Proposal for Harvest Project: La Vie (Learning Adapted Video Information Enhancer)

Rayid Ghani Marko Grobelnik Colin de la Higuera Mitja Jermol
Alfons Juan Matjaž Rihtar John Shawe-Taylor

January, 2012

Project general information

Title: Learning Adapted Video Information Enhancer (La Vie)

Coordinator: Matjaž Rihtar (IJS & K4All)

Management team: Rayid Ghani (K4All), Marko Grobelnik (K4All), Colin de la Higuera (K4All), Mitja Jermol (K4All), Alfons Juan (Valencia) and John Shawe-Taylor (K4All)

Hosting Organisation: Knowledge 4 All Foundation Ltd

Hosting Site: Institut Jožef Stefan

Expected Dates: June – September, 2012 for a period of 12 weeks, possibly interrupted for a holiday period.

Problem Description

PASCAL has been very successful in supporting and promoting the development of the videolectures site (www.videolectures.net). Knowledge 4 All Foundation has been established to carry forward the development of this activity as well as other legacies of the PASCAL Network.

One problem created by the success of videolectures is the difficulty that individual users have in identifying the best video for their needs among the vast range of possibilities afforded by the site. Each video has a particular mix of content and style of presentation with implicit assumptions about background knowledge and level of expertise of its intended audience. On the other hand the video consumer has an approximate understanding of his/her abilities and material that he/she would like to learn about. For example, they may have a background in basic classification methods (eg SVMs) applied to text, some knowledge of probability theory, but not know about Bayesian reasoning. He/she would like to learn about Topic Models. The question of which sequence of videos would be most appropriate to help him/her to attain the desired knowledge would also depend on the style of presentation he/she prefers and so on.

Currently Videolectures provides a contextualisation service that is only based on keywords extracted from the lecture titles. It provides a recommendation of videos that are related to a given video. The relation would typically be based on the topic of the video or the lecturer. Furthermore, the system does not adapt its responses to the interests or background of the user.

Videolectures has begun to collect information about individual users, though it is currently limited to associations with the keywords of lectures that they have viewed. This information is currently only being used for off-line analytics.

Proposed Harvest Project: The La Vie project would develop a proof-of-concept system that would provide users with advice on suitable videos for their needs. The key additional components that the project will bring to videolectures are:

1. topic extraction and modeling based on text extracted from associated slides and audio transcriptions. This will ensure that the devised user models can capture semantic level interests of the users.

2. inclusion of the information currently being logged about individual users in the recommender system running live on the videolectures site.

3. visualisation of the developed recommender system. This will provide a topic landscape that will enable users to see the available videos emphasising those likely to be of interest to the user.

The La Vie system will be broken down into five phases. First the text extraction modules will use text mining methods to extract information from the content and meta-data associated with a particular video. The second phase will involve topic modeling from the extracted text in order to develop a richer semantic representation. Phase three will use the developed topic models and other relevant information (such as formalised ontologies, taxonomies, structured information sources such as wikipedia, linked open data (LOD), and other contextual information) to populate a semantic representation of individual videos and users. The fourth phase information retrieval module will provide relevant recommendations by linking the enriched semantic representations of users and videos. The final phase collects feedback from users about the operation and effectiveness of the recommendations that have been given. This information will be used to update the user and other models. All of these components will be integrated into videolectures through the user interface that will provide visualisation and interactivity. More detailed descriptions of these components are given below.

The project will provide a framework into which new modules can be plugged either to replace old ones or to enhance functionality. For example, a simple k-means algorithm would be used initially for the topic modeling, while clearly more advanced methods should provide improved performance and may be included in the system during the first phase of the project if time permits. As an example of enhancing functionality the representation of user interests as a distribution over meta-data attributes would make it straightforward to include new modules refining these distributions through interactions with a user applying say a bandit algorithm inspired by content-based image retrieval.

