Automated Content Alignment for Adaptive Personalized Learning

Elliot Robson
Eduworks
New York, NY
Elliot.Robson@eduworks.com

Robby Robson
Eduworks
Corvallis, OR
Robby.Robson@eduworks.com

ABSTRACT

Effective learning interventions (online courses, SIMS, live instruction, and self-directed activities) must be strongly aligned with instructional goals. Programs such as the P*ersonal Assistant for Learning* (PAL) being developed by the US Advanced Distributed Learning initiative and the *Generalized Intelligent Framework for Tutoring* (GIFT) developed by the Army Research Lab (ARL) emphasize the Government's investment in learning interventions that adapt to learner goals and preferences. To be practical, such systems must automatically detect and align digital content and other learning intervention with learning goals.

The research reported here addresses one step in this process. It is part of the larger integration effort between GIFT and *Tools for the Rapid Development of Expert Models* (TRADEM), supporting the efforts and goals of the Army Research Lab (ARL). This paper presents techniques that automatically use a set of text-based features to detect pedagogically appropriate topics. These techniques are part of an attempt to automate portions of the front-end analysis and design steps in the tradition "ADDIE" (analysis, design, development, implementation, and evaluation) [Branson et. al., 1975] approach to content creation. This paper sets the context for this work, describes the techniques and algorithms used, and provides data that shows that auto-detection performs well when reviewed by and compared to hand-generated mappings by instructional design experts.

ABOUT THE AUTHORS

Elliot Robson is director of Eduworks Research and Solutions department. His research interests include statistical and graphical modeling, simulation, and applied techniques in education. He comes from a K-12 background where he led research and development of school-side solutions for Amplify Education. He is an active member of the Data Advisory Board for Horizon's National and a member of the board for Yleana Leadership Academy. He holds a Masters in Public Policy from the Ford School at the University of Michigan and has over 10 years of professional experience in educational technology.

Robby Robson is CEO and Chief Scientist at Eduworks Corporation. His research interests include online learning environments, reusable design, and applications of computational linguistics to learning, education, and training. He chaired the IEEE Learning Technology Standards Committee from 2000 – 2008 and has led multiple National Science Foundation and Department of Defense projects that explored the use of emerging technologies. He serves as a consultant to the Institute for Defense Analysis and holds a doctorate in mathematics from Stanford University and has held posts in both academia and industry.

INTRODUCTION

In recent years, intelligent tutors have been shown to be a highly effective method for providing meaningful targeted training in a flexible and distributed fashion [Dodds & Fletcher, VanLehn, 2004]. To support their use in a rapidly changing world where training content must often be updated to reflect new regulations, situations, technologies, and other dynamic forces, and to lower the cost and time required develop intelligent tutors, Government projects such as the *Generalized Intelligent Framework for Tutoring*, or GIFT [Sottilare, Brawner et. al., 2012] are seeking ways to partially automate the authoring process. As has been stated in conjunction with the currently-running Office of Naval Research "STEM Grand Challenge" and in a recent symposium on GIFT that focused on authoring tools, the goal is to produce intelligent tutoring systems in arbitrary domains that approach the two-sigma gain in learning effect exhibited by human tutoring over classroom instruction [Bloom, 1984].

As part of this program we designed and built an automatic feature detection system that extracts a set of meaningful topics and associated instructional content from a *corpus* of content (i.e., a collection of documents and training materials), and then aligns the nuggets with pedagogical frameworks. This work was in part supported by a U.S. Army Small Business Innovation Research (SBIR) project called TRADEM (*Tools for the Rapid Development of Expert Models*) which has been described in [Robson, R., Ray, F., and Cai, Z., 2013] and elsewhere.

