Natural Language Processing in Support of Learning: Metrics, Feedback and Connectivity (NLPsL'09)

Philippe Dessus, Stefan Trausan-Matu, Peter van Rosmalen, Fridolin Wild

AIED 2009: 14th International Conference on Artificial Intelligence in Education

Workshops Proceedings

Editors and Co-Chairs:

Scotty D. Craig
University of Memphis, USA

Darina Dicheva
Winston-Salem State University, USA

July 6-7th, 2009
Brighton, UK

Preface

The supplementary proceedings of the workshops held in conjunction with AIED 2009, the fourteen International Conference on Artificial Intelligence in Education, July 6-7, 2009, Brighton, UK, are organized as a set of volumes - a separate one for each workshop.

The set contains the proceedings of the following workshops:

- **Volume 1: The 2nd Workshop on Question Generation**
  Co-chairs: Vasile Rus & James Lester. *University of Memphis, USA & North Carolina State University, USA*.
  http://www.questiongeneration.org/AIED2009/
- **Volume 2: SWEL'09: Ontologies and Social Semantic Web for Intelligent Educational Systems**
  Co-chairs: Niels Pinkwart, Darina Dicheva & Riichiro Mizoguchi. *Clausthal University of Technology, Germany; Winston-Salem State University, USA & University of Osaka, Japan*.
  http://compsci.wssu.edu/iis/swel/SWEL09/index.html
- **Volume 3: Intelligent Educational Games**
  Co-chairs: H. Chad Lane, Amy Ogan & Valerie Shute. *University of Southern California, USA; Carnegie Mellon University, USA & Florida State University, USA*.
  http://projects.ict.usc.edu/aied09-edgames/
- **Volume 4: Scalability Issues in AIED**
  Co-chairs: Lewis Johnson & Kurt VanLehn. *Alelo, Inc., USA & Arizona State University, USA*.
  http://alelo.com/aied2009/workshop.html
- **Volume 5: Closing the Affective Loop in Intelligent Learning Environments**
  Co-chairs: Cristina Conati & Tanja Mitrovic. *University of British Columbia, Canada & University of Canterbury, New Zealand*.
  http://aspire.cosc.canterbury.ac.nz/AffectLoop.html
- **Volume 6: Second Workshop on Culturally-Aware Tutoring Systems (CATS2009): Socio-Cultural Issues in Artificial Intelligence in Education**
  Co-chairs: Emmanuel G. Blanchard, H. Chad Lane & Danièle Allard. *McGill University, Canada; University of Southern California, USA & Dalhousie University, Canada*.
  http://www.iro.umontreal.ca/~blanchae/CATS2009/

- **Volume 7: Enabling Creative Learning Design: How HCI, User Modelling and Human Factors Help**
  Co-chairs: George Magoulas, Diana Laurillard, Kyparisia Papanikolaou & Maria Grigoriadou. *Birkbeck College, University of London, UK; Institute of Education, UK; School of Pedagogical and Technological Education, Athens, Greece & University of Athens, Greece*.
  https://sites.google.com/a/lkl.ac.uk/learning-design-workshop/Home

- **Volume 8: Towards User Modeling and Adaptive Systems for All (TUMAS-A 2009): Modeling and Evaluation of Accessible Intelligent Learning Systems**
  Co-chairs: Jesus G. Boticario, Olga C. Santos and Jorge Couchet, Ramon Fabregat, Silvia Baldiris & German Moreno. *Spanish National University for Distance Education, Spain & Universitat de Girona, Spain***.**
  https://adenu.ia.uned.es/web/es/projects/tumas-a/2009

- **Volume 9: Intelligent Support for Exploratory Environments (ISEE'09)**
  Co-chairs: Manolis Mavrikis, Sergio Gutierrez-Santos & Paul Mulholland. *London Knowledge Lab, Institute of Education/Birkbeck College, University of London, UK & Knowledge Media Institute and Centre for Research in Computing, The Open University, UK*.
  http://link.lkl.ac.uk/isee-aied09

- **Volume 10: Natural Language Processing in Support of Learning: Metrics, Feedback and Connectivity**
  Co-chairs: Philippe Dessus, Stefan Trausan-Matu, Peter van Rosmalen & Fridolin Wild. *Grenoble University, France; Politehnica University of Bucharest; Open University of the Netherlands & Open University, United Kingdom*.
  http://webu2.upmf-grenoble.fr/sciedu/nlpsl/

While the main conference program presents an overview of the latest mature work in the field, the AIED2009 workshops are designed to provide an opportunity for in-depth discussion of current and emerging topics of interest to the AIED community. The workshops are intended to provide an informal interactive setting for participants to address current technical and research issues related to the area of Artificial Intelligence in Education and to present, discuss, and explore their new ideas and work in progress.

All workshop papers have been reviewed by committees of leading international researchers. We would like to thank each of the workshop organizers, including the program committees and additional reviewers for their efforts in the preparation and organization of the workshops.

July, 2009
Scotty D. Craig and Darina Dicheva

AIED 2009 Workshops Proceedings
Volume 10

Natural Language Processing in Support of Learning:
Metrics, Feedback and Connectivity

Workshop Co-Chairs:

Philippe Dessus
University of Grenoble, France

Stefan Trausan-Matu
"Politehnica" University of Bucharest, Romania

Peter van Rosmalen
OUNL, the Netherlands

Fridolin Wild
Open University, United Kingdom

http://webu2.upmf-grenoble.fr/sciedu/nlpsl/

Preface

In AIED research, providing feedback for learning entails measuring differences among learners; between learners and their desired characteristics (e.g., knowledge, competences, motivation, self-regulation processes); or between learners and their looked-for resources (e.g., web-links, articles, courses) has often been performed by computing and analysing 'distances' using several techniques like factorial analysis, instance-based learning, clustering, and so on. Corpora on which these measures are made are all writing-based, that is, are multiple forms of pieces of evidence such as texts read (written by teachers), spoken utterances, essays, summaries, forum or chat messages. Some of these metrics are based on shallow syntactical and morphological aspects of the interaction and production artefacts (e.g., text length). Others are focused more on semantic and pragmatic aspects. These measures are used for providing various kinds of feedback for supporting learning and connections between learners. For instance, relations between learners' utterances, knowledge, concept acquisition, emotional states, essay scores, and even learners themselves have all been investigated with the help of computing semantic distances.

The purpose of this workshop is to focus on the latter two – semantics and pragmatics – by trying to identify what questions and problems are solved, but also to raise and discuss how well the metrics developed assist in the provision of support and the construction of feedback for learning. What are the most efficient ones? To what extent do they match distances inferred by teachers' assessments?

Presentations on topics like the following ones will fuel the research on NLP in support of learning: automated essay scoring and grading, summarization and writing assistance, methodological issues of distance-based semantic processing techniques, cognitive modelling using distance-based semantic processing techniques, analysis, assessment, and feedback generation of content and inter-animation in CSCL through chats or forums.

July, 2009

Philippe Dessus, Stefan Trausan-Matu, Peter van Rosmalen and Fridolin Wild

Program Committee

Table of Contents

Making Use of Language Technologies to Provide Formative Feedback

Adriana J. BERLANGA[a,1], Francis BROUNS[a], Peter VAN ROSMALEN[a], Kamakshi RAJAGOPAL[a], Marco KALZ[a], Slavi STOYANOV[a]
[a] *Centre for Learning Sciences and Technologies, Open Universiteit Nederland*

Abstract. This paper presents an ongoing research towards the use of Language Technologies to provide lifelong learners with formative feedback. To this end, the paper briefly elaborates the theoretical background of conceptual development and existing Language Technology applications that can be used to identify and approximate learner's conceptual development. It also presents preliminary results of proof of concept tests conducted to demonstrate the use of tools for diagnosing conceptual development and the generation of an expert-model. Finally, the paper provides initial findings towards the design of a conceptual development service.

Keywords. Formative Feedback, Language Technologies, Leximancer, Pathfinder, Expert Model, Conceptual Development

Introduction

As any learner, lifelong learners need to receive feedback on how they are developing their knowledge on the topic of study. Lifelong learners, however, are heterogeneous: they differ on their learning goals, profile, knowledge, and learning paths. This diversity increases the complexity and time required to provide formative feedback: tutors need to position every learner in the curriculum and assess (almost in an individual basis) how she is developing her knowledge. From our point of view, formative feedback can be (semi-)automated using Language Technologies [1, 2].

In the context of the Language Technologies for Lifelong Learning (LTfLL) project we explore how Language Technologies can be used to provide lifelong learners with formative feedback on their conceptual development and with support to overcome conceptual gaps. We hold that a learner's conceptual development can be diagnosed by comparing the manner in which the learner organizes and structures the domain knowledge with how an expert does this.

This assumption is based on research on expertise that has shown differences in the knowledge base development from novice to expert [3]. According to [4] experts and novices differ in their knowledge usage, information processing, and organizing of their knowledge structures. Experts distinguish better between relevant and non-relevant information than novices, who tend to reason on both relevant and irrelevant information [5]. Experts have elaborated, well structured and organized mental frameworks that activate to interpret information and problems and to create a suitable

[1] Corresponding Author: Adriana J. Berlanga, Centre for Learning Sciences and Technologies, Open Universiteit Nederland, PO Box 2960, 6401 DL Heerlen, the Netherlands; E-mail: adriana.berlanga@ou.nl

solution [3, 6], whereas novices do not easily activate their mental frameworks, which are less accurate, complete, organized and structured [7]. Findings in Law [7], Physics [8], Management [4], and Medicine [9] have shown that knowledge is more hierarchically structured with increasing expertise, while novices' knowledge appears to be highly fragmented and concepts loosely connected.

For our research, therefore, we have to use and compare to an "expert model" that is not absolute; it develops as it does in practice [4, 7-9]. We use the term to define the expected set of concepts and relations that represent the domain of knowledge at a specific point in time of the development of a learner.

Others indicate the expert model in advance [10], or include a phase of sampling and negotiating amongst participants and peers which concepts the expert model should have [11]. In our work we go beyond these approaches by deriving the expert model (semi-)automatically. There are three different types of expert model that can underlay this.

1. *Archetypical expert model*; considers expert and state-of the art information (e.g. scientific literature).
2. *Theoretical expert model*; considers particular information (e.g. course material, tutor notes, relevant papers, etc.).
3. *Emerging expert model*; considers the concepts and the relations between those concepts that a group of people (e.g. peers, participants, co-workers, etc.) used most often.

In this paper we concentrate on the theoretical and emerging approaches to identify or approximate the conceptual development of learners and the role of Language Technology tools in this. Next, we explain how existing applications and tools, namely Leximancer [12] and Pathfinder [13], have been used in two different preliminary explorations as proof of concept of the suitability of these approaches. In the final section, we provide initial recommendations for the design of a conceptual development service.

1. Investigating How Formative Feedback Can Be Provided

In order to assess the individual's knowledge of a particular domain, [14] propose a structural approach to determine how the individual organizes the concepts of such a domain. This approach involves three steps: knowledge elicitation, knowledge representation, and evaluation of the representation.

1. *Knowledge elicitation* techniques measure the learner's understanding of the relationships among a set of concepts [15]. Methods that support this activity include card sorting, concept maps, think aloud, or essay questions.
2. *Knowledge representation* reflects the underlying organization of the elicited knowledge [14]. Advanced statistical methods (e.g. cluster analysis, tree constructions, dimensional representations, pathfinder nets) are used to identify the structural framework underlying the set of domain concepts.
3. *Evaluation of the representation* relative to some standard (e.g. expert's organization of the concepts in the domain) using one of the following approaches [14]: qualitative assessment of derived representations; quantifying the similarities between a student representation and a derived structure of the content of the domain; or comparing the cognitive structures of experts and novices.

A data collection protocol was defined to elicit and represent a learner's knowledge. This protocol combines a think aloud procedure with a cognitive map method to provide a suitable and appropriate measure of the learner's representation of the subject matter structure. Concept maps, furthermore, are one of the most common ways of representing cognitive structures. Research evidence demonstrates the appropriateness of concept maps in eliciting knowledge [16] and their superiority for evaluation of learners of different ages compared to classical assessment methods such as tests and essays [17, 18].

There are already a number of tools for the automatic construction and support of concept maps: Knowledge Network Organizing Tool (KNOT, PFNET) [19]; Surface, Matching and Deep Structure (SMD) [20]; Model Inspection Trace of Concepts and Relations (MITOCAR) [21]; Dynamic Evaluation of Enhanced Problem Solving (DEEP) [22]; jMap [23], Leximancer [12], and ProDaX [24] (for a comparison see [1]).

A number of these tools (Pathfinder, Leximancer, Infomap, jMap, MITOCAR, KNOT, and ProDaX) have been explored. Giving the results of this exploration, Leximancer and Pathfinder have been selected for a further proof of concept. Leximancer generates concept maps from a document collection using content analysis (based on co-occurrence) and relational analysis (proximity and concept mapping). Pathfinder can be used to derive and visualize structured (semantic) networks. It is based on proximity measures (similarity, correlations, distances, probability) between pairs of concepts [25].

As a proof of concept these tools have been explored in two different ways. In the first one, a so-called theoretical expert model was identified (considering course and tutor materials) and compared with the concept map of a student. For this purpose, a combination of Leximancer and Pathfinder was used. The second proof of concept, in which only Leximancer was used, explored the generation of an expert model identifying the concepts and relations mentioned by participants in a small-scale pilot. The rest of this paper elaborates further on these explorations.

2. Leximancer and Pathfinder: Generation of a Theoretical Expert Model

An initial exploration has been conducted on how formative feedback could be provided within the formal curriculum of the Manchester Medical School. To this end the following procedure, based on the structural approach described earlier, was defined:

1. *Knowledge elicitation*: The data collection protocol to elicit students' knowledge was used. Next, the think aloud protocols were transcribed.
2. *Knowledge representation*: Leximancer was used to generate concept maps for novices –derived from student-generated think aloud– as well as a theoretical expert concept map –derived from tutor notes and supporting materials–. Next, a correlation matrix of concepts was exported.
3. *Evaluation of the representation*: Pathfinder was used to compare the cognitive structures of the novices and theoretical expert concept map, and identify similarities and differences.

> The cytotoxic P cells are responsible for killing the microorganisms and it's triggered by the binding of TCR to the MAC protein complex, bound to the specific antigen, the antigen peptide fragments, the T helper cells or the CD 4 T cells are essential for the cell-mediated response. They make cytokines for delayed hypersensitivity and help making B cells specific for antigens. T-regulator cells play a role in the negative regulation of the immune system.

Figure 1. Part of transcribed student think aloud

2.1. Procedure

The protocol of data collection was used with first year students of Manchester Medical School. The curriculum is designed according the problem-based-learning approach. Students do not always receive timely feedback or individual feedback. That makes it difficult for them to judge whether they are on track. Students receive lecture notes and a case description. During the think aloud sessions, students were asked to talk about a case they just studied. The sessions were transcribed (see Figure 1 for an example transcription). The transcriptions were used to generate a Leximancer concept map for the students. Similarly, the tutor notes and supporting material were used to derive the theoretical expert model. Figure 2 depicts the concept map for the student (left) and the theoretical expert model (right). The interpretation of both concepts maps is given in the next section. Next, the concept maps were exported as a co-occurrence matrix, which provides the relevance scores for the nodes. These relevance scores represent the conditional probability of co-occurrence for a concept. It is a measure of co-occurrence of two concepts as a proportion of occurrence of the selected concept.