This framework approach to the development of the La Vie project ensures that the risks associated with the project are kept to a minimum. Once the framework is in place and the basic system operational, further incremental developments can be included without risking the overall success of the project results. Several components already exist including Qminer for user modeling, Enrycher for text enrichment, DocumentAtlas for visualisation, etc., that will make this approach likely to deliver a basic working system within approximately six weeks given the manpower envisaged below. Here we have taken into account the time required for the team to accumulate the necessary expertise in the different existing modules as well as the significant effort required to design and program the new framework and modules. In the second part of the project specific advanced modules will be developed chosen based on the skills of the participants and potential for enhancing the overall performance of the system. These modules could include probabilistic topic models, bandit style algorithms for optimising user interaction, improved visualisation algorithms, etc.

Visibility afforded to PASCAL

We believe that the system will combine two goals in promoting PASCAL: firstly improving access to videolectures and secondly demonstrating the use of machine learning in a system designed to enhance a user experience. It will also ensure that the system is compatible with the outputs of the Translectures project, a project that is also helping to secure the PASCAL legacy of videolectures through automatic (cross-lingual) subtitling. It is also anticipated that La Vie would act as a dissemination vehicle for Translectures.

Outline of the system

The system will be architected in a way that ensures its forward compatibility as new components that improve the implementation of different parts of the system are develop or become available. The key design elements must therefore be quite generic and the design team have attempted to second guess a number of different ways in which components may be enhanced, eg cross-lingual access, improved topic models, novel evaluation methodologies, etc. The architecture of the system is built around a pipeline involving five main stages:

1. text extraction modules (extracting text from audio and associated textual documents and slides);

   Input: content + metadata

   Output: text + metadata

2. topic modelling module (building topic modules from the extracted text);

   Input: text

   Output: content × topic matrix, topic × word matrix (cutoff by some threshold), topic title word candidates

3. Enrichment modules (providing enriched representations for both content and users);

   - Content Enrichment Modules: level, scope, quality, didactic, application area, topics and categories

     Input: text, topic matrix

     Output: scalar (level) or vector (score for each category)

   - User enrichment modules:

     Input: userid, topics

     Output: distribution over content metadata, (fixed) user attributes (age, gender, etc.)

4. Information retrieval module (identifying content for a user based on the output of the enrichment modules);

   Input: based on the UI (userid, query, visual feedback, etc.) Intermediate representation: query words + distribution over meta-data attributes

   Output: ranked list of content, meta-data weights according to their importance in retrieval

5. evaluation modules (assessing performance of the system based on user feedback).

   Input: Interactions with users

   Output: Logging of a series of indicators

Underpinning the application will be a user interface including visualistation modules integrated into videolectures.

- User Interface: includes visualization module and interactive module
  - Visualisation module:

    input: ranked list + meta-data importance
    Output: through videolectures
  - Interactive module:

    Input: from the user
    Output: parameters for the context retrieval module or direct to content

Method that will be used to assess performance

The Evaluation module will be developed to enable logging of a range of attributes that could be indicative of usefulness of the system. These might include time spent interacting with the system, number of videos accessed, time spent viewing videos accessed, time to next visit to the site, etc. During the six months following the deployment of the system, this data will be logged and where appropriate compared with previous behaviour of users. By performing a simple principal components analysis and including some data from a small user study, we will identify which indicators are correlated with a positive user experience. In this way we hope to avoid misinterpreting specific attributes such as for example time using the system: this could mean user satisfaction in that it is providing useful information, or dissatisfaction that it is taking so long to locate the right video.

Training and Outreach

The project will be using state-of-the-art methods from text processing, topic modelling, user modelling, visualisation and language modelling for transcription. We will ensure that the project involves regular talks (either in person or from videolectures.net) that provide training in these topics to the team and wider audience. There will be opportunities to discuss the issues involved and understand all aspects of the design. In addition, the project is planning to organise a set of lectures aimed at disseminating the resulting system to targeted user groups such as OCWC (Open Courseware Consortium) and OCC (Opencast Community), both closely associated with Knowledge 4 All Foundation.

Management and effort

The team that will manage the project comprises the following members:

Rayid Ghani was until recently a Senior Researcher and Director of Analytics Research at Accenture Technology Labs, where he led research in the areas of Machine Learning, Data Mining and Text Mining. The Labs' goals include researching and inventing the next wave of business solutions using new and emerging technologies and exploring how these will evolve, converge and shape businesses in the future. Rayid is a member of IEEE and ACM and has published widely in Machine Learning, Data Mining, and Knowledge Management journals and conferences. He has also organized several machine learning and data mining workshops and has been on Program Committees for a variety of major conferences. He is currently leading the development of software systems to support the campaign to reelect Barack Obama.