The main focus of this paper is a multi-layered approach to solving a specific problem related to the semi-automatic creation of meaningful topic maps and alignment of topics with corpus content. Among most challenging problems in this process is automatically labeling machine-generated topics. A new set of algorithms was developed for this purpose. This paper starts with a brief overview of the larger process into which this fits, followed by a discussion of the techniques used to solve the topic labelling problem, followed by measures of the effectiveness of the new algorithms and their impact on the more general problem of generating topic networks. This research contributes to the growing body of knowledge resulting from Department of Defense sponsored projects that can lead to more effective training systems that apply advances in computational power and algorithm development to adaptively present and personalize training content.

OVERVIEW

TRADEM is being developed as part of the GIFT research program cited above as a front-to-back automated intelligent tutor generation system. It ingests a corpus of content (e.g., textbooks, training manuals, existing learning materials, documents, etc.) and walks the user through a front-end analysis workflow (described below). The system generates a suggested topic map for the corpus, associates each topic with auto-selected content and auto-generated questions that the user can augment and edit, and then aids the user in creating intelligent tutors for any topic(s) covered in the corpus. This overall process is diagrammed in Figure 1.

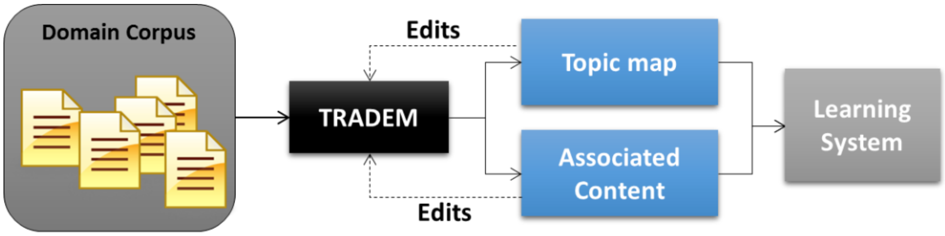

Figure 1: TRADEM Process

Front-End Analysis Workflow

The input for the process in Figure 1 is a corpus of documents. The primary output is a well-structured network of connected topics with associated text, questions, content, and pedagogical strategies. This process uses text mining and natural language processing to support traditional front-end analysis that is part of an instructional system design workflow. The automated (or more accurately, semi-automated) process takes place in five steps (next page):

Step 1: The user selects input corpus of content.

Step 2: The system generates a topic network that fully covers the input corpus. This is accomplished using a combination of text analysis, natural language processing and machine learning methods.

Step 3: The system associates each topic in this network with small chunks of material from the initial corpus based on relevance and its usability as instructional material in an intelligent tutor.

Step 4: A different part of the system transforms the content it has selected into course content for each topic.

Step 5: The system identifies how the teachable course content is aligned to instructional strategies across several pedagogical frameworks using metadata and feature detection algorithms.

Of these, Step 2 (topic generation) is critical because it produces descriptions of topics that are used by later algorithms to select, align, and transform content subsequent steps.

TOPIC NETWORK GENERATION

In the topic generation step the system extracts an editable network of topics from the input corpus. Ideally, this network accurately describes the domain and its pre-requisite structure, even though users have a chance to make edits before continuing. Three types of techniques were developed that led to improved results: preprocessing techniques, modified semantic analysis, and topic definition algorithms.

Preprocessing Techniques

The first step is to clean the input data through automatic and rule-based filters. This uses multiple methods, the first of which is *granulation*.

Granulation: The size of the corpus is reduced to remove unnecessary noise and improve the specificity topic extraction algorithms. This is done by breaking the corpus into small *granules* detected based on natural break points in the text and other factors. Granules vary in size from a single sentence to a full paragraph. For example, a 2 page paper, on average, breaks down into ten granules. The granule level is considered to be the smallest level where meaningful content and topic affiliations can exist, and granulation is necessary to assemble and deliver content in multiple types of learning systems. Thus, granules are the core unit of analysis, and rule-based filters are applied to granules.