First we determined whether the exported co-occurrence matrix could be transformed to a Pathfinder data format, and whether this resulted in a comparable representation of the concepts. To facilitate this process, only the five most used concepts of the Leximancer concept maps for the theoretical expert model and one of the students were exported (see Figure 3 for an example). This was manually transformed into a Pathfinder data format. Best results for these small networks were obtained with the probability data format and with default settings for the parameters.

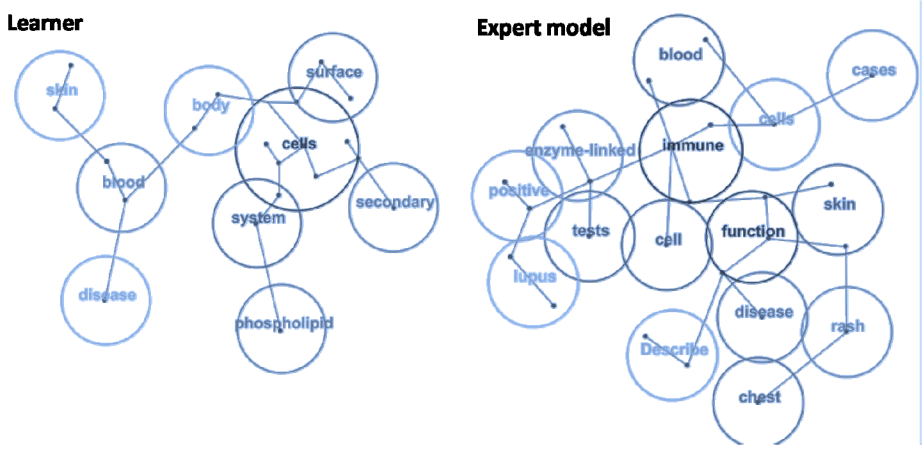

Figure 2. Concept map for a student (left) and the theoretical expert model (right) (Leximancer)

```
<entity colour="#ffffff" freq="21" id="0" kind="WORD" linksVisible="false"
value="cells" visible="true" x="6.94646429487698" y="17.541484109122838">
<relEnts>
  <relEnt id="1" str="0.61904764" />
  <relEnt id="2" str="0.33333334" />
  <relEnt id="3" str="0.23809524" />
  <relEnt id="4" str="0.0952381" />
</relEnts>
</entity>
```

Figure 3. Leximancer matrix XML export

The resulting Pathfinder networks, although not identical, resembled the Leximancer concept maps. Leximancer only allows users to visually inspect concept maps, while Pathfinder can depict and calculate similarities and differences between the student concept map and the theoretical expert model. Figure 4 depicts similarities and differences in the maps of the student and the expert model.

2.2. Initial findings

As initial verification, the Leximancer generated concept maps and the comparison produced in Pathfinder were discussed with an expert. The concept maps of the students and of the theoretical expert model differ on the level of detail. Whereas the student concept map included detailed concepts, the theoretical expert model encapsulated the concepts and gave the panoramic view of the knowledge (as can bee seen in Figure 2 and Figure 4). Interestingly, this suggests that even if the learning material explains the reasons and conditions of a problem ("the why"), novice students represent their understanding by indicating only procedural knowledge, mentioning how to solve a problem ("the how"). This suggests that the tutor notes and learning materials might not be ideal to generate an expert model. The materials are written from a perspective that requires more expertise than the novice student can achieve at that point of time. Consequently, this might not be a good basis for deriving the theoretical expert model, nor for providing formative feedback.

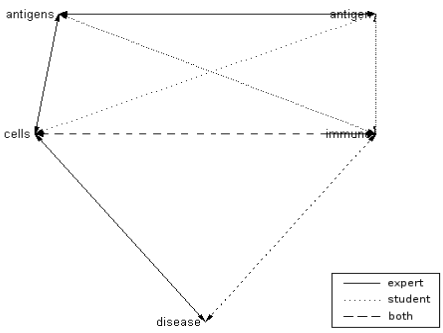

Figure 4. Example of a comparison of a student and expert concept map (Pathfinder)

3. Leximancer: Generation of an Emerging Expert Model

In addition a second proof of concept was conducted. The aim was to test how Leximancer could be used to provide formative feedback to employees in an informal learning situation. To this end the following procedure was defined:

1. *Knowledge elicitation*. The data collection protocol to elicit employees' knowledge was used. Next, the think aloud protocols were transcribed.
2. *Knowledge representation*. The emerging expert model was generated as a single Leximancer concept map based on the transcripts of all think aloud protocols. In addition, concept maps for every speaker were generated.
3. *Evaluation of the representation*. Leximancer was used to compare the cognitive structures of experts and novices, and to identify similarities and differences.

3.1. Procedure

The protocol of data collection was used with employees (n=10) of the Open Universiteit Nederland. They were asked to reflect on the concept Learning Networks (i.e., online social networks where the participants organize their own learning process in line with their needs for competence development), which is the topic of research conducted within the university. Therefore it can be considered as knowledge that is learned and developed at the work place, an informal learning situation.

The sessions were transcribed and coded in a way that Leximancer recognized as interviews. The emerging expert model was derived from a single Leximancer concept single map based on all transcripts (see Figure 5).

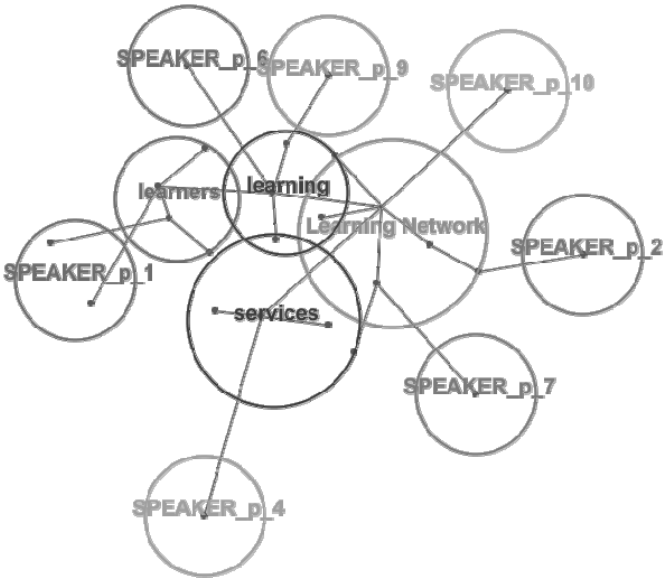

Figure 5. Example of an emerging expert map (Leximancer)

Leximancer discovered the 10 most used concepts and their relevance automatically: learning (47% relevant); services (45%); people (34%); learners (27%), resources (17%); community (17%); social support (15%); participants (12%); course (12%); content (12%). The tool also depicts the relations of each concept with other concepts. Figure 5 depicts the emerging expert model for the concept *Learning* Networks as it arises from all concepts and the relations between concepts. It also visualizes the position of the individual speakers in relation to the model, by indicating which concepts the speaker mentioned.

Further, a concept map was generated for individual employees for whom the 10 most used concepts were identified. These were compared to identify similarities and differences between the emerging expert model and employees' concept maps. It seems feasible to generate individual formative feedback reports that present differences and similarities. Future work involves validation of the reliability of the emerging expert map and the formative feedback report.

4. Conclusions and Discussion

This paper presented our current research in the area of (semi-)automated formative feedback for learners with the help of Language Technologies. To this end, the paper presented two approaches of how Language Technologies can be used and discussed conceptual and technical implications.

There are several ways to generate expert models. We concentrated on two approaches: the theoretical expert model and the emerging expert model. Conceptually, the first approach seems to provide little information to generate a formative feedback report, since the theoretical information is written in a way that might be at a "too high level" for a student at a specific point of time. The second approach, the emerging expert model, seems to solve this issue. The set of concepts that is used by most people at a specific point of time might provide better evidence of the level of abstraction and relations between concepts. This approach, however, will require a better appreciation of the learner's knowledge representation –by contextualizing both the learner's knowledge and the situation in which the knowledge will be applied– and requires mechanisms to keep the model updated.

Acknowledgements

We would like to thank Manchester Medical School for their participation in this work. The work presented in this paper was carried out as part of the LTfLL project, which is funded by the European Commission (IST-2007-212578).

References

[1] A. J. Berlanga, M. Kalz, S. Stoyanov, P. Van Rosmalen, A. Smithies, and I. Braidman, "Using Language Technologies to Diagnose Learner´s Conceptual Development," in *The 9th IEEE International Conference on Advanced Learning Technologies*, Riga, Latvia, 2009. In press.

[2] M. Kalz, A. J. Berlanga, P. Van Rosmalen, S. Stoyanov, J. Van Bruggen, and R. Koper, "Semantic Networks as Means for Goal Directed Formative Feedback," in *Kreativität und Innovationskompetenz*

im digitalen Netz - Creativity and Innovation Competencies in the Web, Sammlung von ausgewählten Fach- und Praxisbeiträgen der 5. EduMedia Fachtagung, Salzburg, Austria, 2009 pp. 88 – 95.

[3] H. P. A. Boshuizen and H. G. Schmidt, "On the role of biomedical knowledge in clinical reasoning by experts, intermediates and novices," *Cognitive Science,* vol. 16, pp. 153 – 184, 1992.

[4] A. J. Arts, W. H. Gijselaers, and H. Boshuijzen, "Understanding managerial problem-solving, knowledge use and information processing: Investigating stages from school to the workplace," *Contemporary Educational Psychology,* vol. 31, pp. 387 – 410, 2006.

[5] H. P. A. Boshuizen and H. G. Schmidt, "The development of clinical reasoning expertise: Implications for teaching," in *Clinical reasoning in the health professions*. vol. 3rd rev. ed., J. Higgs and M. Jones, Eds. Oxford: Butterworth-Heinemann, 2008.

[6] M. van de Wiel, "Knowledge Encapsulation. Studies on the development of medical expertise," Unpublished PhD dissertation. University of Maastricht, 1996.

[7] F. Nievelstein, T. Van Gog, H. P. A. Boshuizen, and F. J. Prins, "Expertise-related differences in ontological and conceptual knowledge development in the legal domain," *European Journal of Cognitive Psychology*, vol. 20, pp. 1043 – 1064, 2008.

[8] R. J. Dufresne, W. J. Gerace, P. Thibodeau-Hardiman, and J. P. Mestre, "Constraining novices to perform expertlike problem analysis: Effects on schema acquisition," *The Journal of the Learning Sciences,* vol. 2, pp. 307 – 331, 1992.

[9] M. W. J. van de Wiel, H. G. Schmidt, and H. P. A. Boshuizen, "Knowledge Restructuring in Expertise Development: Evidence From Pathophysiological Representations of Clinical Cases by Students and Physicians," *European Journal of Cognitive Psychology,* vol. 12, pp. 323-355, 2000.

[10] R. Clariana and P. Wallace, "A Computer-Based Approach for Deriving and Measuring Individual and Team Knowledge Structure from Essay Questions," *Journal of Educational Computing Research,* vol. 37, pp. 211 – 227, 2007.

[11] V. J. Shute, A. C. Jeong, J. M. Spector, N. M. Seel, and T. E. Johnson, "Model-Based Methods for Assessment, Learning, and Instruction: Innovative Educational Technology at Florida State University," in *2009 Yearbook Educational Media and Technology*, M. Orey, Ed.: Greenwood Publishing Group, submitted.

[12] A. E. Smith and M. S. Humphreys, "Evaluation of unsupervised semantic mapping of natural language with Leximancer concept mapping," *Behavior Research Methods,* vol. 38, pp. 262 – 279, 2006.

[13] R. W. Schvaneveldt, Pathfinder associative networks: Studies in knowledge organizatioN. Nordwood,NJ: Ablex, 1990.

[14] T. E. Goldsmith, P. J. Johnson, and W. H. Acton, "Assessing structural knowledge," *Journal of Educational Psychology,* vol. 83, pp. 88 – 96, 1991.

[15] D. H. Jonassen, K. Beissner, and M. Yacci, *Structural knowledge: techniques for representing, conveying, and acquiring structural knowledge*. Hillsdale, NJ: Lawrence Erlbaum Associates, 1993.

[16] J. Nesbit and O. Adesope, "Learning with concept and knowledge maps: A meta-analysis," *Review of Educational Research,* vol. 76, pp. 413 – 448, 2006.

[17] J. D. Novak, Learning, creating and using knowledge: concept maps as facilitative tools in schools and corporations. Mahwah, NJ: Erlbaum 1998.

[18] D. Jonassen, T. Reeves, N. Hong, D. Harvey, and K. Peter, "Concept mapping as cognitive learning and assessment tools," *Journal of interactive learning research,* vol. 8, pp. 289 – 308, 1997.

[19] R. Clariana, R. Koul, and R. Salehi, "The criterion-related validity of a computer-based approach for scoring concept maps," *International Journal of Instructional Media,* vol. 33, pp. 317 – 325, 2006.

[20] D. Ifenhaler and N. M. Seel, "The measurement of change: learning-dependent progression of mental models," *Technology, Instruction, Cognition and Learning,* vol. 2, pp. 317 – 336, 2005.

[21] P. Pirnay-Dummer, "Expertise and model building. MITOCAR," Freiburg: Unpublished doctoral dissertation. University of Freiburg, 2006

[22] J. M. Spector and T. A. Koszalka, "The DEEP methodology for assessing learning in complex domains," Syracuse University 2004.

[23] A. Jeong, "jMap v. 104," Florida State University, 2008.

[24] R. Oberholzer, S. Egloff, S. Ryf, and D. Läge, "Prodax. Datenauswertung im Bereich der Skalierung. Benutzerhandbuch," 2008.

[25] R. Clariana and R. Koul, "A computer-based approach for translating text into concept map-like representations," in *Proceedings of the First International Conference on Concept Mapping*, Pamplona, Spain, Sep 14-17, 2004, pp. 131– 134.

Lexical similarity metrics for vocabulary learning modeling in Computer-Assisted Language Learning (CALL)

Ismael ÁVILA[a,1] and Ricardo GUDWIN[b]
[a] *University of Campinas and CPqD Research and Development Center*
[b] *University of Campinas*

Abstract: This paper discusses a technique for measuring lexical similarity in terms of its effect on the perceptual ability of learners in recognizing L2 words with the help of L1. This technique can be used in many modules of an ITS CALL implementation, in particular in the initialization of the learner model based on his/her native language and in the diagnose of errors due to interference from L1. The rationale for such an implementation is discussed and a brief description of the technique is given.

Keywords: natural language processing, cross-linguistic influence, interference (language), learner errors (language), learner model.

Introduction

The very particular nature of second language teaching comes from the fact that the language itself is the learning goal, the main instructional resource and the key aspect defining learners' background knowledge. This contrasts neatly with other teaching areas, indicating the need for an adequate understanding of second language learning and demanding implementation techniques capable of capturing its richness. Hence, the Computer-Assisted Language Learning (CALL) field demands very specific instructional tools and strategies as well as accurate techniques for learner modeling. For instance, it is well known that the first language (L1) can create a basis for learning the vocabulary of the second language (L2), since the already acquired L1 lexicon can help the learner to infer the meanings of words in L2, most of all if both languages have lexical similarities. In order to model (qualify and quantify) this cross-linguistic influence, techniques to compare the lexical distance between L1 and L2 are required. This comparison can be done in terms of how similar is the form of semantically related words in L1 and L2, so that the ITS can know in advance which lexical units from L2 will be more easily learned due to transfers from L1 and which ones are likely to produce interferences. The ITS can use the results of this comparison to initialize the learner model or, by means of a similar technique, to continuously assess the learning process by measuring how distant the learner's answers are from the right answers.

