

MLeXAI: BIOMEDICAL TERM CLASSIFICATION*

*Vasile Rus
Department of Computer Science
Institute for Intelligent Systems
The University of Memphis
Memphis, TN, 38152
901 678-5259
vrus@memphis.edu*

*Ingrid Russell
Department of Computer Science
The University of Hartford
West Hartford, CT 06117
860 768-4191
irussell@Hartford.edu*

*Zdravko Markov
Department of Computer Science
Central Connecticut State University
New Britain, CT 06050
860 832-2711
markovz@ccsu.edu*

ABSTRACT

Machine Learning is an important area of Artificial Intelligence which is generally applicable to almost any field of science. Early exposure of students to the potential of machine learning could have a positive impact on their attitude towards Artificial Intelligence in particular and computer science in general. In this paper, we present a semester long machine learning project that was incorporated in an Introduction to Artificial Intelligence course. An assessment of the impact of the project on students' learning of core concepts of Artificial Intelligence and on their plans to pursue more or advanced classes related to Artificial Intelligence is reported. The conclusion of our study is that the semester-long machine learning project had extremely positive effects on students learning and perception of Artificial Intelligence.

* Copyright © 2009 by the Consortium for Computing Sciences in Colleges. Permission to copy without fee all or part of this material is granted provided that the copies are not made or distributed for direct commercial advantage, the CCSC copyright notice and the title of the publication and its date appear, and notice is given that copying is by permission of the Consortium for Computing Sciences in Colleges. To copy otherwise, or to republish, requires a fee and/or specific permission.

INTRODUCTION

In this paper, we present a machine learning project called *Biomedical Term Classification* that was implemented in the *Introduction to Artificial Intelligence* course taught at The University of Memphis during Fall 2008. All the details about this project, including resources such as data sets, can be found at <http://www.cs.memphis.edu/~vrus/ML-AI/>.

The *Biomedical Term Classification* is part of the larger NSF-funded project *Machine Learning Experiences in Artificial Intelligence* (MLeXAI; <http://uhaweb.hartford.edu/compsci/ccli/>) whose main purpose is to incorporate machine learning [5] as a unifying theme for Artificial Intelligence (AI) courses. The main motivation for the MLeXAI project was the need to address the seemingly disconnected core AI topics (e.g., problem solving by search, first-order logic, uncertainty and probabilistic reasoning) that are typically covered in AI courses. Machine learning is inherently connected with the AI core topics and provides methodology and technology to enhance real-world applications within many of these topics. The basic idea of our approach is to expose students to machine learning through a set of hands-on lab projects.

The machine learning approach we adopted is an alternative to previous attempts to offer a unifying theme to AI courses. Previous attempts are the agent-centered [6] and robotics-centered [2, 3, 4] approaches. One advantage of our approach is that a machine learning application can be rapidly prototyped as opposed to, for instance, robotics-centered approaches that teach through hard experience that a simulated world is often much easier to handle than the real one. Machine learning allows learning to be grounded in engaging experience without limiting the important breadth of an introductory course. Our machine learning emphasis acknowledges that intelligent systems are best taught through their application to challenging problems.

The larger NSF-funded project MLeXAI encompasses a suite of machine learning projects (e.g., recognizing DNA sequences, creating robotic intelligent agents, route planning) proposed by more than a dozen faculty members throughout the country [7]. The full list of individual projects is available at the website of the MLeXAI project: <http://uhaweb.hartford.edu/compsci/ccli/>. In this paper, we focus on the project developed and implemented at The University of Memphis: *Biomedical Term Classification*. The project was implemented in the *Introductory to Artificial Intelligence* course which is a senior and graduate course open to both computer science (COMP4/6720) and electrical and computer engineering majors (EECE4/6720). The major topics covered in the course are: problem solving by search, informed and heuristic search, games playing, knowledge-based agents, first order logic, knowledge representation, uncertainty and probabilistic reasoning, machine learning, and natural language processing.

BIOMEDICAL TERM CLASSIFICATION

BioNLP [1] is a new area of research at the intersection of biomedical and Natural Language Processing (NLP) technologies. A major problem in BioNLP is that same biomedical term can be frequently used with different meanings in biological texts. For instance, the biomedical term SBP2 can refer both to a protein or a gene [7]. If a researcher studies the SBP2 gene, but not the protein, and s/he searches MEDLINE for

articles published recently on the SBP2 gene then there is a high chance s/he would also get articles related to the protein. The researcher would have to manually sort out the gene-related articles from the rest. It would be extremely beneficial if an automated method were available that would classify occurrences of the token SBP2 into genes and proteins. One automated method is to combine NLP with machine learning (ML) techniques to build software tools that classify biomedical terms. In particular, we focused on two well-known machine learning algorithms: decision trees and naïve Bayes. Both algorithms are relatively easy to understand which makes them appropriate for an introductory AI class.