Granulation is critical to everything that follows, but other rules are added to improve the performance of algorithmic topic detection. For example, when working with a collection of combat medic training content (Army Military Occupation Specialty 68W training), we discovered that acronyms caused problems, and this generalizes to most military content. To address this we added acronym detection filters to preprocessing. These enabled the system to better recognize and match phrases in both their acronym and expanded form.

Throughout the development process, improvement in pre-processing yielded significant improvements in the quality of the topics extracted from a corpus. These enhancements, according to our pedagogical expert, substantially increased the quality of the topics. As an example, out of the twenty-four automatically generated topics in the 68W training, three topics that previously failed to provide useful pedagogical coverage of the source material were replaced with high quality, meaningful, topics.

Modified Semantic Analysis

Once the preprocessing filters have been applied, we use semantic analysis to detect topics. The granules are split into *n*-grams (i.e. sequences of *n* words) of multiple sizes and data mining techniques are applied. These techniques output topics in the form of ordered word lists ranked by relative importance of the word to the topic. Previously, these lists themselves were considered to be topics. This is consistent with standard practice in other fields that use similar techniques [Blei, Ng, *et al*, 2002], but we found that ordered lists of words were too abstract to be interpreted by users as genuine topics and did not have pedagogical meaning. For example, our original system generated topics that looked like the following:

- blood >pressure > cuff = reading >systolic …
- medication >> dose > prescribed >medications …
- burns > skin > burned > degree > water …

Topics in this raw form can be used by machine algorithms, but they are not acceptable to a user trying to create a lesson plan or understand the pedagogical coverage of a corpus. To transform these raw word lists into meaningful topics that align with pedagogical needs and user expectations, a set of new algorithms was created to transform this raw analytic output into real, human-readable, topics.

Topic Labeling Algorithms

These topic labelling algorithms were a major breakthrough and are described here.

Machine generated topics can be correctly labeled in many different ways, but only a handful of the potential labels will be appropriate and useful as part of a front-end analysis or for generating adaptive learning systems. To find a good set of topic labels, we first generate a candidate list. Other researchers [Shen, Zhai et. al.; Magatti, Calegari, et. al., 2009] have worked on this problem, but their approaches assume that the nature of the corpus is known and that a structured glossary or table of contents is available. In our experience, cases where the corpus comes with a pre-defined dictionary of potential topics are rare. One of the key requirements for GIFT and for other DOD projects such as the ONR STEM Grand Challenge and the ADL PAL project is that they be applicable to new domains in which formalized instruction has not been developed, or to old domains with significant new information. In general, it is impossible to “assume that the set of candidate labels can be extracted from a reference text” as is done by [Shen, Zhai, et.al, 2007].

With a pre-defined list of topic labels to guide us, we needed to define criteria for determining a good set of labels. We defined “good topic labelling” as one that both *described* and *disambiguated* the underlying word lists that defined topics. “Describe” means that each topic label should accurately encompass the hierarchical word list output from our topic detection techniques. “Disambiguates” means that each label should be as specific to its word list as possible and distinguish different lists as much as possible.

To meet the first requirement (“describe”), the base word list in each topic was compared to the full corpus to create a very large set of potential topic labels. This was done by creating a list of every phrase from the full corpus that contained a word from the topic list, and this phrase list was expanded to include words with similar meanings. For example, if the list contained the word “education,” a potential topic list might include the phrases “elementary education,” “specialized education,” “targeted training,” etc. Using these phrases as a base set, a relevance score was assigned to each topic label based on the frequency and rank of the words in machine-generated topic. The score also considered weights for label length and specificity of words. This generated a per-topic preference pool of potential topic labels. The following is an example of labels generated for a topic:

Topic:	blood >pressure > cuff = reading >systolic …
Labels:	(1) blood pressure (2) systolic blood pressure (3) medical equipment etc...