The lexical distance can be relevant to a greater or a lesser extent depending on the adopted instructional strategy. If teachers decide to organize the lexical units based on their frequency of use, teaching first the most used words, they can rely on objective metrics that refer only to L2, and which can be established in terms of some ranking of most frequent words (for example, in everyday vocabulary, or in some particular area of interest such as business, tourism, etc.). If, on the other hand, the lexical units are to be organized in terms of their easiness for the learners, this is indeed a relative criterion that will depend on the L1(s) of the target audience(s). In this case, the easiness of each

1 Corresponding Author.

lexical item will strongly depend on its resemblance with the corresponding word in L1, and the use of some metrics for quantifying this similarity would be desirable.

In this paper we present a work in progress, in which we are applying a technique for measuring lexical similarity in terms of its effect on the perceptual ability of learners in recognizing L2 words with the help of L1.

1. Lexical distance as a predictor of transference likelihood in ITS CALL systems

In L2 learning, it is possible and even inevitable that the learner's L1 lexicon will influence the easiness she/he will have assimilating L2 vocabulary. If the involved languages are closely related, many L2 words will probably be more easily learned since they look similar to their counterparts in L1, usually because they share a same origin (cognate words). This is, for instance, the case of words such as the Spanish "corazón" and the Portuguese "coração". Such lexical similarities may occur even between not so closely related languages, such as English and French (e.g. "liberty" – "liberté") or German and French (e.g. "blau" – "bleu"). Lexical similarities may even be found in totally distant languages due to borrowings (e.g. Japanese "arigato" and Portuguese "obrigado") or to accidental coincidences (Greek "oikia" and Tupi "oca").

Regardless of the origin of these similarities, from the didactic point of view this is an aspect that impacts the entire language learning process and therefore needs to be carefully accounted for by the ITS. This implies evaluating the level of similarity, classifying its dimensions and assessing its potential effects (beneficial or detrimental): similarity it is not always a facilitating feature, since in the case of false friends it tends to induce cross-linguistic interferences rather than correct inferences (transferences).

The level of lexical similarity can be used in many modules of an ITS CALL. For example, to determine the learner's background knowledge, and then to initialize parts of the learner model. Also knowing how distant a learner's answer is from the correct answer to a question is something that can be used to quantify and qualify the learning results and, in case of discrepancies, be a clue to diagnose causes of error (interference from L1, overgeneralization, etc.). Measuring word-level dissimilarities regarding right answers or similarities to common errors is a valuable tool in educational applications.

The similarity level has two main parallel dimensions: orthographic and phonetic. Each of them may vary from a level of "no similarity" to a level of "absolute match". For instance, the English and French words "direction" share the same spelling, but somewhat distinct pronunciations (and slightly different meanings), whereas English "house" and German "Haus" present "partial orthographic match" but have similar pronunciations (and meanings). Therefore, in order to correctly evaluate the proximity between lexical units in L1 and L2, or between learner's answer and the right answer, the CALL system needs to distinguish and compare these dimensions while applying quantitative metrics of similarity.

In our ITS CALL application we employed a multidimensional similarity measure based on perceptual criteria, involving correspondences such as letter-by-letter match, same initials, equivalent consonant order and phonetic distance. The calculation of the similarity use weights to balance the influence of orthographic and phonetic features in the overall similarity and can be used in combination with AI algorithms, such as those discussed in [1], in order to classify or cluster errors in terms of their most likely causes. Our ITS CALL application is applied in a web-based language course. As the (L2) learning object of the course we chose the international language Esperanto for two reasons: (i) it has a compact lexicon; (ii) its lexicon is based on international roots. But we believe that to some extent the achieved results will be also valid for any other languages. In the next sections we discuss these implementation in detail.

3. Methods for calculation of lexical similarity

According to [1], the manipulation of symbolic data, such as words and sentences, has usually been outside the focus of the research on neural networks and related learning algorithms, which have mainly dealt with numeric data. This was due to the fact that sensory data from real world information processing are generally numeric by definition. When it comes to numerical data, the average and the similarity are easily computed in terms of arithmetical mean and inverse distance, respectively. Although, for non-numerical data, like letter strings, both measures tend to be more complicated to compute, both calculations for letter strings can also be based on a distance measure, just like their numerical counterparts, by means of techniques such as the Levenshtein or the Feature distance. Consequently, the average of a set of strings can be obtained as a string with the smallest distance from all strings in the set, whereas the similarity can be defined as the inverse or negative distance between the strings [1]. And with those two measures and substituting reference vectors by reference strings one can construct self-organizing maps of letter strings.

As pointed out by [1], a letter string cannot be represented by a numerical vector, since a coding in which numerical differences between the codes reflect dissimilarities among corresponding letters is hard to achieve, and even more difficult when one tries to compare strings of different lengths, or when one string is derived from another by insertion or deletion of letters, something that is very common in the case of cognate words in different languages.

Hence, distance measures suited for letter strings are required. One such measure is the *Levenshtein distance*, defined as the minimum number of basic transformations – insertion, deletion and substitution of letter – to transform one string into another [2]:

$$LD(s_1, s_2) = \min (n_{ins} + n_{del} + n_{subst})$$

Derived from it is the *weighted Levenshtein distance* [3], also known as edit distance [4], where different costs are assigned to each edit operation.

The *Feature distance* [4] is given by the number of features in which two strings differ. In Feature distance, N-grams (substrings of N consecutive letters) are the usual choice for features, and if one string is longer than the other, the unmatched N-grams are also counted as differences [1]:

$$FD(s_1, s_2) = \max (N_1 + N_2) – m(s_1 + s_2)$$

Where N_1 and N_2 denote the number of N-grams in strings s_1 and s_2 and $m(s_1 + s_2)$ is the number of matching N-grams [1].

The *Levenshtein distance* leads, according to [1], to slightly better classification accuracy than the *Feature distance*, but the latter allows for much faster searching. It is worth noting that these general-purpose methods are not aimed at specific applications. Thus, in some cases, betterments have been proposed to make these calculations more suited to real world problems. In [5], for instance, the authors applied Levenshtein Distance to measure language distances so as to produce phylogenetic trees of language families based on the similarities of their basic vocabularies. However, so as to account for the fact that one letter change has more relevance in short words than in long ones, the authors developed a normalized version of LD.

Regarding the use of the lexical similarity as a parameter to determine language proximity, the authors argue that the grammatical differences would be too hard to compute, and also point out that an automated method avoids the subjectivity that is inherent when these comparisons are made by humans. Subjectivity arises because

scholars tend to see similarity in remote kinship linking cognate words even when the current word forms look very different one from another, such as the Spanish word "*leche*" and the Greek "*gala*" [5]. It is worth noting that in our course we are interested in measuring effective similarity rather than in linguistic kinship, since from a didactic viewpoint, similarity, even if accidental, is what matters for learning easiness. Thus, L2 word recognition is, in such a learning context, a shortcut to vocabulary learning.

An instructional application requires similarity measures that encompass the main features that facilitate the recognition (and memorization) of a given L2 word on the basis of its alikeness with the corresponding word in L1. This measuring could involve some sort of letter-by-letter comparison, as discussed above. However, from a semiotic standpoint, the recognition of an L2 word due to its similarity to a semantically correlated L1 word is a kind of inference that is essentially based on diagrammatic (iconic) features, although both words are symbols (arbitrary signs) rather than icons. Then, in this case the similarity points from an L2 symbol (word) to a corresponding L1 symbol, contrary to ordinary icons, whose similarity (such as the picture of a car) links to physical features of an actual object. So as to emphasize the particular nature of this phenomenon we have coined the expression "intersymbolic iconicity or similarity".

As in the case discussed in [5], this requires objective criteria, based on effective similitude, rather than subjective ones, founded on remote etymological kinship. Thus, the calculation of a letter-by-letter similarity is a good starting point. Nevertheless, the evaluation of a level of similarity is not limited to an orthographic correspondence. It implies assigning more weight to key features such as correspondence of initials or coincidence in the positions of consonants, considering that the consonants in general, and initials in particular, form a diagrammatic image of any given word. This fact has a lot of support in the area of perceptual psychology, since a written or printed word is a visual stimulus in the first place [6].

According to [6], for instance, for the vast majority of people, the left hemisphere is more important than the right hemisphere for language processing, what makes the word recognition slightly easier after fixation of the leftmost than the rightmost letter of a word (in languages that are read from left to right the leftmost letter is the initial), simply because information in the right visual half-field is projected directly onto the left cerebral hemisphere whereas information in the left visual half-field requires inter-hemispheric transfer to reach the left cerebral hemisphere. Another reason for the strong word-beginning advantage in words that are read from left to right is related to the fact that fixation on the leftmost letter makes the whole word fall in the right visual half-field, which has direct connections to the dominant left hemisphere.

Word processing accuracy and speed depend on two factors: (i) perceptibility of the individual letters as a function of the fixation location and (ii) the extent to which the most visible letters isolate the target word from its competitors [6].

These word recognition factors are also applicable as a common sense technique to create word abbreviations: *tks* (*thanks*), *pg* (*page*), *cmd* (*command*) or *ctrl* (*control*). For this reason, the matching of initials and consonants is more likely to enable word recognition than matching a comparable number (i.e. same LD) of other letters without the initial or with vowels included (resp. *tak*, *ae*, *oma*, *coto*). Hence, in our technique we assign more value to the diagrammatic role of consonants than to other matchings and emphasize the function of consonants and initials, as indicated in the next section.

But these similarities can be realized also in a more phonetic level, even when the spelling rules are not equivalent (as in the case of English "physics", Czech "fyzika", Polish "fizyka", Italian "fisica", Afrikaans, "fisika" and French "physique"). According to [6], it is now clear that reading and word recognition are not simply based on orthographic information but involve the activation of phonological codes. This has

been shown, for example, by [7] and [8]. In our technique the overall similarity score combines orthographic and phonetic features. It includes a *grapheme → phoneme* conversion (normalization) prior to calculating phonetic similarity of words, since a more straightforward mechanism for computing the phonetic similarities would depend on a support for the international phonetic alphabet (IPA) in the simulation tool at hand, what is not always true.

4. Calculation of intersymbolic similarity

The calculations involved in measuring word similarity in our application attempt to capture the features that matter when a learner first encounters a new L2 lexical unit. As discussed in the previous section, the main features are:

Orthographic (in order of importance):
-Initials
-Consonants (in the order they appear)
-Vowels (in the order they appear)
Phonetic:
-Phonemes (in the order they appear)

A phoneme match implies equal pronunciation even if written with different graphemes such as "c" and "k"; phonemes are considered similar in cases such as "s" and "z", "r" and "l", etc., but the similarity will depend on the languages involved, and thus a previous mapping of phonetic correspondence between L1 and L2 is necessary.

The orthographic criteria are modulated by the phonetic ones, in such a way that, if the orthographic rules of L1 use one letter to represent the same phoneme that in L2 is represented by two or three letters (e.g. Czech "š", English "sh" and German "sch"), the phonetic matching should cause the system to treat the consonantal cluster in L2 as a surrogate for the one letter initial in L1, and vice-versa. This solution tends to be more accurate in representing the similarity perceived by learners than a letter-by-letter comparison, which, by the way, could incur distortion of the similarity measure due to the risk of comparing the final letter(s) of the consonantal cluster in L2 word with the second letter/phoneme in L1. Therefore, the first step in the method deals with the segmentation of the strings in order to establish the L1–L2 grapheme/phoneme pairs. The second step evaluates distances between paired segments. The third step calculates the total intersymbolic distance, assigning weights to the parameters in the equation so that the final result is contained between 0 (match) and 1 (no match).

The equation for intersymbolic similarity is:

$$IS = \alpha(\gamma_1 I + \gamma_2 C + \gamma_3 V) + \beta P \qquad (1)$$

Where: IS: intersymbolic similarity (maximum =1, minimum = 0)
I: initials
C: consonants
V: vowels
P: phonemes (can be decomposed as the orthographical part: $\gamma_4 I + \gamma_5 C + \gamma_6 V$)
α: weight of the orthographical similarity (adjusted according to the context)
β: weight of the phonetic similarity (adjusted according to the context)
γ_n: weights of factors of similarity (e.g. γ_1=0.4; γ_2=0.4; γ_3 =0.2)
$\alpha + \beta = 1$ and $\gamma_1 + \gamma_2 + \gamma_3 = 1$ and $\gamma_4 + \gamma_5 + \gamma_6 = 1$

Note 1: Weights of the equation are adjusted so that the maximum similarity is 1 (for totally matching words) and the minimum is 0 (for totally different words).
Note 2: Weights of the orthographic features can be adjusted to assign more relevance

to initials and consonants while preserving some of the effect of the vowels (e.g. γ_1=0.4; γ_2=0.4; γ_3 =0.2). The phonetic factors can be adjusted differently, if necessary.

Note 3: While initials are compared one-to-one, the comparisons of the consonant or vowel sequences consider letter groupings such as "cntrl" or "oo". The values assigned to each individual letter will depend on the length of the corresponding sequence in the original (L2) word. If the reference consonant sequence is, as in the example below, formed by "tmp", and the maximum similarity is valued as 1, each matching letter will be assigned the value of 0.33. Therefore, if the L1 word has the sequence "tm", the total score for consonant similarity will be 0.66. It goes without saying that the order of the letters is important. An alternative sequence such as "mt" would be valued 0 since it does not retain a diagrammatic representation of the L2 word morphology, and then would not have the same effect in facilitating word recognition. Here we think of the isolated role of these middle letters in the overall process of word recognition, in spite of the fact that the swap of middle letters does not impede the recognition of the word as a whole if the first and the last letters of the word are correct [9].

Note 4: In the comparisons, it may be necessary to normalize consonants and clusters to a same notation: for instance, "š", "**ŝ**" and "sch" to "sh". Depending on the required transformations in the normalization, different similarity values can be assigned:

-Total match = 1: Exactly the same letter(s)
-Equivalent = 0.9: Letters have closely the same function (e.g. "š" and "ŝ");
-Similar = 0.8: One letter corresponds to a consonant cluster (e.g. "š" and "sch").

Note 5: Depending on the context of the implementation, developers may neglect the phonetic similarity. In our case, however, given the multimedia nature of a Web-based course, the phonetic similarity can provide an effective basis for L2-word recognition.

Note 6: Although the final letter of a word can also play a role in its diagrammatic recognition, in our technique we decided not to emphasize final letters because in our target language the final letter is not part of the word root, but a syntactical marker. This does not preclude other developers to adapt the technique to other languages.

The algorithm for word comparison (implemented in Matlab) has the following steps:

- Identification of L1 (in order to identify the orthographic and phonetic rules)
- Segregation of initials, consonants and vowels
- Conversion of consonant clusters (normalization)
- Comparison of initials, consonants and vowels
- Calculation of the final similarity score

Obs.: All these steps were implemented as a function that can be called by other algorithms, such as AI applications for classification or clustering of data (SOM).