Problem Description

The Biomedical Term Classification project involves the development of classifiers that categorize words or groups of words that denote a biomedical term into five classes: *DNA, RNA, protein, cell_line, and cell_type*.

There are two issues related to classifying biomedical terms in scientific articles. First, the terms must be delimited from the surrounding words in the texts/sentences in which they occur. This first step is called *biomedical term recognition* and is a non-trivial step, if automated. Second, the delimited terms must be *labeled with/classified into* a biomedical class such as DNA or RNA. In our project, we assumed the biomedical terms were already delimited and thus our focus was only the classification step.

Project Phases

As many projects in machine learning, this project was split into three major parts: *data collection, feature extraction, and machine learning*.

Phase 1 – Data Collection

The first phase consisted of collecting a set of sentences containing biomedical terms belonging to the five biomedical term classes we selected: DNA, RNA, protein, cell_line, and cell_type. Because of the nature of the domain, i.e. biomedical, expertise is needed to be able to identify such sentences. While we did not ask students to become experts in biomedicine, we did want to expose students to what it means to collect data. We asked students to query MEDLINE for articles using three keywords: *human, blood cell, or transcription factor*. Students had to identify in these articles sentences containing biomedical terms (of any type not only belonging to the five classes). An example of such a sentence is given below.

IL-2 gene expression and NF-kappa activation through CD28 requires reactive oxygen production by 5-lipoxygenase.

The deliverables for Phase 1 was a list of 5 sentences annotated with delimited and labeled biomedical terms. It was fine if there were annotation errors in these 5 sentences because students were not biomedical experts. These sentences are not used in later phases of the project and thus the annotation errors will not negatively impact the performance of the induced classifiers. Instead, a dataset of 300 sentences has been put

together by the instructor to make sure appropriate data was used in later phases of the project. The 300 sentences contain biomedical terms belonging to the five classes mentioned above. These sentences are a subset of the standard dataset used in the Shared Task of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications

(<http://research.nii.ac.jp/~collier/workshops/JNLPBA04st.htm>).

Phase 2 – Feature Selection and Extraction and Data Preparation

In this phase, every biomedical term delimited and labeled in the data collection step from Phase 1 was mapped onto a features vector representation (explained later), which in turn was used as input for the machine learning phase. The mapping process is detailed in the three steps below.

Step 1: Feature Selection

We characterized each biomedical term occurrence by the sentential context in which it appeared. In other words, we characterized a biomedical term occurrence by the company it keeps, i.e. surrounding words. The number of surrounding words to characterize biomedical terms does matter. A large number of surrounding words would have more discriminative power but would fail to generalize. Too few surrounding words would generalize too much and lead to less discriminative power. We used a window of three words before and after the target term to characterize each instance. Each such word is viewed as a feature of our problem of classifying biomedical terms and the resulting representation is called the features vector representation. In this vector representation, each instance is represented as a vector with n dimensions where n is the number of features used, six in our case (the three words before and the three words after). Because we adopted a supervised learning approach, we must also provide for each instance in the data the correct label.

As an example, the term CD28 in the example sentence below was mapped in the feature vector $\langle B, activation, through, requires, reactive, oxygen, PROTEIN \rangle$, where last entry in the vector is the correct class/label of the term.

$\langle DNA \rangle IL-2 \langle /DNA \rangle$ gene expression and $\langle protein \rangle NF-kappa B \langle /protein \rangle$ activation through $\langle protein \rangle CD28 \langle /protein \rangle$ requires reactive oxygen production by $\langle protein \rangle 5-lipoxygenase \langle /protein \rangle$.

For terms that occur at the beginning or the end of sentences, e.g., *IL-2* in the example sentence above, the previous words or following words are missing. In such cases, we assigned for the corresponding features default values such as *beg1*, *beg2*, *beg3*, or *end1*, *end2*, *end3*, respectively.

The deliverable for Step 1 of Phase 2 was the vector representation of all the biomedical terms that occurred in a set of sentences. The set consisted of five sentences hand-picked by the instructor from the 300-sentence standard data set. For these five sentences the mapping had to be done manually by students. The role of this manual mapping task was to give students the opportunity to try themselves the mapping of biomedical term instances onto the vector representation. Also, the instructor had the

opportunity to check students' level of understanding of this important task and offer the necessary feedback.

Step 2: Feature Extraction

In this step, the whole data collection provided in Phase 1 needed to be mapped onto the features vector representation. Students were asked to write a computer program that for each biomedical term in the 300-sentence data set would generate the vector representation. The deliverable for this step was the entire data set mapped into features vector representation.