Generating such per-topic pools of labels is useful, but many of the labels generated are over-generalized and do a poor job of disambiguating among topics. For many topics, the most common and most highly ranked words appear in many phrases, causing large blanket topic labels to be ranked more highly than is desired. For example, when we used the I/ITSEC 2013 paper set as a corpus, *training simulation* appeared as a highly ranked label for almost 20% of the topics. It is true that many of the I/ITSEC papers were about training and simulation, but this label is not useful for a user trying to understand the differences among topics. At the same time, there are often large overarching topics within a corpus, and forcing them into a narrow or specific label can lead to topic labels that do not match the actual content associated with the topic.

To satisfy the requirement for disambiguation, we used a lottery system for topic labels. Each topic was allowed to assign a preference for each label in its pool of potential labels. Topics were allowed to do this in an order related to the word lists that defined the topics. These algorithms used to produce these lists also output “strengths” for each word in a list. We ranked topics based on the strength of their words and allowed the topics to pick labels based on this ranking. Topic labels are removed from the overall candidate pool in a round-robin fashion with the most highly ranked topics picking first and least common topics picking last. This let overarching and powerful topics find the broad topic labels that match them well while removing these labels from the pool so that more specific topics were labelled in easily understood and highly discriminating ways.

TOPIC DETECTION RESULTS

Two methods were used to test the automated topic naming strategy discussed above. The first method compared automatically generated topic labels with labels manually generated by an expert with knowledge of the domain. The second method asked an expert with an understanding of the domain pedagogy to give each machine-generated label a topical relevance score. The first experiment confirmed that the machine-generated generated labels were similar to those an expert would choose in most cases. The second verified that the labels are appropriate and useful to an end user. The results of these tests are reported next.

Test 1: Comparison with Expert Labels

We used thirty-four topics extracted from a 68W Army medical training corpus for this test. This corpus was provided to us by Engineering and Computer Simulations, Inc., who had developed simulations using this corpus [Sotomayor, 2010]. An expert who had not been involved in any of the algorithm design or automated topic generation was asked to hand-label all thirty-four of the topics and assign a numerical score to each label representing how well the auto-generated label covered the topic’s ranked word list. As is common in experiments of this type [Manning, Raghavan, Schütze, 2008], we consider the expert opinion to be the ground truth. An example of the results obtained is given in Table 1 on the next page.

Table 1; Comparison of Expert Labels to algorithm-generated Labels

Topic Word List	Expert Label	Algorithm Label
blood >pressure > cuff = reading >systolic …	Measure Blood Pressure (8)	● blood pressure ● systolic blood pressure ● systolic blood ● direct pressure
medication >> dose > prescribed >medications …	Medication Prescriptions (9)	● correct medication ● medication administration ● topical medication ● medication sheet
burns > skin > burned > degree > water …	Burns (7)	● third degree burns ● second degree burns ● burned area ● serious burns

Comparing all thirty-four topics, there was a label that the expert rated as equivalent to or better than their label within the top four automatically generated labels in twenty eight of the cases. Thus, 82% of the time, the machine-generated label agreed with and matched expert expectations. In four of the six remaining cases, the expert noted that, although the algorithmically generated labels were not thematically similar to the hand generated labels, they were high quality labels and potentially preferable given the domain of the content. The second example above is one instance of this. After examining the algorithmically generated labels, the expert felt that *Medication Prescriptions* is a useful and correct label, but looking at the actual content of the corpus showed a context-based preference for a label like *Medication Administration*. In this case, the corpus is of a size that makes expert checking possible, but this is not the case for many of the thousand-document corpora that have been analyzed.

Overall, an 82% top level agreement, with an additional 12% of cases where the expert preferred algorithm-based results given the corpus content, is an excellent result. A 94% accuracy measure is higher than initially expected, so we considered that this may have been partially due to the coherence of the corpus chosen. To investigate this possibility, the same test was performed again using a less focused and larger corpus.