Example: The intersymbolic similarities of the Italian word "tempo" respectively to speakers of Portuguese, Spanish, English, German and Finnish are:

L1 (tempo)→L2 (tempo): Initials: t=t; Consonants: tmp=tmp; Vowels: eo=eo
IS = 0.6*(0.4*1+0.4*1+0.2*1)+0.4*1 = **1**
L1 (tempo)→L2 (tiempo): Initials: t=t; Consonants: tmp=tmp; Vowels: eo≈ieo
IS = 0.6*(0.4*1+0.4*1+0.2*0.66)+0.4*0.9 = **0.92**
L1 (tempo)→L2 (time): Initials: t=t; Consonants: tmp≈tm; Vowels: eo≠ie
IS = 0.6*(0.4*1+0.4*0.66+0.2*0)+0.4*0.4 = **0.48**
L1 (tempo)→L2 (Zeit): Initials: t≈Z(ts); Consonants: tmp≈Zt; Vowels: eo≈ei
IS = 0.6*(0.4*0.5+0.4*0.16+0.2*0.33)+0.4*0.2 = **0.28**
L1 (tempo)→L2 (aika): Initials: t≠a; Consonants: tmp≠k; Vowels: eo≠aia
IS = 0.6*(0.4*0+0.4*0+0.2*0)+0.4*0 = **0**

5. Experimental results

In order to evaluate the proposed technique we took the word "physics" and some of its synonyms in other languages, such as mentioned in Section 3, and compared the scores of similarity with the results produced by one of the existing distance measures, in this case the Levenshtein Distance. In LD, i = insertion, s = substitution, x = no change, one insertion counts 1, whereas one substitution counts 2 (since it means one deletion + one insertion) as follows:

Original word: "physics" transformations

to Czech "fizyka"	(sisssss)	LD=13
to Polish "fyzika"	(sixsxss)	LD=9
to Afrikaans "fisika"	(sisxxss)	LD=9
to Italian "fisica"	(sisxxxs)	LD=7
to French "physique"	(xxxxxssi)	LD=5

The results for intersymbolic similarity are:

$IS_1 = 0.6*(0.4*0.8 + 0.4*0.65 + 0.2*0.8) + 0.4*0.8 = 0.764$
$IS_2 = 0.6*(0.4*0.8 + 0.4*0.65 + 0.2*0.9) + 0.4*0.8 = 0.776$
$IS_3 = 0.6*(0.4*0.8 + 0.4*0.72 + 0.2*0.8) + 0.4*0.8 = 0.781$
$IS_4 = 0.6*(0.4*0.8 + 0.4*0.80 + 0.2*0.8) + 0.4*0.8 = 0.800$
$IS_5 = 0.6*(0.4*1.0 + 0.4*0.90 + 0.2*0.9) + 0.4*0.8 = 0.884$

In comparison with LD, which produced totally different distances, ranging from 5 to 13, we can see that the intersymbolic similarity technique produced similar scores for the five L2 words, arguably because the technique can capture the fact that all the L2 words are more or less recognizable based on the knowledge of the original word.

Conversely, we can have an opposite situation in which two words produce smaller Levenshtein Distance, but score worse on intersymbolic similarity, such as the case of the English word "glamour" and the French "amour", whose LD=2scores better than the synonyms in the example above, but whose IS=0.52 indicates less actual similarity.

Table 1: Similarity levels for different words and languages

Language	Word 1	IS	Word 2	IS	Word 3	IS
Esperanto	floro	-	ĉokolado	-	cirko	-
English	flower	0.91	chocolate	0,79	circus	0,84
French	fleur	0.90	chocolat	0,81	cirque	0,88
Spanish	flor	1.00	chocolate	0,81	circo	0,94
Portuguese	flor	1.00	chocolate	0,81	circo	0,94
Italian	fiore	0.90	cioccolata	0,81	circo	0,94
Romanian	floare	0.91	ciocolată	0,81	cirk	0,86
German	Blume	0.11	Schokolade	0,88	Zirkus	0,88
Dutch	bloem	0,29	chocolade	0,88	circus	0,84
Afrikaans	blom	0,31	sjokolade	0,88	sirkus	0,84
Polish	kwiat	0,12	czekolada	0,83	cyrk	0,89
Indonesian	bunga	0,00	cokelat	0,48	sirkus	0,84
Russian	Цветок (tsvetok)	0,10	Шоколад (shokolad)	0,88	Цирк (cyrk)	0,89
Hindi	फल (fool)	0,65	चॉकलेट (chākleṭ)	0,57	सर्कस (sarkas)	0,65
Arabic	زَهْرَة (zahra)	0,00	شوكولاتة (shūkulāta)	0,60	سيرك (zirk)	0,63
Japanese	花 (hana)	0,00	チョコレート (chokorēto)	0,79	サーカス(sākasu)	0,31
Chinese	花 (huā)	0,00	巧克力 (qiǎo kē lì)	0,42	馬戲 (mǎ xì)	0,00

In order to further test the proposed technique, we selected three words from the basic lexicon of our L1 (Esperanto) and calculated their respective levels of similarity to corresponding words in 16 other (L2) languages, from different families, as shown in Table 1. For languages that do not use Latin script, we used a phonetic transcription of the words in question. The results are presented in the form of total similarity scores. The difference of writing systems, as illustrated in the lower rows of Table 1, can be an additional difficulty in the learning process. In a Web-based context, however, one can assume that many of the learners from those cultural regions will likely be already used with the Latin script. For other contexts one could, for instance, represent the different scripts as a reduction factor in the calculation of word similarity (equation 1).

6. Discussion of the results and conclusions

We believe that the technique provides similarity values that capture the crucial features that make a word more easily recognizable by learners whenever their L1s contain a lexical unit that favors such iconic inference. In terms of effective word recognition, we conjecture that the higher the level of similarity between L1 and L2 words, the higher the probability of correct recognition (and easier memorization). Furthermore, we can assume that there is a threshold below which the recognition will no longer be possible (at least based on intersymbolic iconicity). The identification of the specific thresholds for speakers of each L1 is something that could be done in tests involving a significant number of individuals of each linguistic group. This was not in the scope of this paper. However, a field study with a reasonable number of individuals is being designed so that we can investigate how this threshold relates to the linguistic knowledge of each subject, such as the lexicon of L1 or other known languages (what is especially relevant in the cases of native speakers of languages with little lexical similarity with the target L2, if those speakers have some basic skills of another L2 more closely related to the target language).

Still related to the iconic link to L1 vocabulary, a pertinent question is how the word recognition process could be affected by other similar derivative words, such as, for example, the case of the word “episkopo”, that has weak similarity with its English translation, “bishop”, but a very high similarity with the corresponding adjective in English, “episcopal”. A full-fledged implementation should be capable of considering such indirect similarities in the calculations, for instance, by measuring the distance not only to the direct counterpart, but the average distance to all correlated word, and maybe assigning different weights to similarities with less used words (such as in the case of “episcopal”, that is less frequent than “bishop”).

The purpose of this technique is to offer a practical word-level similarity metric to compare symbols from different languages so that this measure can be used as an input to initialize the learner model or to evaluate word-level errors in the context of CALL applications. It is not aimed to replace formalisms such as HPSG [10], neither to create new computational treatments of lexical rules, such as those discussed in [11, 12, 13].

In what refers to the performance of the described technique, we need to point out that calculation speed was not a primary concern, since we are more interested in the accuracy in capturing intersymbolic similarity. Furthermore, in the particular context of our ITS CALL, such lexical (dis)similarities can be used to initialize the learner models *a priori*, and then the processing load of the technique can be less relevant because it is used offline. And even in the case of the error module, responsible for comparing learner answers with the right answers, much of the calculation can be done offline, if one uses the technique to create a list of common cross-linguistic errors for every

learner L1 profile, leaving to the online processing the more simple task of finding the applicable error case from among a limited list of preprocessed error types.

As discussed in [1], once the similarity (and then the distance) values are known, it is possible to apply some kind of classification or clustering algorithm, such as self-organized maps, to classify new strings. In our application we are developing a SOM, which will be used to classify word-level errors in terms of their similarities with common error types, including interferences caused by influence from L1, in which case we expect to see such errors clustered around the position that represents the corresponding L1-word.

References

[1] Fischer, I., Zell, A.: Processing Symbolic Data with Self-Organizing Maps. *Workshop SOAVE*, (2000).

[2] Levenshtein, L.I.: Binary codes capable of correcting deletions, insertions and reversals. Soviet Physics-Doklady, 10 (8) (1966).

[3] Kohonen, T.: Self-Organization and Associative Memory, *Springer Series in Information Sciences*, vol. 8, Springer Berlin Heidelberg (1988).

[4] Wagner, R.A., Fischer, M.J.: The string-to-string correction problem, *Journal of the ACM* 21 168-173 (1974).

[5] Petroni, F., Serva, M.: Language distance and tree reconstruction. Journal of Statistical Mechanics: Theory and Experiment. IOP Publishing Ltd and SISSA (2008).

[6] Brysbaert, M., Nazir, T.: Visual constraints in written word recognition: evidence from the optimal viewing-position effect, *Journal of Research in Reading*, v.28, i.3, pp. 216-228 (2005).

[7] Drieghe, D., Brysbaert, M. Strategic effects in associative priming with words, homophones, and pseudohomophones, *Journal of Experimental Psychology: Learning, Memory and Cognition*, 28, 951-961.

[8] Harm, M.W., Seidenberg, M.S., Computing the meaning of words in reading: Cooperative division of labor between visual and phonological processes. *Psychological Review*, 111, 662-720 (2004)

[9] Grainger, J., Whitney, C.: Does the huamn mnid raed wrods as a whole? *Trends in Cognitive Science*, 8, 58-59.

[10] Pollard, C., Sag I. A.: *Information-based Syntax and Semantics*. Volume 1: Fundamentals. Stanford: CSLI Publications (1987).

[11] Meurers, W.D.: Towards a semantics for lexical rules as used in HPSG. University of Tübingen (1995).

[12] Meurers, W.D., Minnen, G.: A computational treatment of lexical rules in HPSG as covariation in lexical entries. Association for Computational Linguists (1997).

[13] Dickinson, M., Meurers, W.D.: Detecting errors in part-of-speech annotation. In Proceedings of EACL, pages 107–114, Budapest, Hungary (2003).

Cohesion, Semantics and Learning in Reflective Dialog

Arthur WARD, John CONNELLY, Sandra KATZ, Diane LITMAN, Christine WILSON
Learning Research and Development Center, University of Pittsburgh

Abstract. A corpus of reflective tutorial dialogs was tagged for cohesive relationships between student and tutor. We describe our tagging scheme, and show that certain cohesive features of tutoring dialog are correlated with learning in our corpus. In particular, our *semantic* cohesive relationship tags are significant predictors of learning, while our *lexical* tag is not. We find that "abstractive" dialog moves, in which the student or tutor repeats the other's previous utterance but at a greater level of generality, are significant positive predictors of learning. We also find that tutor moves which repeat the student's previous utterance but in a less abstract way predict learning in our corpus. These findings suggest that tracking student dialogue moves can enhance student modeling and guide planning of effective natural-language dialogues.

Keywords. Intelligent Tutoring, Learner Modeling, Discourse Analysis

1. Introduction

Interactive tutorial dialog with a human tutor has been shown to be a very effective form of instruction [1, 2]. Many researchers have hypothesized that the very *interactivity* of tutorial dialog contributes to the effectiveness of one-on-one tutoring, and there is substantial empirical support for this hypothesis [3–5]. Although we have some idea what interactivity looks like from the perspective of exchange level analysis [5, 6], we know little about what specific discourse mechanisms contribute to interactivity, or how they affect learning. Identifying discourse mechanisms that correlate with learning might help us both to improve our tutoring system dialogs, and also to improve our student models by helping us recognize knowledge gaps and learning during tutoring.

Based on previous work [7, 8], we suspect that "cohesion" is an important discourse mechanism in tutoring. Following others [9], we consider cohesion to be the connectedness of a text. Cohesive devices such as pronoun reference and word repetition tell us what elements to include in our mental model and how to connect them. Zwaan and Radvansky [10] consider text to be a "set of processing instructions on how to construct a mental representation of the described situation" (p. 162). The result of following these instructions *may* be a coherent mental model. However, work in textual cohesion has shown that not all readers respond to these processing instructions in the same way. In a series of experiments (e.g.: [11, 12]), McNamara and her colleagues have shown that low knowledge readers gain in both comprehension and recall from reading a high, but not

a low cohesion text. On the other hand, high knowledge readers, particularly those with low comprehension skill, show better comprehension gains when given a low cohesion text.

Cohesion in *dialog* can be considered a record of the participants' "collaboration toward coherence" [13]. Dialog participants use various cohesive devices to establish common ground [14], negotiate references [15], and coordinate their mental models [16]. Just as high cohesion text can indicate more detailed instructions for building a mental model (relative to low cohesion text), high cohesion dialogue may signal more detailed collaboration between dialog participants, in building a shared mental model.

Our previous work provided some evidence that cohesion in tutorial dialog interacts with student prior knowledge in a way similar to that of cohesion in expository text. We have found that simple automatically detectable cohesive devices such as word and word-stem repetition between tutor and student predicted learning for low knowledge students, but not for high knowledge students [7]. We later found that also counting cohesive ties between words that were lexically different but semantically related in a hypernym/hyponym hierarchy improved the correlation with far-transfer learning for high pre-testers [8]. In that study far-transfer learning was evaluated using questions that were non-isomorphic to the tutored problems. We discuss a related definition of far transfer for the current corpus in Section 3.

The previous, automatically detectable cohesive ties fit under Halliday and Hasan's [9] category of "lexical cohesion," which includes word, synonym and superordinate-class reiteration. Our implementations, however, were limited to recognizing simple lexical relations between single words. In the current work we use a similar tag which counts simple lexical repetition (our "Exact" tag, Section 2). However we also count more sophisticated *semantic* relations, and recognize ties between multiple-word spans. We find that in our corpus, these manually tagged measures are in fact better predictors of learning than the simple lexical measure. Specifically, we find that tags indicating tutor or student abstraction are significant predictors of learning in our corpus. A tag indicating tutor specialization is also a significant predictor of learning. Similarly to our previous work [7, 8], we find that student response to cohesion varies with both student preparedness and with the type of learning being measured. Our results suggest that abstraction and specialization are important cohesive devices in tutorial dialog. We argue in Section 5 that this has implications for both student modeling and dialog planning.

We describe our tagging scheme in Section 2, our corpus in Section 3 and correlations between tags and learning in Section 4.

2. Tagging for Cohesion

Our previous, automatically computable tags attempted to identify when the tutor and student were referring to each other's contributions. When selecting our expanded tag set, we again focused on ties between tutor and student contributions which might indicate their types of interactivity.

Our final tag set is largely a subset of Halliday and Hasan's [9] taxonomy of cohesive devices. Tags and their brief definitions are listed below. The bracketed numbers after the tag name indicate the tag's total count in tutor (**T:**) and in student (**S:**) turns.

- **Exact** [T:899 S:512] is used when one utterance and the next contain the same word, either in identical or inflected forms.
- **Synonym** [T:67 S:36] is applied when two words with similar meanings are used.
- **Paraphrase** [T:605 S:205] is used for phrase repetitions with word substitution or with different word order.
- **Pronoun** [T:327 S:153] repetition is used when a pronoun such as "it" in one utterance refers to a discourse entity in the previous utterance.
- **Superordinate-class** [T:236 S:50] is used when one speaker uses a more general or abstract referring expression. Examples from our corpus include "force" as a more general reference to "weight," and "velocity" when it follows the more specific "horizontal components of velocity."
- **Class-member** [T:206 S:214] is used when a *more* specific word or phrase such as "horizontal" is used after a less specific one such as "direction."
- **Collocation** [T:121 S:55] is the use of lexical items that regularly co-occur. We follow Halliday and Hasan (who refer to collocation as "the most problematical part of lexical cohesion") and emphasize collocations that stand in some relation of complementarity, such as "left-right" and "up-down." Although collocations are often between individual words, we also recognize the relationship between phrases when they have the complementarity relation.
- **Negation** [T:46 S:25] is used when one speaker directly contradicts the previous speaker.