Step 3: Data Preparation

This last step of Phase 2 was about creating a data set in the format required by the Weka 3 machine learning toolkit [9]. Weka 3 Data Mining System is a free Machine Learning software package implemented in Java, which includes decision trees and naïve Bayes learning algorithms.

Weka requires the input be in the Attribute-Relation File Format (ARFF) format. An ARFF file is a text file which defines the attribute/feature types and lists all biomedical term feature vectors along with their class value (the biomedical term class). Details of the ARFF format can be found in [9]. We provide below an example of an attribute/feature definition corresponding to the third previous word feature of our modeling of the biomedical term classification problem and a feature vector for the biomedical term instance *CD28* that appeared in the example sentence shown earlier. The bold words are ARFF-required keywords.

```
...
@attribute PREVIOUS3 {beg3, oxygen, B, decreased, levels, ...}
...
@data
B, activation, through, requires, reactive, oxygen, PROTEIN
...
```

The major task in this step is to generate the ARFF file for all 300 sentences. Students had to write a program to generate the ARFF file. The ARFF file would serve as input to Weka in the machine learning phase, i.e., Phase 3, of the project.

There were two deliverables for Step 3 of Phase 2. First, the ARFF data file containing the feature vectors for all instances of biomedical terms in the 300-sentence data set was required. Second, a description of the ARFF data file was needed that would include: an explanation of the correspondence between the surrounding words and the attribute declaration part of the ARFF file (the lines beginning with **@attribute**) and an explanation of the data rows (the portion after **@data**). For example, students had to choose a tuple, which is the corresponding representation in the ARFF format of the vector representation of an instance, and explain what the feature values mean for the biomedical term that the tuple represented.

Phase 3 – Machine Learning

In this phase, machine learning algorithms were used to induce classifiers that could then be applied to new instances of biomedical terms in order to classify them. The classifiers were induced based on the biomedical term instances in the training set (200 sentences out of the 300 sentences in the provided data set) and then evaluated on new instances from the test set (the remaining 100 sentences). For both purposes we used the Weka 3 Data Mining System. The steps involved were: (1) loading the ARFF files created in phase 2 of the project, (2) using the Weka's decision tree algorithm (J48) students had to examine the decision tree generated from the training data set and identify the most important features (the ones appearing on the top of the tree), (3) repeating the previous steps using the Naïve Bayes algorithm, and (4) using the Weka test set option students were asked to classify the term instances in the 100-sentence test data set, analyze and report the obtained performance. In an additional step, students were asked to consider all 300 sentences as a single set and evaluate the induced classifiers using 10-fold cross-validation. In *k-fold cross-validation*, the available data is divided into k folds. One fold is then chosen as testing data and the remaining $(k-1)$ folds used as training data. This process is repeated k times for each fold and then the average of the performances obtained for each chosen fold is taken. When $k=10$ we have 10-fold cross-validation.

The deliverables of this phase were: an explanation of the decision tree learning algorithm (Weka's J48), a detailed report on the experiments performed including descriptions of the experiments (input data, Weka outputs), and suggestions of what students thought could have been done to improve on the classification.

COURSE IMPLEMENTATION DETAILS

The Biomedical Term Classification project was implemented in the *Introduction to Artificial Intelligence* course that was taught during Fall 2008 in the Department of Computer Science at The University of Memphis. Table 1 provides the Syllabus and where in the semester each phase of the project was added. Whenever a phase or major step of the project was added to the regular topics, a 30-minute time slot was reserved to cover the details of that phase or step. While there were some challenges with integrating the various phases of the project throughout the semester, e.g. adding a phase in a week in which the regular topic was somehow distant to machine learning, we were able most of the time to find some overlap between the regular topic and a project phase. For instance, Step 2 of Phase 2 was about mapping the biomedical term occurrences onto the features vector representation. Because the vector representation is a form of knowledge representation it fit well with the topic of Week 8, which is *Knowledge Representation* (see Table 1). Another driving force behind the proposed schedule was the goal to have the data sets in the ARFF format, i.e., ready to be loaded in Weka, before Weeks 11 and 12, the *Machine Learning* weeks (see Table 1). This way students could focus on experimenting with the decision trees and naïve Bayes algorithms during these weeks. We have learned that the schedule shown in Table 1 does not cause major problems in student learning experience and was also relatively easy to implement for the instructor. One challenge with implementing the *Biomedical Term Classification* project in the AI course was the processing of natural language texts in LISP (the main programming language used in the course). LISP is not the best language for that purpose (e.g., to

locate certain strings such as biomedical terms in input sentences) and in consequence students were allowed to use any language they preferred to process the texts in order to accomplish the various tasks needed in the project, e.g. to generate the features vector representation from the original annotated textual sentences.

