single comma measures deviate from the central word mean tendency per sentence, X14 excessive use of commas and X15 amount of commas.

2. Use of language grading model

2.1 ***Procedure***

17 human experts were asked to grade *use of language* on a 1 to 10 scale on 17 summaries written in Basque. The summaries had previously been gathered from university students, second language learners (L2) and primary and secondary school pupils (17x17 grades). The goal was to obtain a wide range of different scenarios involving *use of language* in summarisation. Each grader evaluated every summary gaining an agreement of $r = 0.75$ and $p < 0.05$. The task for expert graders consisted on reading the text based on which the summaries were written, read each summary and produce global grades on *use of language*. Automatically computed *use of language* measures were compared to graders' mean scores.

2.2 ***Results***

A multiple linear regression analysis was run to observe which of the previously described measures could best predict human *use of language* grading. The best predictive model was obtained with error tagging *(X5)* and excessive use of commas *(X14)* as predictors, significantly disambiguating more than half of the total grading variance *(*$R^2 = 0.51$, $F(2, 13) = 6.964$, $p = 0.008796$*),* a large effect size $f^2 = 0.71$ [21] and post hoc power $1\text{-}\beta = 0.8013$.

Error tagging (ETG_i) showed a β_1=- 0.44 t=-3.33 and p = 0.0054 and excessive use of commas *(UC_i)* showed a β_2=-0.45 t=-1.92 and p = 0.077. Therefore, the *use of language* grades, $Grade_i$, is obtained based on the equation:

$$Grade_i = \beta_0 + \beta_1 ETG_i + \beta_2 UC_i + \varepsilon_i$$

2.3 ***Discussion***

Among the available methods two of the variables were most salient predicting more than half of the total variances involved in grading *use of language*. However, the method shows still ambiguity which has not been resolved by $Grade_i$. Authors believe that including error diagnosis methods such as subject-verb agreement diagnosis, grammar check, a more in depth orthography check and anaphora resolution could help to obtain better prediction measures. It might also allow further corpus based grammar error detection possibilities (*e. g.* [20]). Nonetheless, in comparison to humans, $Grade_i$ has been capable to gain what is considered a large effect size in Behavioural Sciences [21].

3. Learner Summary Corpus

From the psychological point of view, a key issue in summary assessment is the difference between a poor and a good summary. Many researchers in this area make a clear distinction between immature and mature summarisers. But, how do we identify a mature or an immature summary? Garner [2] distinguishes between low-efficient and high-efficient summarisers. She argues that high-efficient summarisers recognise true information that does not appear in the source text in a higher proportion than low efficient summarisers. Therefore, immature summarisers' difficulties are mainly related to comprehension and remembering. Students have great difficulties differentiating super-ordinate from subordinate information [22]. Language proficiency is a relevant point that can make the difference between a mature and an immature summary. For instance, second language (L2) summarisers are faced with comprehension failure and lack of grammar and lexicon knowledge that summarisers do not have in their first language. Although the final result is similar to a monolingual immature summariser's the reasons behind have shown to be different [23-24]. Overall deficits were that relevant information was less well selected; a less efficient language processing and the poorer role of language on summarisation and recall. Based on these premises, authors expected that the $Grade_i$ measure should be able to significantly discriminate between first language (L1) summarisers and L2 learners. Then, L2 learners should obtain significantly lower *use of language* error grade than L1 learners. On the other hand, considering differences found in overall summarisation assessment between mature and immature summarisers, differences might also appear in *use of language* grading.

3.1 ***Procedure***

A corpus was developed distinguishing summarisation expertise and language proficiency. The corpus (*Learner summary corpus*) was compound by a total of 37 summaries −14 L2 learner summaries (7029 words), 11 primary and secondary education learner summaries (934 words) and 12 university student summaries (4762 words). Finally, all the summaries were automatically processed to gain $Grade_i$.

*3.2 **Results***

A one way analysis of variance was run with the aim to observe if $Grade_i$ measures were capable to discriminate summariser writing maturity levels. The $Grade_i$ method identified significant *use of language* differences between the three different summariser maturity groups; $F(2, 34) = 20.509$ and $p < 0.001$, post hoc power $1\text{-}\beta=0.99$. In order to observe the specific maturity level differences, a Tukey's Honestly Significant Difference (HSD) post hoc mean comparison analysis was applied. Significant differences in *use of language* grades were found between L2 and L1 summaries. L2 learners showed significantly ($p < 0.001$) lower measures than primary and secondary, Hedge's $g = 2.44$, and university learners, Hedge's $g = 1.79$. No significant differences were found between L1 summaries; however, a medium effect size [21] Hedge's $g = 0.65$ was found.

*3.3 **Discussion***

$Grade_i$ was capable to discriminate summary maturity differences in *use of language*. There was a clear differentiation between L1 and L2 learners, which is consistent with [23-24]. Nonetheless, differences between mature (university students) and immature summarisers (primary and secondary education) were not found to be significant in this data. Still, the obtained medium effect size shows that there are differences in lexical error and orthography between mature and immature summarisers.

4. Conclusions

From its early beginnings automatic evaluation of summaries written by students is an open issue in the area of Computer Supported Learning Systems. Nowadays, and due mainly to advances produced in the area of NLP techniques and cognitive modelling, it is possible to face this difficult task. LEA is an automatic summary evaluation grading environment developed with the aim of automatically grading students' summaries. Among other relevant variables LEA considers summary related discourse measuring variables such as adequacy, coherence, cohesion, and comprehension. The concrete goal of the study presented along this paper has been the treatment of *use of language,* with a twofold goal: to find the best currently available method for *use of language* grading of learner summaries written in Basque on one hand, and to test the efficiency of the method on the other. First, a *use of language* grading model has been obtained; regression modelling showed that *lexical error tagging* and *excessive use of commas* measures best predict human grading. The model disambiguated more than half of the total ambiguity resolved by human graders and gains a large effect size. In addition, an experiment has been carried out to test the accuracy of the method. In future work, authors expect to increase the proportion of disambiguation including further grammar error diagnosis as part of the model.