In choosing this tag set, we selected cohesive devices from Halliday and Hasan (H&H) [9] which could identify common reference between tutor and student, and which seemed to be present in our corpus. We combined some devices which had been distinct in H&H but which were poorly represented in our corpus. For example, our "pronoun" category includes devices such as "nominal reference" ("this") and other types of substitution (e.g. "one"). Our categories of "exact," "synonym," "superordinate-class" and "class-member" correspond to types of lexical reiteration in H&H. Our "paraphrase" tag, however, has no corresponding device in H&H. It is designed to recognize when tutor and student use entire phrases to refer to the other's contribution, and can often contain other types of ties, such as ellipsis, synonym and collocation.

Table 1 contains examples of most of these tags taken from our corpus, edited slightly for clarity. A tutor utterance and the student utterance that followed it are shown at the top of the table. Below them are shown the spans identified in each utterance and the tags given to those spans. In the middle of the table the student utterance and the tutor utterance that followed it are shown. Again, the spans identified in each utterance are shown below them, with their tags. For example, there are two cohesive ties shown[1] between the first two utterances: "superordinate-class" and "exact."

As can be seen from the above definitions and examples, many of our tags required the identification of spans of words that were being paraphrased, elided or otherwise referred to. Identifying spans turned out to be difficult. Spans can be split (as when the referents of "those" are in separated clauses of the preceding utterance), and can even overlap. An example of overlapping spans is in Table 1, where "coming down" is tagged as a collocation, and is also part of the paraphrase "faster coming down." An important and difficult part of applying this set of tags is therefore identifying appropriate spans.

[1] Other ties were removed from the example for clarity.

Using this tag set, two coders tagged a training corpus of 518 student and tutor turns, iteratively refining tag definitions and re-tagging. During this initial tagging phase, the coders relied largely on lexical features. That is, a cohesive tie would be tagged if the words in one span, taken by themselves, could be construed to have a cohesive relationship to the words in the other span. No attempt was made at this stage to judge if the spans referred to the same discourse item or if the relationship made sense in context.

Following this initial coding, we performed a second coding pass in which we re-evaluated spans which had been previously tagged "superordinate-class" "class-member" or "collocation." The remaining tags will be checked later, as time allows. In the new pass we required that the ties previously selected using only lexical features also make sense, and we eliminated the ones that didn't. Ties were eliminated when their spans seemed to have mis-matched topics or referents. Ties were also eliminated if they were not original to the second speaker. For example, if the first speaker had used both "weight" and "force," and the second speaker had also used "force," we would no longer count a superordinate-class tie between "weight" and "force" in the second utterance. We did this in order to distinguish between lexical repetition and knowledge co-construction or elaboration on the part of the second speaker.

Tutor: Good. And the effect on the water is the same. What about the horizontal components of the velocity (of the ball or of the water - either?)		
Student: Velocity is in the same direction as acceleration so the ball is faster coming down		
Tut. Span	**Stu. Span**	**Tag**
horizontal components of the velocity	velocity	superordinate-class
ball	ball	exact
Student: Velocity is in the same direction as acceleration so the ball is faster coming down		
Tutor: It slows down going up and it speeds up coming down - but all the time the horizontal components of the velocity stay unchanged. Horizontal components of velocity are unaffected by gravity. Ok?		
Stu. Span	**Tut. Span**	**Tag**
the ball	it	pronoun
faster coming down	speeds up coming down	paraphrase
velocity	horizontal components of the velocity	class-member
same	unchanged	synonym
coming down	going up	collocation
direction	horizontal	class-member

Table 1.Example Cohesion Tags

One coder re-tagged all instances of these three tags, and a second tagger coded a randomly selected 10% of them for agreement analysis. Kappa on these tags was .57. Agreement was counted when both taggers identified the same textual span *and* applied the same tag to it. Due to the difficulty of reaching agreement on spans, they were counted as the same if they had substantial overlap (no more than one word different at either end, not counting stop-words).

3. Corpus

Our corpus was collected as part of a study of the effectiveness of post-practice, reflective discussions [17]. This study had three conditions. In each condition, students solved a

series of physics problems in the Andes physics tutoring system [18]. After the Andes session, the students were asked "reflection questions" that invited them to elaborate on a specific part of the solution. For example, the following reflection question changes one variable in a previous problem about a jumper hanging motionless from a bungee cord:

> Suppose the maximum tension that the bungee cord could maintain without snapping was 700 N. What would happen to the bungee jumper if he hung stationary on the cord?

After answering reflection questions, students in two conditions were given either canned text feedback or no feedback. In the third condition, however, students used a chat interface to engage a human tutor in dialogue about the reflective questions. Our corpus is taken from these dialogs. The experimental procedure used and the resulting dialogue corpus can be summarized as follows.

Sixteen students answered a questionnaire, then were given a math test and a physics pre-test. After the pre-test, students reviewed a workbook chapter on kinematics developed for the experiment, and received training in the use of the Andes tutoring system. They then solved 12 physics problems in Andes. Following each problem, they were given between three and eight reflection questions. They would type their answer to each question, or state that they could not answer the question, and then engage a human tutor in reflective dialog about the answer, using a chat interface. Fifteen students participated in 60 reflective dialogs each, while the sixteenth participated in 53 dialogs. This created a corpus of 953 reflective dialogs, containing 2,218 student turns and 2,136 tutor turns. A post-test was given after the reflective dialogs. The pre- and post- tests covered the same topics and contained 36 questions: 9 quantitative mechanics questions similar to those that the students worked on in Andes, and 27 qualitative questions that tested their ability to apply mechanics concepts and principles to new problems that were dissimilar to those tutored under Andes. The pre- and post- tests were administered in a counterbalanced order. Overall, the researchers found that students in both dialog treatment conditions learned more than students in the no-dialogue control condition, as measured by pre-test to post-test gain score, but the canned feedback and human feedback conditions did not differ significantly.

In Section 4 we show that some of our tags predict learning measured by the qualitative but not the quantitative questions. Because the qualitative questions were less similar to the training problems, we argue that they measure farther transfer of learning than do the quantitative questions. While this transfer is probably not all that "far" in the taxonomy described by Barnett and Ceci [19], it seems probable that success on these problems required the construction of a more abstract representation of the material than was needed for the quantitative problems.

As noted in Section 1, students of different knowledge levels may respond to cohesive cues differently, so we ran statistics on our "high" and "low" pre-testers separately, as divided by their median pre-test score. When using the quantitative questions, this division results in eight low and eight high pre-testers. Using either the qualitative questions or the set of all questions results in seven low and nine high pre-testers.

4. Results

We used linear models to look for relationships between each of our cohesion tags and learning, as measured by pre- and post-test scores (total scores as well as quantitative and

	All Questions											
	All Students				**Low Pre-test**				**High Pre-test**			
Tag Name:	**Pre**	**Mth**	**Tag**	**Mod**	**Pre**	**Mth**	**Tag**	**Mod**	**Pre**	**Mth**	**Tag**	**Mod**
S:Super-Ord	.061	.070	*.054*	**.005**	.259	.562	.152	.109	.466	.146	.420	.272
	Qualitative Questions											
	All Students				**Low Pre-test**				**High Pre-test**			
Tag Name:	**Pre**	**Mth**	**Tag**	**Mod**	**Pre**	**Mth**	**Tag**	**Mod**	**Pre**	**Mth**	**Tag**	**Mod**
S: Super-Ord	.274	.006	**.005**	**.002**	.987	.072	*.066*	.111	.966	.159	.229	.295
T: Class Mem	.205	.006	**.015**	**.005**	.983	.620	.794	.629	.537	.097	**.032**	*.059*
T: Super-Ord	.296	.049	.262	**.045**	.029	.011	**.017**	**.032**	.488	.185	.748	.550
	Quantitative Questions											
	All Students				**Low Pre-test**				**High Pre-test**			
Tag Name:	**Pre**	**Mth**	**Tag**	**Mod**	**Pre**	**Mth**	**Tag**	**Mod**	**Pre**	**Mth**	**Tag**	**Mod**
T: Class Mem	.035	.671	.873	*.093*	.058	.527	**.050**	.153	.283	.533	.614	.155

Table 2. P-values for individual predictor variables (Pre-test, Math and tag count) and whole linear model

qualitative subscores). We regressed post-test score (or relevant sub-score) on pre-test score (or relevant sub-score), math test score, and normalized tag count. We included pre-test scores as predictors because they are significantly correlated with post-test scores in our corpus[2], and math test scores because they were shown to be a significant predictor of learning in previous work with the Andes tutor [17]. We use normalized tag counts to control for the effect of longer tutorial dialog on learning. For example, the tag "Student Superordinate-class" (S:Super-Ord) is the total count of "superordinate-class" tags for a student, divided by that student's total number of turns.

Results are shown in Table 2. The table is divided horizontally by which measure of learning gain is used in the regression model. The models at the top use all 36 questions, the models in the middle use the 27 qualitative questions, and the models at the bottom use the 9 quantitative questions. For each measure of learning, the tags that were significant predictors (or strong trends) in at least one model are shown in the left most column. P-values for the linear models that use that tag are shown on the same row. For each linear model, the table shows the individual p-value for each predictor variable: "Pre" = relevant pre-test score, "Mth" = math score, "Tag" = the cohesion tag used in the model, "Mod" = the p-value for the whole model. Significant tag and model p-values are shown in bold, and trends are italicized. For each tag, we ran regressions using the entire student sample as well as on the subgroups of high and low pre-testers.

The only tag that approached significance when predicting total learning gains (under "All Questions") was Student Superordinate-class. This tag had a p-value of .054 for All Students, but was not even a trend for the low or high pre-testers taken separately. Similarly only one tag, "Tutor Class-member," was significant in predicting quantitative learning gain, but it occurred in a non-significant model (p = .153).

For qualitative learning, however, three different tags proved to be significant predictors. As shown in the center rows of Table 2, the Student Superordinate-class tag was

[2]The pre- to post-test correlation for all questions is .67 (p =.0045), for the quantitative (near) questions: .63 (p = .009), and for the qualitative (far) questions: .51 (p = .043)

significant for all students and achieved a trend in the sub-group of low pre-testers, although in a model which was not significant overall. Similarly, the Tutor superordinate-class tag was a significant predictor for low pre-testers. Tutor class-member was a significant predictor for all students. It was also significant for high pre-testers, but in a model that fell slightly short of being significant ($p = .059$).

5. Discussion and Future Work

The goal of this study was to expand upon our previous work which had suggested the importance of cohesion in tutorial dialog. Those studies had found that automatic measures of interactivity, which measured when tutor and student used similar words or words that were related in WordNet's is-a hierarchy, are correlated with student learning. The shallow measures we used then, however, could only provide limited insight about exactly what mechanisms account for the value of interactivity. Our new tags capture semantic relationships between phrases which were invisible when counting only shallow lexical relationships between individual words.

The current study has broadened and reinforced our earlier work by showing that different measures of tutor-to-student cohesion also positively predict learning in a new corpus of tutorial dialogs. In addition, it provides insight into exactly what mechanisms are involved. By tagging manually rather than automatically, we were able to recognize a broad set of semantic relationships between tutor and student utterances. Some of these cohesive relationships did not seem to be related to learning in our corpus. The ones that *were* related involve changes in the level of concreteness being used. In particular, tutor or student abstraction seems to be a particularly valuable cohesive device. We suggest that this type of cohesive tie tends to happen when the student is building a more abstracted mental model. This model is then more useful in answering far-transfer questions, as shown by our results in predicting "qualitative" learning.

These results can also be seen as adding detail to previous work by Katz et al. [17], who found that the number of reflective dialogs that abstracted from the previous problems were correlated with learning.

It is also interesting to note that the "Exact" tag was not significant in any model, even though this tag is similar to the lexical reiteration measure which correlated with learning in other corpora [7]. In the current corpus of reflective dialogs, only tags that were sensitive to the semantic content of the utterances were significant predictors of learning.

Our results using the Tutor Class-member and Tutor Superordinate-class tags suggest that we might be able to improve learning by manipulating tutor and student utterances. In future work we hope to test this experimentally, by making tutor utterances more concrete or more abstract at appropriate places in the tutorial dialog, and by prompting students to do the same. Further work will be required to tell where those places are.

Our results using the Student Superordinate-class tag suggest that we might be able to improve student modeling in our tutor by measuring student abstraction during tutoring. We hope to explore this possibility by using cohesion within certain dialog segments to predict correctness on particular post-test questions. If we can in fact build better student models through more sophisticated automated cohesion analysis, this model could then be used to guide more effective tutorial dialog planning.

6. Acknowledgments

Funding for this work was provided by the Office of Naval Research (ONR), Grant Number N000140710039, and by the Learning Research and Development Center. These organizations do not necessarily endorse the data or views presented here.

References

[1] B. S. Bloom. The 2 sigma problem: The search for methods of group instruction as effective as one-to-one tutoring. *Educational Researcher*, 13:4 – 16, 1984.

[2] A. Corbett. Cognitive computer tutors: Solving the two sigma problem. *Proceedings of the Eighth International Iser Modelling Conference*, pages 137 – 147, 2001.

[3] Michelene T.H. Chi, Stephanie A. Siler, Heisawn Jeong, Takashi Yamauchi, and Robert G. Hausmann. Learning from human tutoring. *Cognitive Science*, 25:471 – 533, 2001.

[4] Michelene Chi, Marguerite Roy, and Robert Hausmann. Observing tutorial dialogues collaboratively: Insights about human tutoring effectiveness from vicarious learning. *Cognitive Science: A Multidisciplinary Journal*, 32:301 – 341, 2008.

[5] Arthur C. Graesser, Natalie Person, and Joseph P. Magliano. Collaborative dialogue patterns in naturalistic one-to-one tutoring. *Applied Cognitive Psychology*, 9:495 – 522, 1995.

[6] Kate Forbes-Riley, Diane Litman, Alison Huettner, and Arthur Ward. Dialogue-learning correlations in spoken dialogue tutoring. In *Proceedings 12th International Conference on Artificial Intelligence Education (AIED),Amsterdam, Netherlands*, July 2005.

[7] Arthur Ward and Diane Litman. Cohesion and learning in a tutorial spoken dialog system. In *Proceedings of the 19th International FLAIRS Conference (FLAIRS-19)*, pages 533–538, May 2006.

[8] Arthur Ward and Diane Litman. Semantic cohesion and learning. In *Proceedings 9th International Conference on Intelligent Tutoring Systems (ITS)*, pages 459–469, Ann Arbor, June 2008.

[9] M. A. K. Halliday and Ruqaiya Hasan. *Cohesion in English*. English Language Series. Pearson Education Limited, 1976.

[10] Rolf A. Zwaan and Gabriel A. Radvansky. Situation models in language comprehension and memory. *Psychological Bulletin*, 123:162 – 185, 1998.

[11] Danielle S. McNamara and Walter Kintsch. Learning from text: Effects of prior knowledge and text coherence. *Discourse Processes*, 22:247–287, 1996.

[12] Danielle McNamara. Reading both high-coherence and low-coherence texts: Effects of text sequence and prior knowledge. *Canadian Journal of Experimental Psychology*, 55:51–62, 2001.

[13] *Coherence in Spontaneous Text*. Typological Studies in Language, 31. John Benjamins, Philadelphia, 1995.

[14] Hebert H. Clark and S. A. Brennan. Grounding in communication. In L. B. Resnick, J. M. Levine, and S. D. Teasley, editors, *Perspectives on socially shared cognition*. 1991.

[15] Peter A. Heeman and Graeme Hirst. Collaborating on referring expressions. *Computational Linguistics*, 21:351–382, 1995.