Table 1. Class Syllabus with the phases of the project highlighted.

Week	Topic
Week 1	<i>Introduction to AI [Introduction to the Project]</i>
Week 2	<i>Problem Solving by Search</i>
Week 3	<i>Informed/Heuristic Search [Phase 1]</i>
Week 4	<i>Games Playing</i>
Week 5	<i>Knowledge-based Agents [Phase 2, Step 1]</i>
Week 6	<i>First-order Logic</i>
Week 7	<i>First-order Logic</i>
Week 8	<i>Knowledge Representation [Phase 2, Step 2&3]</i>
Week 9	<i>Knowledge Representation</i>
Week 10	<i>Uncertainty and Probabilistic Reasoning</i>
Week 11	<i>Machine Learning [Phase 3]</i>
Week 12	<i>Machine Learning Logic</i>
Week 13	<i>Natural Language Processing [Survey]</i>
Week 14	<i>Natural Language Processing</i>
Week 15	<i>Review</i>
Week 16	<i>FINAL EXAM</i>

Assessment

In Week 14, we gave students a survey in which we asked about various aspects of their learning experience in the class as reflected by the addition of the machine learning project. For each questions, students had to provide an answer on a scale of 1-5, with 1 meaning *strongly disagree* and 5 *strongly agree*. We received 100% participation (7 students) in the survey. All subjects were male with 72% of the participants being seniors and 28% being graduate students (the course was COMP4/6720 which is open both to undergraduate and graduate students).

From the survey we learned that students found the project interesting to work on (average score of 4.28), it contributed to their overall understanding of the material in the course (4.42), the project was an effective way to introduce machine learning concepts (4.28), and it made them interested in learning more about machine learning (4.28) and AI (4.28). Overall students had a good learning experience in the class (4.71). The least average score (3.71 corresponding to *Somehow Agree*), and – importantly - the only score below 4.00 (all questions were positively formulated), was with respect to the difficulty level of the project compared with student background and programming knowledge (3.71). On the other hand, students considered the project took a reasonable amount of time to complete (4.28).

CONCLUSIONS

The *Biomedical Term Classification* project that we introduced in the *Introduction to Artificial Intelligence* course had a very positive impact on student understanding of core concepts in Artificial Intelligence and machine learning. The plan for the future is to give students a choice of machine learning projects from the large MLeXAI project and group them in teams to work on these projects. The team will distribute the effort necessary to develop the projects as less time will be available to the instructor to address each individual project in detail while still covering the regular topics needed to be covered in an *Introduction to Artificial Intelligence* course. The next phase of MLeXAI, which *Biomedical Term Classification* is a part of, will further investigate the impact of theme and problem based learning on student outcomes and examine whether results of this earlier phase can be extended and generalized by 20 faculty across the country.

ACKNOWLEDGEMENTS

This work is supported in part by the National Science Foundation under Grant Numbers DUE-0409497, and DUE-0716338

REFERENCES

- [1] Cohen, B.K., Hunter, L. (2004). Natural Language Processing and Systems Biology, In Dubitzky and Pereira, *Artificial intelligence methods and tools for systems biology*, Springer Verlag, 2004.
- [2] Dodds, Z. et al. (eds). (2006). *Robots and Robotics in Undergraduate AI Education*, Special Issue of AI Magazin, 27 (1), AAAI Press, 2006
- [3] Kumar, A., Kumar, D., Russell, I. (2006). Non-traditional Projects in the Undergraduate AI Course, Proceedings of the 37th Annual SGICSE Technical Symposium on Computer Science Education, ACM Press, March 2006.
- [4] Kumar, D., and Meeden, L. (1998). A Robot Laboratory for Teaching Artificial Intelligence, Proceedings of SIGCSE, ACM Press, New York, NY, 1008, pp. 341-344.
- [5] Mitchell, T.M. (1997). *Machine Learning*, McGraw Hill, New York, 1997.
- [6] Russell, S. and Norvig, P. (2002). *Artificial Intelligence: A Modern Approach*, 2nd edition, Prentice Hall, 2002.
- [7] Russell, I. Markov, Z., and Coleman, S. (2007). Project MLeXAI: Applying machine learning to web documebnt classification. J. of Computing Sciences in Colleges 23(2).
- [8] Hatzivassiloglu, V., Doboue, P., Rzhetsky, A. [2001]. Disambiguating proteins, genes, and RNA in text: a machine learning approach, *Bioinformatics*, Vol. 17, pages S97-S106, 2001
- [9] Witten, I.H. and Frank, E. (2000). *Data Mining: Practical Machine Learning Tools and Techniques (Second Edition)*, Morgan Kaufmann, 2005, ISBN: 0-12-088407-0.