Acknowledgements

This work has been partially supported by the Spanish Ministry of Education (TIN2009-14380), the University of the Basque Country UPV/EHU (UE09/09) and the Basque Government (IT421-10).

References

[1] Bartlett, F. C. 1932. Remembering; a study in experimental and social psychology. Cambridge University Press.

[2] Garner, R. 1982. Efficient text summarization. Costs and benefits. *Journal of Educational Research*, 75(5):275–279.

[3] Kintsch, W., Patel, V., & Ericsson, K. 1999. The role of long-term working memory in text comprehension. *Psychologia*, 42, 186–198.
[4] Clancey, W. J. 1982. Tutoring rules for guiding a case method dialogue. In D. Sleeman & J. S. Brown (Eds.), *Intelligent tutoring systems*, 201–226.
[5] Ford, L. 1988. The appraisal of an icai system. In *Artificial intelligence and human learning*, London: Chapman and Hall, Ltd., 109–123.
[6] Woolf, B. P. 1988. Representing complex knowledge in an intelligent machine tutor. In J. Self (Ed.), *Artificial intelligence and human learning*, London: Chapman and Hall, Ltd., 3–28.
[7] Winkels, R., & Breuker, J. 1989. Discourse planning in intelligent help systems. In C. Frasson & G. Gauthier (Eds.), *Intelligent tutoring systems: At the crossroad of artificial intelligence and education*, Norwood, New Jersey: Abblex Publishing Corporation, 124–139.
[8] Díaz de Ilarraza, A. 1990. Gestión de diálogos en lenguaje natural para un sistema de enseñanza inteligente. *Unpublished doctoral dissertation*, University of the Basque Country. In Spanish.
[9] Khuwaja, R. A., Evens, M. W., A., M. J., & A., R. A. 1994. Architecture of CIRCSIM-tutor. In Proceedings of the 7th annual ieee computer-based medical systems symposium.
[10] Schulze, K. G., Shelby, R. N., Treacy, D., Wintersgill, M. C., VanLehn, K., & Gertner, A. 2000. Andes: A coached learning environment for classical newtonian physics. *The Journal of Electronic Publishing,* 1 (6).
[11] Graesser, A., Person, B., & Harter, D. 2001. Teaching tactics and dialog in autotutor. *International Journal of Artificial Intelligence in Education*, 12, 257–279.
[12] Zinn, C., Moore, J. D., & Core, M. G. 2002. A 3-tier planning architecture for managing tutorial dialogue. In S. A. Cerri, G. Gouardéres, & F. Paraguau (Eds.), Proceedings of the 6th international conference on Intelligent Tutoring Systems Biarritz, France and San Sebastian, Spain: Springer-Verlag, 574-584.
[13] Robertson, J., & Wiemer-Hastings, P. 2002, June. Feedback on children's stories via multiple interface agents. In S. A. Cerri, G. Gouardéres, & F. Paraguau (Eds.), Proceedings of the 6th international conference on Intelligent Tutoring Systems Biarritz, France and San Sebastian, Spain: Springer-Verlag, pp. 923–932.
[14] Zipitria, I., A. Arruarte, and J. A. Elorriaga. 2008a. LEA: Summarization web environment based on human instructors' behaviour. In *Proceedings of 8th International Conference of Advanced Learning Technologies*, pages 564–568.
[15] Zipitria, I., P. Larrañaga, R. Armañanzas, A. Arruarte, and J. A. Elorriaga. 2008b. What is behind a summary-evaluation decision? *Behavior Research Methods*, 40(2):597–612,May.
[16] Zipitria, I., A. Arruarte, and J. A. Elorriaga. 2006. Observing lemmatization effect in LSA coherence and comprehension grading of learner summaries. In Ashley, K. and M. Ikeda, editors, *Proceedings of Intelligent Tutoring Systems*, Lecture Notes in Computer Science, Jhonghli, Taiwan. Springer,
[17] Cassany, Daniel. 1993. *Did´actica de la correcci´on de lo escrito*, volume 108 of *Serie lengua*. Editorial Gra´o, de IRIF SL, Spain. In Spanish.
[18] Aduriz, I., I. Alegria, X. Artola, N. Ezeiza, K. Sarasola, and M. Urkia. 1997. A spelling corrector for Basque based on morphology. *Literary and Linguistic Computing*, 12(1):31–38.
[19] Aduriz, I., M. Aranzabe, J. Arriola, A. Diaz de Ilarraza, K. Gojenola, M. Oronoz, and L. Uria. 2004. A cascaded syntactic analyser for basque. In *Proceedings of Computational Linguistics and Intelligent Text Processing*, pages 124–135.
[20] Leacock, C. and M. Chodrow. 2003. Automated grammatical error detection. In Shermis, Mark D. and Jill C. Burstein, editors, *Automated essay scoring: a cross-disciplinary perspective*, pages 186–197. Lawrence Erlbaum Associates.
[21] Cohen, J. 1988. *Statistical Power Analysis for the Behavioral Sciences*. Lawrence Erlbaum Associates.
[22] Taylor, B. M. 1982. Text structured and children's comprehension and memory for expository material. *Journal of Educational Psychology*, 74:323–340.
[23] Long, J. and E. Harding-Esch, 1978. Language Interpretation and Communication, pages 273–287. Plenum Press.
[24] Kozminsky, E. and N. Graetz. 1986. First vs second language comprehension: some evidence from text summarizing. *Journal of Research in Reading*, 9(1):3–21.