[16] Martin J Pickering and Simon Garrod. Toward a mechanistic psychology of dialogue. In *Behavioral and Brain Sciences*, volume 27, 2004.

[17] Sandra Katz, David Allbritton, and John Connelly. Going beyond the problem given: How human tutors use post-solution discussions to support transfer. *International Journal of Artificial Intelligence in Education*, 13:79 – 116, 2003.

[18] K. VanLehn, C. Lynch, K. Schulze, J.A. Shapiro, R. Shelby, L. Taylor, D. Treacy, A. Weinstein, and M. Wintersgill. The andes physics tutoring system: Lessons learned. *International Journal of Artificial Intelligence and Education*, 15:147 – 204, 2005.

[19] Susan Barnett and Stephen Ceci. When and where do we apply what we learn?: A taxonomy for far transfer. *Psychological Bulletin*, 128:612 – 637, 2002.

Speling Mistacks and Typeos: Can your ITS handle them?

Adam M. RENNER[a], Philip M. McCARTHY[a], Chutima BOONTHUM[b], and Danielle S. McNAMARA[a]
[a]*University of Memphis*
[b]*Hampton University*

Abstract. Guided feedback is a critical component of learning environments that emphasize constructive and active learning processes. In an ITS environment, accurate evaluation of the user's input is essential when feedback is provided by the system. The main purpose of this study was to evaluate the robustness of an established ITS in its ability to evaluate user input that may be ill-formed (e.g., containing spelling and syntax errors). The corpus, called User-Language Paraphrase Corpus (ULPC), consisted of 1998 student responses collected during the paraphrase module in iSTART, an ITS that provides metacognitive reading strategies and self-explanation training. iSTART relies on a computational feedback system in order to interact with and scaffold instruction for the learner. The unedited student responses were evaluated by iSTART, along with a parallel version of the corpus in which each response was corrected for typographical and grammatical errors. The results indicated that error correction significantly affected the feedback produced by iSTART. Our results suggest that preprocessing of typed user input would allow ITSs to more strictly scrutinize such input and provide more accurate feedback.

Keywords. Intelligent tutoring systems, ITS, paraphrase evaluation, feedback, natural language processing, NLP

Introduction

Current research in the educational sciences indicates that guided feedback and scaffolded training is an effective means of instruction. Computer programs afford an effective and efficient means of implementing this type of teaching method [1]. Intelligent Tutoring Systems (ITSs) provide one approach to doing so by engaging users in one-on-one, adaptive tutoring. One way adaptive tutoring is achieved by some ITSs is through conversational dialogue, which relies on computational linguistic algorithms to translate and respond to natural language input from the user [2, 3]. Because ITSs depend on computerized interpretation of the unedited input of the user to determine the feedback response, the effectiveness of dialogue-based ITSs partially depends on the accuracy of its underlying Natural Language Processing (NLP) system. The accuracy of the evaluation of a user's input is important because it affords appropriate individualized instruction that simulates interactions with a human tutor. Successful and accurate interactions between the user and the ITS are those where the ITS is able to ascertain what the user *intended*, which increases the likelihood of cognitive and learning gains and results in motivational advantages [4].

The last two decades of research in computational linguistics have led to major advances in the NLP technologies that provide the backbone for ITSs. For instance, text-relatedness indices such as Latent Semantic Analysis (LSA) cosines [5] and shallow overlap measurements [3] have contributed to effective assessment designs within many of the systems that evaluate natural language input (e.g., iSTART) [6]. In addition, entailment methods have also recently reported considerable success [7].

Although such research is certainly of great importance, the dominant focus of many of these indices has been to evaluate edited, polished texts; an endeavor that has had considerable success [cf. 8, 9]. In contrast, research is relatively sparse with respect to the computational assessment of textual relatedness in unedited ITS user input, or *user-language*, which is riddled with typos and grammatical errors (e.g., [3]).User-language is the typed natural input of a user interacting with an ITS. In fact, it can be a major challenge for natural dialogue ITSs to accurately evaluate user-language [10]. A few studies have reported on techniques for correcting errors in user-language, but such approaches may be limited to testing on artificial datasets or with students who are comparatively proficient [e.g., 11, 12]. Thus, the present study addresses the challenge of evaluating natural language input within the context of an authentic, error-saturated ITS environment, and the problems that may be subsequently encountered in providing feedback to the learner. Although issues with user errors are readily addressed in word processors and online applications (e.g., spell check in email services), ITSs necessitate that corrections be made *silently*, so as not to distract from the intended learning task.

It may be impractical to assume that students using ITSs will type flawlessly. In practice, user-language has a high rate of misspellings, inadvertent keying mistakes, and dubious syntactic choices [18]. Conventional text relatedness indices (e.g., LSA) may have limited accommodation for such issues, and consequently responses that contain misspelled words or other typographical errors may lead to lower similarity values. A lower value would produce unhelpful feedback that reflects the misspelling rather than the student's grasp of the key concept [13]. These consequences are of concern because many ITS interventions have been shown to be of greatest benefit for low domain knowledge and low ability learners, who make more of these types of errors [14]. Additionally, systems that engage with nonnative speakers have dealt with similar issues [15]. While some ITSs have implemented automatic spell-checking techniques (e.g., Why2-Atlas [16], CIRCSIM-Tutor [12]), many more ignore errors altogether. Furthermore, some existing spell-checking methods utilize a lexicon, which may be computationally expensive and ill-suited for real-time feedback. Moreover, it requires that the language produced by the user be highly predictable and constrained.

The present research focuses on characterizing and evaluating the User-Language Paraphrase Corpus [17], according to the types of errors commonly found in user-language. The corpus comprises 1998 target/response pairs, collected from interactions with the paraphrasing module of iSTART, an ITS that provides students with self-explanation and reading strategy instruction [7].

Many ITSs may not address user errors because they use indices that appraise the overlap over the whole text, which is thought to make their evaluations resistant to individual errors. Conversely, our previous research [18] has revealed that measures such as LSA, Overlap Indices, and Entailment changed significantly after correcting user errors. This finding is not surprising, given that these indices are trained on large bodies of texts where the errors are relatively few, but the overwhelming majority of responses in the present ITS data (1968 out of 1998) comprise a single sentence. Our goal in the current study is to appraise the feedback algorithm used by iSTART, which

applies some of the same computational components. The ULPC also provides human ratings of the responses on paraphrase quality, which are gold standard ratings. Because one goal of iSTART is to offer feedback that is comparable to human tutors, this study will also examine how human ratings of paraphrase quality differ when errors are corrected (for further detail on the ULPC, see [17]).

1. iSTART

iSTART (Interactive Strategy Training for Active Reading and Thinking) is a web-based ITS that uses animated pedagogical agents to teach self-explanation and reading strategies to high school (grades 9-12) and college students [6]. The reading strategies include *comprehension monitoring*, *paraphrasing*, *elaboration*, and *making bridging inferences*. The ULPC was collected during student interactions with the iSTART module that focuses exclusively on generating paraphrases. For example, the following sentence, called Target Sentence (TS), is from a science textbook and the student input, called paraphrase (P), is reproduced from the ULPC corpus. The samples in this study are replicated as typed by the student.

TS: *Over two thirds of heat generated by a resting human is created by organs of the thoracic and abdominal cavities and the brain.*

P: *a lot of heat made bya lazy person is made by systems of your stomack and thinking box.*

The computational aspect of the system is based on a match between the student's paraphrase and the target sentence, which determines the feedback response. The same algorithm is used for paraphrases as for all other strategies in the iSTART system. That is, the algorithm is designed to fit the entire range of reading strategies covered in the iSTART curriculum, not only paraphrases. First, *frozen statements* (e.g. *I think this sentence is saying…*) are responded to if they comprise more of the self-explanation than the explanatory material, or they are filtered from the remaining content of the response. If the response has little in common with the target sentence (i.e., IRR), the student is asked to add relevant information to the response. This procedure provides an important filter for the evaluation of paraphrases because the paraphrase must be suitably different from the target sentence yet still be relevant to it. Appropriateness of length is also assessed, such that if the self-explanation is too short (SH), more information is requested from the student. If it is not too short, similarity to the target sentence is assessed. If the response is highly similar (i.e., SIM1), the student is asked to revise the response. If the response is somewhat similar, then the system asks for a revision in the context of the self-explanation practice module. Lastly, the response manager in the paraphrasing module treats responses that are adequate paraphrases (i.e., SIM2) or better responses (i.e., OK) as good responses, resulting in positive feedback.

The iSTART feedback system integrates word matching and LSA in its feedback algorithm. Words are matched against words in a benchmark in two ways: (1) literal matching – compare character by character, and (2) soundex matching [19] – in which vowels are eliminated and similar consonants are mapped to the same soundex symbol (e.g., *b*, *f*, *p*, *v*). At the end of this matching process, the total matching word count is computed (literal match count plus soundex match count) for each benchmark. Words that do not match any of these benchmarks are counted as new words. LSA provides a measure of semantic overlap between the student's response and the benchmark by computing cosines between their vector representations in a high dimensional semantic

space. The LSA cosine value ranges from 0 to 1 and represents the degree of conceptual overlap between the linguistic elements. Detailed descriptions and evaluations of the iSTART algorithm can be found in [3].

2. Procedure and Predictions

Two trained expert evaluators identified and categorized the errors in a subset of the ULPC corpus, and one of the raters completed this procedure for the full corpus. Inter-rater agreement was assessed for a subset of the data (10%; $n = 200$). Cohen's Kappa for overall error identification was $\kappa = .70$. With overall agreement established, we can be relatively confident that a single rater's evaluation of the data is consistent and valid. Additionally, the judgments were based on validated models of grammaticality [cf. 20]. Each error was labeled according to its type and the error was corrected. Revisions were made such that the corrected version preserved the original intent of the student's response as the evaluator discerned. The full list of errors and revisions is provided in [18]. These errors included spelling, capitalization, spacing, punctuation, and agreement (verb, article, preposition, determiner, conjunction, possessive). Among these categories, 83% contained some form of error, 52% contained some form of spelling error, and 63% of those spelling errors were *internal spelling errors* (i.e., words the student could see in the target sentence).

We speculated that the simple soundex match may not provide adequate corrections for some of the atypical words found in the iSTART curriculum. In other words, iSTART cannot compensate for *major* misspellings. Because word matching is a large part of the overall computational framework, we anticipated that a significant portion of responses generated by iSTART would change. For instance, we expected to see a response change from SIM1 to SIM2 (or vice versa) more often than from IRR (i.e., irrelevant) to SIM2. Because we expected the accuracy of the algorithm to improve overall, we also predicted that these improvements would consequently produce higher correlations with human ratings of paraphrase quality.

3. Results

A marginal homogeneity (MH) test was conducted to detect whether the two paired categorical measures (original paraphrase/corrected paraphrase) differed significantly. The MH test assesses the significance of the difference between two dependent samples when the variables are multinomial. Out of six distinct categories (frozen, IRR, SH, SIM1, SIM2, OK) and 1998 paraphrases, 244 (12.2%) responses were found to change as a result of error correction. The changes were both positive and negative. That is, a change from SIM1 to SIM2 is a positive change because it increases, and SIM2 to SIM1 is a negative change because it decreases. The cross tabulation of the original and corrected paraphrases is presented in Table 1. A Cramer's correlation coefficient was generated in order to indicate the strength of the relationship between the two categorical variables. Cramer's V for the strength of the association between original and corrected paraphrases was significant, $V = .849$, $p < .001$. However, the MH test suggested significant discrepancies in the feedback scores, MH = 5.892, $p < .001$. The results indicate that the feedback responses for the original and corrected paraphrases differ significantly because of the number and degree of the errors.

Table 1. Cross tabulation of iSTART feedback responses to user paraphrases

		iSTART response - Corrected Paraphrases						
		Meta	IRR	SH	SIM1	SIM2	OK	Total
iSTART response original paraphrases	Frozen	16	0	0	0	0	0	16
	IRR	7	120	6	0	0	6	139
	SH	1	2	206	1	0	11	221
	SIM1	0	0	0	527	7	7	541
	SIM2	0	0	0	98	194	12	304
	OK	0	0	4	37	45	691	777
	Total	24	122	216	663	245	727	1998

The results reflect a general trend that the feedback values were generally *lower* once the paraphrases were corrected for errors because 194 of the changes in the scores were negative, while only 50 were positive. This result indicates that error correction enables a more stringent evaluation of the student's attempt. For instance, of the 777 paraphrases that were evaluated as OK (beyond paraphrase), 45 of them moved one place lower in estimated quality to SIM2 (good paraphrase) and 37 moved two places lower to SIM1 (too similar). For example:

TS: *Scanty rain fall, a common characteristic of deserts everywhere, results from a variety of circumstances.*

P: *a characteristic of dearts is scanty rainfaal that causes circumstance* (OK).

P(corrected): A *characteristic of deserts is scanty rainfall that causes circumstance*. (SIM2)

In this instance, the corrections enabled the system to more accurately assess the response as only a paraphrase and nothing more, whereas originally it was assessed as better than a paraphrase (e.g., includes elaboration). Conversely, 36 previous evaluations of IRR, SH, SIM1, and SIM2 improved to OK. For example:

TS: *During vigorous exercise, the heat generated by working muscles can increase total heat production in the body markedly.*

P: *musclescan increase total heat producation in the body*. (SH)

P (corrected): *Muscles can increase total heat production in the body.* (OK)

In this example, it is apparent that the original paraphrase was just below the length threshold. By correcting the two words that ran together, the evaluation was more adequately assessed, resulting in more positive feedback. Thus, correcting for errors allows the iSTART system to make a more accurate evaluation of paraphrases.

The second purpose of this research was to assess the degree to which the iSTART algorithm was predictive of human ratings of paraphrase quality. Because our investigation is concerned with instances that had the *potential* to alter the student's response, we filtered out those cases that either required no corrections or consisted of random garbage keying. Thus, 328 cases were withheld from the analysis. Separate ANOVAs were conducted for each condition (i.e., original, corrected), including Paraphrase Quality as the dependent variable and the iSTART score as a fixed factor. The ANOVAs showed that Paraphrase Quality was significantly different as a function of the iSTART feedback category for the original paraphrases, F (5, 1636) = 53.324, p < .001, part. η^2 = .138, and for the corrected paraphrases, F (5, 1636) = 58.543, p < .001, part. η^2 = .15. Thus, for a fine-grained analysis, we conducted separate pairwise comparisons for each condition (see Table 2).

Table 2. Pairwise Comparisons of Human-Rated Paraphrase Quality.

		Original			Corrected		
		Mean Diff.	SE	Sig.[a]	Mean Diff.	SE	Sig.[a]
Frozen	IRR	.152	.402	1	.081	.361	1
	SH	-.776	.370	.581	-.922	.299	.032
	SIM1	-1.955	.363	< .001	-2.176	.288	< .001
	SIM2	-2.071	.366	< .001	-2.421	.297	< .001
	OK	-1.897	.361	< .001	-2.106	.288	< .001
IRR	SH	-.918	.209	< .001	-1.002	.245	.001
	SIM1	-2.107	.196	< .001	-2.257	.231	< .001
	SIM2	-2.223	.203	< .001	-2.502	.242	< .001
	OK	-2.0249	.192	< .001	-2.187	.231	< .001
SH	SIM1	-1.189	.115	< .001	-1.255	.112	< .001
	SIM2	-1.305	.127	< .001	-1.500	.133	< .001
	OK	-1.131	.111	< .001	-1.185	.111	< .001
SIM1	SIM2	-.116	.103	1	-.245	.107	.331
	OK	.058	.082	1	.070	.077	1
SIM2	OK	.174	.097	1	.315	.106	.044

[a] Adjustment for multiple comparisons: Bonferroni.