This time, the full paper list from I/ITSEC 2013 was used. The corpus of I/ITSEC papers resulted in over 150 distinct machine-generated topics, so a random subset of 30 was chosen for a different expert to label. As expected, the results for this corpus showed a greater variation between auto-generated labels and expert hand-labels. The expert, unable to scan through the entire corpus and weight each word and phrase based on frequency, relied heavily on the top few words in the distribution to label the topics. This example highlights the difference in the labels generated, as shown in Table 2:

Table 2: Comparison for a Topic from the Corpus of 2013 I/ITSEC Papers

Topic Word List	Expert Label	Algorithm Label
autonomous >different > systems >agent> control …	Autonomous Systems	● autonomous vehicles ● unmanned vehicle ● general architecture ● proposed architecture

The expert had no way of knowing that these word lists primarily came from a set of paragraphs discussing the architecture for unmanned vehicles. As a result, the human generated the label *Autonomous Systems*. Although sensible based on the top ranked words in the distribution, this label missed the true topic as it existed in the corpus. Based on this, and many similar results, a second test of the topic naming algorithm was proposed.

Test 2: Labelling Goodness of Fit

The first test shows that reliance on human experts may not be ideal. This second test asked a pedagogical expert to simply rate the algorithmically generated results against the underlying content. The corpus used was the Army's 68W training corpus used above, which was of a manageable size for the expert. The expert was asked to rate the top four labels for each topic on a 1-5 scale, with 1 being a weak label and 5 being a very strong label given the content of the 68W training. Table 3 shows some sample results:

Table 3: Pedagogical Expert Ratings of TRADEM-generated Labels

Topic Word List	Auto-Generated Label	Label Score (1-5)
blood >pressure > cuff = reading >systolic …	● blood pressure ● systolic blood pressure ● systolic blood ● direct pressure	● 4 ● 3 ● 3 ● 2
medication >> dose > prescribed >medications …	● correct medication ● medication administration ● topical medication ● medication sheet	● 2 ● 4 ● 2 ● 2
burns > skin > burned > degree > water …	● third degree burns ● second degree burns ● burned area ● serious burns	● 3 ● 2 ● 3 ● 5

In 31 of the 34 topics at least one of the algorithm's labels scored a 4 or higher. It is encouraging that our algorithms produced at least one high-quality label for over 90% of the topics, but in only 15 of these cases was the preferred label the top ranked label. This indicates to us that additional improvements are necessary to identify which of the candidate labels is the best. As a result, the actual system we have implemented suggests the top ranked topic label but also provides the user with all of the top four labels so that the user can select one of the alternatives or manually enter a better label if the top suggestion doesn't fit the user's needs.

EFFECT OF AUTOMATED LABELLING ON TOPIC NETWORK GENERATION

Topic labeling algorithms provide a solution for aligning content with topics that works in complex corpora that are larger than a human could process, as well as in smaller corpora. There are still improvements to be made to move the very best labels up in the per-topic rankings, but the current results are better than those produced by any other methods we have seen reported. Overall, algorithmically generated topics have discriminating and useful topic labels. We discovered that this improved the results of subsequent algorithms that produced topic networks. Prior to the improved automatic labeling algorithms, the topic networks were only considered satisfactory by experts after they were hand edited. Using the new labels, we asked an external instructional designer, an internal instructional designer, and an internal pedagogical expert to rate the coverage and alignment of the produced curricular flow on a scale of 1 (poor) to 10 (excellent). The following Tables 4 and 5 (next page) show results:

Table 4: Alignment as Rated by Experts

Coverage of Corpus Content	Original Topic Model	Updated Topic Model
External Instructional Designer	7	9
Internal Instructional Designer	6	10
Internal Pedagogical Expert	7	9

Table 5: Coverage as Rated by Experts

Alignment with Corpus Goals	Original Topic Model	Updated Topic Model
External Instructional Designer	5	9
Internal Instructional Designer	4	8
Internal Pedagogical Expert	6	9

While this is not a rigorous study and ratings were given by only three experts, it is a strong indication that updated topic labeling algorithms have helped the output meet expectations. Based on the strength of the indicators, a more rigorous study design has been proposed.