The results inform us as to how well the iSTART algorithm differentiates between values of Paraphrase Quality for both the original and corrected paraphrases. For example, the algorithm can distinguish significant differences of Paraphrase Quality scores on evaluations that were paraphrases (SIM1, SIM2) and beyond paraphrases (OK) from frozen, IRR, and SH; however, the algorithm does not significantly distinguish Paraphrase Quality comparing paraphrases (SIM2) and better self-explanations (OK) when errors are present. The significant change, when typographical errors are corrected, is that the algorithm then distinguishes differences in Paraphrase Quality between paraphrases (SIM2) and responses that are beyond paraphrases (OK) (p = .044). Additionally, the algorithm detects differences in Paraphrase Quality between frozen expressions and responses that are too short (SH) (p = .032). The ability to make these distinctions is important because it means that the corrected feedback responses show a better association with human ratings. Thus, these results were in line with our predictions, indicating that error correction allows the iSTART algorithm to better discern mere paraphrases from better self-explanations. When errors are corrected, the feedback from iSTART is more comparable to human ratings; thus, the ITS achieves greater similarity to a human tutor.

Discussion

Our goal in this study was to assess the effect of error correction on the ability of computational algorithms to accurately evaluate paraphrases in iSTART. Overall, the results of this study show that when error correction is incorporated into user-language, the algorithms in the NLP system can provide more appropriate feedback to users. We

can also reasonably deduce that ITSs that are unlike iSTART, not using any system of correction (e.g. AutoTutor, [2]) will be more affected. Although there was agreement across most of the compared paraphrases, approximately 12% of the paraphrases were misclassified. Misclassifications can have both positive and negative ramifications. From a motivational standpoint, it may be better to evaluate a paraphrase too highly, so that the feedback generated to the user is more positive and encourages the user to continue. However, from an accuracy standpoint, the algorithm misses something important, because user responses that are too similar (i.e., mere paraphrases) can pass as sufficiently different because of the errors. Thus, mere paraphrases may be misevaluated as more masterful self-explanations during other practice modules.

Results in our previous study [18] indicated that much of the variance was attributable to misspelled words from the target sentence. These *internal misspellings* were the most frequent error type ($N = 665$) and were the most frequent error present in evaluations that changed ($N = 158$ out of 244 changes; 65%). We likewise found that internal misspellings accounted for a large portion of the variance in the LSA index (approx. 35%). The LSA component of the iSTART algorithm is a benchmark approach, requiring correct spelling to correctly match the input and the target concept in the LSA space. As previously mentioned, one reason that ITSs might not correct for spelling is because the statistical approaches such as LSA are expected to be more resistant to individual misspelled words. Our results indicate that LSA and other similarity indices are affected by user errors, contrary to this assumption. As for the word matching component, the literal match undoubtedly gained from correcting misspelled words, as did the soundex match. Contemporary computational approaches are trained on edited data sets and presumably applied to ITS under the assumption that there is sufficient text for the approach to supply appropriate feedback. The results of this study suggest that those algorithms and approaches might be problematic when applied to systems that typically expect short responses, because those responses are keyed in by users with relatively poor writing skills, and because the responses are relatively short, meaning that there is less opportunity for "good" text to washout the effect of "bad" text. Our results may also apply to other ITSs that use comparable matching techniques or are intended for similar populations. Future research should address these issues.

The practical implication from this study is that ITSs can benefit from simple and inexpensive spell-checking programs. Although this conclusion is certainly well-preceded [10, 11, 12], many previous studies were conducted with artificial datasets or with datasets from more proficient populations. In contrast, the present corpus represents a user population that is more typical of producing the kind of spelling that ITSs and their respective algorithms will have to address. It is yet to be seen whether more erroneous data would prove problematic for existing correction techniques (e.g., [11, 12]). Because we attribute the bulk of the observed effects to misspellings of words that are in the target sentence (e.g., rather than grammar), our future research will apply these results to develop an approach for automated spelling correction based on the target.

This research is important because assessing the relative accuracy of NLP algorithms makes both theoretical and applied contributions to the field. However, these tests need to be conducted in contextually valid ways. If real-world input is different from idealized input, then researchers need to know if that factor influences conclusions about NLP accuracy. Otherwise, those conclusions may be invalid. Our results suggest a need for additional research on automatic error correction in NLP.

Acknowledgements

This research was supported by NSF (IIS-0735682) and IES (R305A080589, R305G020018-02). Opinions, findings, conclusions, or recommendations expressed in this material do not necessarily reflect the views of NSF or IES. The authors also thank Vasile Rus, Ben Duncan, Christina Sinquefield, and Sidney D'Mello.

References

[1] R. Azevedo & R. M. Bernard, A meta-analysis of the effects of feedback in computer-based instruction, *Journal of Educational Computing Research* **13** (1995), 111-127.

[2] A.C. Graesser, D.S. McNamara, & K. VanLehn, Scaffolding deep comprehension strategies through Point&Query, Autotutor, and iSTART, *Educational Psychologist* **40** (2005), 225-234.

[3] D.S. McNamara, C. Boonthum, I.B. Levinstein, & K.K. Millis, Evaluating Self-Explanation in iSTART: Comparing Word-based and LSA Systems, In T. Landauer, D.S. McNamara, S. Dennis, & W. Kintsch eds, *Handbook of Latent Semantic Analysis*, Lawrence Erlbaum, Mahwah, NJ (2007), 227-241.

[4] K.R. Koedinger & J.R. Anderson, Intelligent tutoring goes to school in the big city, *International Journal of Artificial Intelligence in Education* **8** (1997), 30-43.

[5] T. Landauer, D.S. McNamara, S. Dennis, & W. Kintsch eds, *Handbook of Latent Semantic Analysis*, Lawrence Erlbaum, Mahwah, NJ, 2007.

[6] D.S. McNamara, I.B. Levinstein, & C. Boonthum, iSTART: Interactive strategy trainer for active reading and thinking, *Behavior Research Methods, Instruments, and Computers* **36** (2004), 222-233.

[7] V. Rus, M. Lintean, P.M. McCarthy, D.S. McNamara, & A.C. Graesser, Paraphrase identification with lexico-syntactic graph subsumption, In *Proceedings of the 21st International Florida Artificial Intelligence Research Society Conference* (2008), 201-206.

[8] P.W. Foltz, S. Gilliam, & S. Kendall, Supporting content-based feedback in online writing evaluation with LSA, *Interactive Learning Environments* **8** (2000), 111-129.

[9] P.W. Foltz & A.D. Wells, Automatically deriving readers' knowledge structures from texts, *Behavior Research Methods, Instruments & Computers* **31** (1999), 208-214.

[10] V. Rus, P.M. McCarthy, D.S. McNamara, & A.C. Graesser, Natural language understanding and assessment, In J.R. Rabuñal, J. Dorado, A Pazos eds., *Encyclopedia of Artificial Intelligence*, Idea Group, Inc., Hershey, PA (2008), 1179-1184.

[11] D. Fossati & B. Di Eugenio, I saw TREE trees in the park: How to correct real-word spelling mistakes. In *Proceedings of the 6th International Conference on Language Resources and Evaluation* (2008), 896-901.

[12] M.A. Elmi & M. Evens, Spelling correction using context, *Proceedings of the 17th International Conference on Computational Linguistics* (1998), 360-364.

[13] P.M. McCarthy, V. Rus, S.A. Crossley, S.C. Bigham, A.C. Graesser, & D.S. McNamara, Assessing entailer with a corpus of natural language, In *Proceedings of the 20th International Florida Artificial Intelligence Research Society Conference* (2007), 247-252.

[14] K. VanLehn, A. C. Graesser, G. T. Jackson, P. Jordan, A. Olney, & C. P. Rose, When are tutorial dialogues more effective than reading, *Cognitive Science* **31** (2007), 3-62.

[15] D. Schneider & K.F. McCoy, Recognizing Syntactic Errors in the Writing of Second Language Learners. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and the 17th International Conference on Computational Linguistics* **2** (1998), 1198-1204.

[16] K. VanLehn, P.W. Jordan, C.P. Rosé, D. Bhembe, M. Böttner, A. Gaydos, M. Makatchev, U. Pappuswamy, M. Ringenberg, A. Roque, S. Siler, & R. Srivastava, The architecture of Why2-Atlas: A coach for qualitative physics essay writing, In *Proceedings of the Intelligent Tutoring Systems Conference* (2002), 159-167.

[17] P.M. McCarthy & D. S. McNamara, (2008). The user-language paraphrase challenge. https://umdrive.memphis.edu/pmmccrth/public/Paraphrase Corpus/Paraphrase_site.htm.

[18] A.M. Renner, P.M. McCarthy, & D.S. McNamara, Computational considerations in correcting user-language in an ITS environment. In Proceedings of the 22nd International Florida Artificial Intelligence Research Society Conference. (2009), pp. 278-283. Menlo Park, CA: The AAAI Press.

[19] D.E. Knuth, *The art of computer programming, Vol.3: Sorting and searching*, 3rd ed., Addison-Wesley, Reading, MA, 1998.

[20] J. Foster & C. Vogel, Parsing ill-formed text using an error grammar, *Artificial Intelligence Review: Special AICS 2003 Issue* **21** (2004), 269-291.

The episodic memory metaphor in text categorisation with Random Indexing

Yann Vigile HOAREAU[a,] and Adil EL GHALI[b] and Denis LEGROS[a]
[a] *CHArt –Laboratory of Human & Artificial Cognition*
Ecole Pratique des Hautes Etudes – University of Paris 8
2, rue de la Liberté 93526 St Denis Cedex 02- France

[b] *Edelweiss Project Team*
French National Institute for Research in Computer Science and Control
2004, route des Lucioles 06902 Sophia Antipolis

Abstract. A model of episodic memory is derived to propose algorithms of text categorization with semantic space models. Performances of two algorithms are contrasted using textual material of the text-mining context 'DEFT09'. Results confirm that the episodic memory metaphor provides a convenient framework to propose efficient algorithm for text categorization. One algorithm has already been tested with LSA. The present paper extends these algorithms to another model of Word Vector named Random Indexing.

Keywords. Random Indexing, episodic memory, text-mining, categorization

Introduction

Since its early introduction, the model that is now named Latent Semantic Analysis (Landauer & Dumais, 1997) has been proposed as a method of matrix reduction and vectorial representation of information for indexing textual documents. The model was known as Latent Semantic Indexing (Deerwester, Dumais, Furnas, Landauer & Harshman, 1990) at that time. Originally only concerned by indexing tasks, LSA has been extended to the area of human memory simulation. Researchers in cognitive psychology got interested in it and then proposed it as a plausible model of human behavior in different tasks such as synonymy test (Landauer & Dumais, 1997) and problem solving (Quesada, Kintsch & Gomez, 2002). The most famous application in cognitive psychology is the coupled CI-LSA model of text comprehension (Kintsch, 1998), which combines the previous "Construction-Integration" model of reading (Kintsch, 1988) with LSA as model of semantic memory. Whereas research involving LSA has been split in two main fields with the text-mining on the one hand and cognitive psychology on the other hand, our paper deals with both of those fields. Discussions of MINERVA 2 model of human episodic memory (Hinztman, 1984, 1988) allow proposing an operative algorithm for texts categorization.

LSA has been known to perform in synonymy test and other equivalent thematic classification tasks (Landauer & Dumais, 1997). The model has been recently successfully applied on opinion judgment task (Ahat, Lenhart, Baier, Hoareau, Jhean-

Larose & Denhière, 2007). There are very important differences between thematic classification, and opinion judgment classification. Firstly, thematic classification is directly connected to the distributional hypothesis, which states that "words that appear in similar contexts have similar meanings". Here is the reason why LSA is able to find words that *share the same thematic, ie "appear in equivalent contexts"*. Secondly, in opinion judgment classification, different thematics could possibly belong to the same category of opinion. For example, I have a good opinion of different movies, which do not deal with the same topic. If I write texts in which I give my opinion of each movie, those texts will be influenced by the topic of the movie for a part, as well as by my motivation to exhibit how and why I loved them for another part. In consequence, the basic application of the distributional hypothesis cannot account for judgment opinion task.

In this paper, we will explore two lines of investigation. In the first line, we will propose the paradigmatic breakthrough that has been realized to find a solution to the limitation of the basic application of the distributional hypothesis. This breakthrough consists in switching from the semantic memory research field to the episodic memory metaphor to drive the similarity comparison stage. The episodic memory metaphor has been tested with LSA (Ahat, & al, 2007). The second line that will be developed in this paper will consist in testing the episodic memory metaphor with an alternative method of Words Vectors construction, named Random Indexing. The proposed algorithm will be tested to categorized large-scale corpus in function of the subjectivity or objectivity they express. Typically a text that expresses facts is considered as objective and a text that expresses opinions is considered as objective. The capability to detect if a text deals with facts or opinions constitutes original application in learning resources classification. For example, on a one hand texts that should be read to learn *what people think about the global warming*, and on the other hand, texts that should be read to learn *what is the global warming*.

Abstractive versus non-abstractive models of memory

In the debate within cognitive psychology about the distinction between "abstractive" versus "non-abstractive" models of memory (Rousset, 2000; Tiberghien, 1997), LSA has been proposed as belonging to the abstractive family (Bellissens, Thérounane, & Denhière, 2004). This proposition is congruent with the affirmation by Landauer, Foltz and Laham that "the representations of passages that LSA forms can be interpreted as abstractions of "episodes", sometimes of episodes of purely verbal content such as philosophical arguments, and sometimes episodes from real or imagined life coded into verbal descriptions" (Landauer, Foltz, & Laham, 1998: 15). Tiberghien considers that "it would be more precise and theoretically more adequate, to consider that all the models are 'abstractive' but, for some of them this abstractive process happens during encoding and for some others it happens during retrieval" (Thiberghien, 1997: 145). Because the abstractive process occurs during encoding, LSA and other Word Vector models are categorized as belonging to the abstractive model family.

A model like MINERVA 2 or other Multiple-Trace models are considered as "non-abstractive" because the abstractive process occurs during retrieval. According to MINERVA 2, memory consists of events or episodes that are represented and stored as vectors. The activation value of each coordinate stores features of episodes. Each

vector corresponds to an episode in the system's life. Retrieval consists of a two stages calculation. First, a similarity calculation is carried out between the probe-vector and all the episode-vectors in memory (see Eq 1). Episodes that are most similar will be affected by a higher level of activation than episodes that are least similar. Second, a calculation is made to compare the level of activation of each feature and this corresponds to the "echo" phenomena of memory. The "echo" calculation produces a new vector that inherits the features of the most activated vectors, even those parts that did not actually exist in the probes. The "echo" has two components: intensity which is denoted *I* (see Eq 2), and content which corresponds to the sum of the content of all traces in memory, weighted by their activation level (see Eq 3). "Echo" constitutes the process of abstraction that Rousset (2000) qualified as "re-creation.

$$S_i = \sum_{j=1}^{N} \frac{P_j T_{i,j}}{N_i}$$

Eq 1 Similarity of a trace i, where P_j is the value of feature j in the probe, and $T_{i,j}$ the value of feature j in trace i

$$\mathbf{I} = \sum_{i=1}^{M} A_i, \text{ where } A_i = S_i^3$$

Eq 2 Intensity of the « echo »

$$C_j = \sum_{i=1}^{M} A_i T_{i,j}$$

Eq 3 The content of the « echo »

The episodic memory metaphor for similarity judgment algorithm

The algorithm used by Ahat & *al* (2007) in Deft07 to identify opinion judgment expressed by unknown texts, consisted in creating a target vector for each type of opinion that should be identified. These target vectors are created by the sum of vectors of all documents that belong to a given category of opinion. For example, the target vector that was used to identify "good critics of movies" was a summed vector of all documents known to be a "good critic of movie". In-comings "text-to-be-indexed" were compared to the target vectors of each category of opinion. Then, the text was categorized with the opinion of the target vector to which it was the more similar. The comparison of similarity used the calculation of the cosine of the angle between the vector of the "text-to-be-indexed" and the target vector. The use of cosine calculation makes it possible to compare the very large target-vectors (hundreds of documents) to the very small text-to-be-indexed vector (one document).