CONCLUSIONS AND FUTURE WORK

The results reported here demonstrate an automated approach to labelling machine-generated topics. Newly developed topic labelling algorithms were successful in tests, generating quality labels for more than 90% of topics. In practice, this led to much better results for automatically generated topic networks as well. Automated generation of topic networks is at the heart of a program to reduce the time and cost for doing front-end analysis in new knowledge domains and for generating adaptive learning systems such as GIFT tutors for those domains. High quality topics align better with pedagogical goals and require less editing by the user. A more rigorous validation is necessary, but the results justify further investigation in the quality and time savings produced by automated support for upfront analysis, especially in corpora that would take instructional designers weeks or months to analyze. A rigorous user study is planned to expand and verify the results reported in this paper.

Despite the progress reported, improvements to the topic label selection must still be made. In only 44% of the topics was the best label also the most highly ranked label. Additional research is necessary to increase that number to around 70%, a reasonable target based on the literature [Shen, Zhai, et. Al, 2007].

ACKNOWLEDGEMENTS

The research described herein has been sponsored by the U.S. Army Research Laboratory - Human Research & Engineering Directorate (ARL-HRED). Statements and opinions expressed in this paper do not necessarily reflect the position or the policy of the United States Government, and no official endorsement should be inferred.

The authors also thank Keith Brawner of the Army Research Lab for his guidance, Brent Smith of Engineering and Computer Simulations, Inc. for providing combat medic content and advice, and to Brent Smith and Howard Mall for acting as an external scorers for combat medic content.

References

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2002). Latent Dirichlet allocation. *Advances in neural information processing systems*, *1*, 601-608.

Bloom, B. S. (1984). The 2-sigma problem: The search for methods of group instruction as effective as one-to-one tutoring. Educational Researcher, 13(6), 4-16.

Branson, R.K., Rayner, G.T., Cox, J.L., Furman, J.P., King, F.: Interservice procedures for instructional systems development. Executive summary and model. Florida State University, Tallahassee (1975)

Dick, W., Carey, L., & Carey, J.O. (2001). The systematic design of instruction (5th ed.). New York: Addison-Wesley, Longman.

Dodds, P. and Fletcher, J.D., 2004. Opportunities for New "Smart" Learning Environments Enabled by Next-Generation Web Capabilities. Journal of Educational Multimedia and Hypermedia 13, 4, 391-404.

Manning, C. D., Raghavan, P., & Schütze, H. (2008). Introduction to information retrieval (Vol. 1, p. 6). Cambridge: Cambridge university press.

Magatti, D., Calegari, S., Ciucci, D., & Stella, F. (2009, November). Automatic labeling of topics. In *Intelligent Systems Design and Applications, 2009. ISDA'09. Ninth International Conference on* (pp. 1227-1232). IEEE.

Mei,Q., Xuehua, S., Zhai,C.,2007. Automatic Naming of Multinomial Topic Models, Proceedings of 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, University of Illinois at Champaign Urbana

Robson, R., Ray, F., and Cai, Z., 2013. Transforming Content into Dialogue-based Intelligent Tutors. In The Interservice/Industry Training, Simulation & Education Conference National Training and Simulation Association, Orlando, FL.

Sotomayor, T. M. (2010). Teaching Tactical Combat Casualty Care using the TC3 SIM Game-Based Simulation: A Study to Measure Training Effectiveness. Studies in Health technology and Informatics, 154, 176-179

Sottilare, R. A., Brawner, K. W., Goldberg, B. S., & Holden, H. K. (2012). The generalized intelligent framework for tutoring (GIFT). *Orlando, FL: US Army Research Laboratory–Human Research & Engineering Directorate (ARL-HRED)*

VanLehn, K. (2011). The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. Educational Psychologist, 46(4), 197-221.