The intuition that was underlying the construction of these very large target-vectors was that the classical distributional hypothesis approach has to be derived to perform in opinion judgment task. The idea was to sum vectors of all documents corresponding to a given opinion category to take advantage of the great number of documents to draw a vector that (i) would not correspond to any topic in particular, and in contrast, (ii) would hold information that would correspond to the linguistic way a given opinion is statistically expressed in numbers of texts. Applying the Multiple-Trace approach specifically to the stage of similarity comparison makes it possible to consider a target vector as an episodic memory that should behave like MINERVA 2

model predicts. Indeed, in considering each document as a specific episode, target-vectors become episodic memories, which are constituted of different episodes of the same category of opinion. As described above, the calculus of "echo" of MINERVA 2 predicts that the more a probe is similar to great numbers of episodes, the more the memory system would react by a strong value of "echo". It is neither mathematically nor psychologically wrong to consider that the value of "echo" in MINERVA 2 and the value of the cosine in LSA behave and can be interpreted in the same way. In consequence, MINERVA 2 gives a theoretical basement to our first intuitive method of vector target construction. The large size target vector method functioned pretty well and contributed to rank second in the Deft07.

Target-vectors as homogeneous episodic memories

The use of the episodic memory metaphor accounts for the limitation of the basic application of the distributional hypothesis for opinion judgment task. In creating these large target vectors, we are creating episodic memories, which behaviors became understandable with the MINERVA 2 model. Predictions concerning "echo" involve that the episodic memories will be more sensitive to probe episodes that are well represented in the memory and less sensible to probe episodes that are less represented. In other words, target vectors will be more sensible to typical documents and less sensible to non-typical documents. Theories of categorization (Rosch & Mervis, 1975) showed that some items are typical of the category they belong, others are not. The typicality of an item is generally defined as (i) a high similarity with items of a given category and (ii) a low similarity with items of other categories.

Target vectors have been produced with the aim of creating episodic memories, which would hold the statistical linguistic signature of a given category of opinion. "Echo" predicts that target vectors will not identify non-typical documents as well as typical documents. We assume that a homogeneous episodic memory, which holds non-typical documents of a given category will be more sensitive to non-typical documents than a heterogeneous episodic memory, which holds typical and non-typical documents, all blended.

Our hypothesis has been implemented for the DEFT09. The aim of the task 1 was to classify texts that express facts or opinions, respectively corresponding "objective" *versus* "subjective" categories.A category of text is represented by a single prototype or different sub-prototypes. In the first case, the prototype vector of a docum ent's category is simply the sum of all document vectors of a category. In the second ca se, a category is represented by a set of sub-prototypes, the number of sub-prototypes is a parameter of the model *t*. To obtain these sub-prototypes, we first split the set of all th e documents ordered by the distance between a document and the single prototype of th e category *C* into *t* sub-sets. This partition ensure that each sub-set contains uniform do cument regarding their distance to the category prototype. A sub-prototypes vector is th en obtained by the sum of the vectors of the corresponding sub-sets. The similarity bet ween the i^{th} sub-prototype P_C^i and a document *d* is given by the cosine between the vect ors representing respectively the document and the sub-prototype.
In our model, the categorization of a given document *d* between *N* categories is given b y the higher similarity between the document *d* and all *t* sub-prototypes of all *N* categor ies, as follows: for each *i* in [1,*t*] we compute the similarities between our document *d* a

nd the i^{th} sub-prototype P_{Ck}^{i} of each category C_k for k in [1,N]. We order these similariti es and assign to each category C_k a duel score for the rank i corresponding to its positio n in the reversly ordered set of similarities. We define the score of a category C_k as the sum of its duel scores. The document d is assigned to the category having the highest sc ore. The figure 1 shows un example of the application of the method with two categorie s and three sub-prototypes

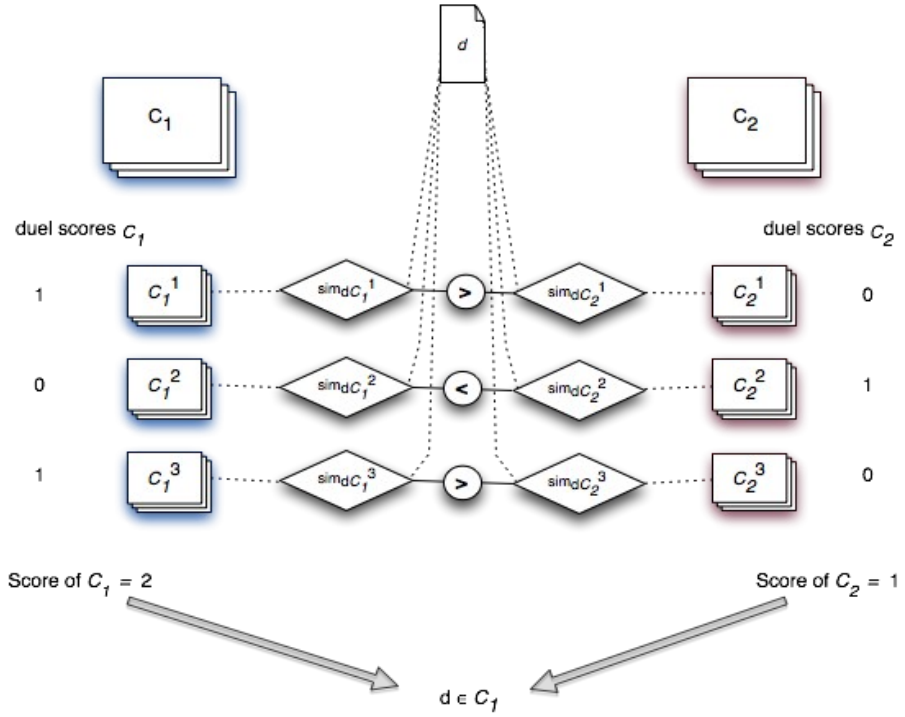

Figure 1. *Application of the sub-algorithm vector with two categories and three sub-prototypes*

Random Indexing

Word vectors correspond to a family of models in which LSA is the most known. Several principles are common to all of these models (see Sahlgren, 2006):

They are based on the distributional hypothesis

They involve a method of counting words in a given unit of context

They have a statistical method, which abstracts the meaning of concepts from large distributions of words in context

They use a vectorial representation of word meaning.

As we will see, Random Indexing is not a typical item of its category. In the other models, the list of principles enounced above is also the stages of a semantic space construction. Particularities of the Random Indexing (RI) model are that (i) it does not create co-occurrence matrix (but it is possible if needed) and (ii) it does not need heavy statistical treatments like SVD for LSA. Contrary to the other Word Vector models, RI

is not based on statistics but on random projections. The construction of a semantic space with RI is as follows:

> Create a matrix A (*d* x *N*), containing *Index vectors*, where *d* is the number of documents or contexts and *N*, the number of dimensions (*N* > 1000!) decided by the experimenter. *Index vectors* are sparse and randomly generated. They consist in small numbers +1 and -1 and thousands of 0.
>
> Create a matrix B (*t* x *N*), containing *term vectors*, where *t* is the number of different terms in the corpus. Set all vectors with null values to start the semantic space construction.
>
> Scan each document of the corpus. Each time a term *t* appears in a document *d*, accumulate the randomly generated *d-index vector* to the *t-term vector*.

At the end of the process, *term vectors* that appeared in similar contexts have accumulated similar *index vectors*. There is a training cycle option in the model. When the scan has been computed for all documents, the matrix B is charged for all *term vectors*. Then a matrix A' (*d'* x *N*), with *d' = d* can be computed with the output of *term vectors*. The number of training cycle is a parameter in the model. The training process output is consistent with what has been described for neural network learning. The RI model has performed in TOEFL synonymy test (Kanerva et al., 2000; Karlgren and Sahlgren, 2001) as well as in text categorization (Sahlgren & Cöster, 2004).

Experiment

The experiment reported here has been realized for the learning stage of the task 1 of the DEFT09[1]. The purpose of the task 1 was the detection of the subjectivity or objectivity character of a text. As described by the committee, "the reference is established by projecting each section on both the subjective and the objective dimension. For instance, the Letter from the editor, which usually states an opinion, has the type subjective, while the News, describing actual facts, have the type objective"[2]. In the learning stage, 60% of the total corpus is given to each team engaged to allow them to implement algorithms that will then be applied on the 40% of uncategorized documents during the test stage. We realized our learning session in using 90% of the learning corpus to build our "machine". The other 10% were used to test and upgrade our categorization algorithm (see Table 1).
Table 2 resumes parameters of the semantic spaces built and tested with two different algorithms of similarity comparison: the *target vector* algorithm *versus* the *sub-target vector* algorithm.

[1] http://code.google.com/p/semanticvectors/

[2] http://deft09.limsi.fr/index.php?id=1&lang=en

Material

Table 1. Number of documents and size of each types of documents used for learning and test.

	Learning stage		Test	
	Number of docs	**Size (Ko)**	**Numbers of docs**	**Size(Ko)**
Objective	18757	97768	2080	10780
Subjective	3901	23460	437	2596

Results

Table 2. Precision performances in function of the numbers of dimensions, cycles and sub-target vectors.

Parameters of algorithms			Precision		
Number of dimensions	**Number of cycles**	**Number of sub-target vectors**	**Objective**	**Subjective**	**Total**
Target vector algorithm					
1000	10	1	27%	97%	40%
1000	15	1	33%	97%	45%
Sub-target vector algorithm					
1000	10	3	94%	54%	86%
1000	10	5	93%	55%	87%
1000	15	5	94%	55%	88%
1500	15	5	95%	57%	89%
1000	10	7	85%	72%	82%
1000	10	9	5%	97%	25%

Precision performances are higher for *Sub-target vector algorithm* than for *Target vector algorithm*. The *Sub-target vector algorithm* gives best results with 5 sub-targets. Those results involve that there is an optimum threshold for the number of sub-target vectors. Considering Multiple-Trace approach, this threshold corresponds to the moment where episodic memories or sub-targets are the most homogeneous.

First, this experiment demonstrated that the episodic memory model provides a good theoretical framework to the *target vector algorithm* that has been proposed for the DEFT07. Second, as predicted by Minerva 2 model, the more targets are homogeneous, the more they perform. According to that, *Sub-target vectors algorithm* performed better than *Target vector algorithm*. Third, whereas *Target-vector algorithm* has been applied with LSA, we applied both *Target vector* and *Sub-target vector algorithms* with Random Indexing.

Conclusion

Target vector algorithm consisted in creating a very large vector composed of each and every documents of a given category as target vector used to identify the category a document belongs to. The proposed theoretical switching from abstractive to non-abstractive model of memory has been described and successfully tested to account for the *Target-vector algorithm*. Those large target vectors have been considered as episodic memories and MINERVA 2 has been used as a metaphor to predict and

interpret behaviors of such episodic memories. Computing the *Sub-target algorithm* with different numbers of more homogeneous sub-targets has confirmed predictions derived from the "echo" calculation of Minerva 2.

Whereas the *Target vector algorithm* has been tested with LSA, both *Target* and *Sub-target vectors algorithm* have been extended to Random Indexing.

The principle of using MINERVA 2 to drive the comparison stage in Word Vector models should not be limited to opinion categorization task. In the field of automated essay scoring, the example of speech register identification is a very close task where the episodic memory metaphor and the *Sub-target vector algorithm* could easily be applied.

References

M. Ahat, W. Lenhart, H. Baier, Y. V. Hoareau, S. Jhean-Larose, & G. Denhière,(2007) "Le concours DEFT'07 envisagé du point de vue de l'Analyse de la Sémantique Latente (LSA)". In *Proceedings of the conference DEFT'07*, Grenoble, 2007.

C. Bellissens, P. Thérouane, & G. Denhière, Les modèles vectoriels de la mémoire sémantique : description, validation et perspectives, *Le Langage et L'Homme*, **34** (2004), 101-122.

S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, & R. Harshman, Indexing By Latent Semantic Analysis, *Journal of the American Society For Information Science*, **41**(1990), 391-407.

D. L. Hintzman, MINERVA 2: A simulation of human memory, *Behavior Research Methods, Intruments, & Computers,* **16** (1984), 96-101.

D. L. Hintzman, Judgments of frequency and recognition memory in a Multiple-Trace Model, *Psychological Review,* **95** (1988), 528-551.

P. Kanerva, J. Kristoferson, & A. Holst, Random Indexing of Text Samples for Latent Semantic Analysis, In L.R. Gleitman, and A.K. Josh (Eds.), *Proceedings of the 22nd Annual Conference of the Cognitive Science Society*, Lawrence Erlbaum Associates, Mahwah, 2000.

J. Karlgren, & M. Sahlgren, From Words to Understanding, In Y. Uesaka, P. Kanerva, & H. Asoh (Eds.) *Foundations of Real-World Intelligence*, CSLI Publications, Stanford, 2001.

W. Kintsch, The role of knowledge in discourse comprehension : a construction-integration model, *Psychological Review,* **95** (1988), 163-182

W. Kintsch, *Comprehension: a paradigm for cognition,* Cambridge University Press, Cambridge, 1998.

T. K. Landauer, LSA as a Theory of Meaning. In T. K. Landauer, D. McNamara, S. Dennis, W. Kintsch (Eds*). The Handbook of Latent Semantic Analysis,* Lawrence Erlbaum Associates, Mahwah, 2007.

T. K. Landauer, & S. T.Dumais, A solution to Plato's problem: The Latent Semantic Analysis theory of the acquisition, induction, and representation of knowledge, *Psychological Review,* **104** (1997), 211-240.

T. K. Landauer, P. W. Foltz, & D. Laham, Introduction to Latent Semantic Analysis, *Discourse Processes,* **25** (1998), 259-284.

J.F. Quesada, W. Kintsch, & E. Gomez, A theory of Complex Problem Solving using Latent Semantic Analysis, In W. D. Gray & C. D. Schunn (Eds.) *Proceedings of the 24th Annual Conference of the Cognitive Science Society,* Lawrence Erlbaum Associates, Mahwah, NJ, 2002.

E. H. Rosch, & C. B. Mervis, Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology,* **7** (1975)**,** 573-605.

S. Rousset, Les conceptions "système unique" de la mémoire : aspect théorique, *Revue de neuropsychologie,* **10** (2000)*,* 30-56.

M. Sahlgren, The Word-Space Model: Using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces. *Ph.D. dissertation*, Department of Linguistics, Stockholm University, (2006).

M. Sahlgren, & R. Cöster, Using Bag-of-Concepts to Improve the Performance of Support Vector Machines in Text Categorization, *Proceedings of the 20th International Conference on Computational Linguistics*, Geneva, 2004.

G. Tiberghien, *La mémoire oubliée*, Mardaga, Liège, 1997.