Marko Grobelnik is researcher and manager of research group of 15 people at the department of Knowledge Department working primarily in the areas of text-mining and social network analysis. He is coauthor of several books and numerous scientific papers. Marko is a technical director of FP6 IST World project on analysis of European research, a member of management board of several FP6 & FP7 projects (FP6 SEKT-IP, FP6 NEONIP, FP7 ACTIVE-IP) and participates in W3C standardizing committees. He co-organized over 10 international workshops and tutorials on text mining and link analysis at prominent conferences like IJCAI, ACM-KDD, IEEE-ICDM. Marko also closely collaborates on research projects with Microsoft Research, Cycorp Europe, Carnegie Mellon University, Cornel University.

Colin de la Higuera is a director of K4All. He is Professor at Nantes University (France), a member of the PASCAL2 Network and is involved in a number of research themes, including algorithmics, formal language theory, pattern recognition and machine learning. He would contribute to the project through the identification of bottlenecks, the proposal of evaluation measures and protocols.

Mitja Jermol is head of the Centre for knowledge Transfer at IJS working in the area of e-learning and dissemination and promotion of research results. His main research area is Knowledge Management enriched with the modern analytic techniques in the context of improved business processes for (networked) organizations. Among others, he is IJSs representative in the leading FP6 ECOLEAD-IP and FP7 COIN-IP in the area of networked organizations and virtual enterprises. Before joining the Institute, Mitja was heading the research group for distance education and e-learning at Slovenian major publishing house.

Alfons Juan is the PASCAL2 UPV site manager and coordinator of Translectures, an EC research project. His main research areas of interest are pattern recognition and machine learning, and their application to speech recognition, machine translation and image processing. His contribution would be in helping to specify the system architecture and ensuring its potential as a showcase for PASCAL.

Matjaž Rihtar has extensive experience with project management working at Hermes Softlab in Ljubljana. He has experience in leading large software development projects in the area of massive data storage, business analytics and business processes. He is currently employed at the Jožef Stefan Institute as a Technical Officer and has worked with many of the components used in software behind the videolectures site. He is employed on the Translectures project that will be linked with La Vie specifically as a source of additional information about videos through audio transcription and subtitling.

John Shawe-Taylor is a director of K4All and the Scientific Coordinator of the PASCAL2 Network. His main areas of research are theoretically motivated approaches to machine learning with an emphasis on statistical learning theory and kernel methods. His contribution would be in helping to specify the system architecture and ensuring its potential as a showcase for both PASCAL and K4All.

Complementarity of competencies:

	text processing	topic modelling	user modelling	visualisation	language modelling	software dev. & int.
Rayid Ghani	X	X	X			X
Marko Grobelnik	X	X	X	X	X	X
Colin de la Higuera	X			X	X	
Mitja Jermol		X	X		X	
Alfons Juan	X	X		X	X	
Matjaž Rihtar		X	X		X	X
John Shawe-Taylor	X		X	X	X	

Marko, Mitja and John are responsible for Knowledge 4 All Foundation, videolectures operation and are therefore also representing the user side.

Management of IPR: The IPR arising from the project will be assigned to Knowledge 4 All Foundation as the PASCAL2 legacy organisation. However, the software will be open source which complies with the strategy of Knowledge 4 All and its charitable objectives.

Risk Management: All of the management team are committed to the project and we have identified local participants. We therefore believe that the risks associated with recruitment of the team are minimal. While this is at heart an implementation of existing technologies, there are some risks associated with their ability to deliver value in this application. We do not anticipate that this will be a significant impact, but to the extent that this does occur the project will initiate interesting open research questions. Hence, should the application prove problematic, the project can potentially create a challenge corpus for developments of topic modelling for real world applications.

Effort breakdown: We have assessed that the effort required is approximately 7 persons working over a 3 month period (three of these people will be local participants from Institute Jozef Stefan and University of Ljubljana and hence incur no costs). We have broken this down into work on the various components of the system:

- Management (1 person)
- Developers/partners
  - Overall architecture: Videolectures developer (1 person)
  - Modules: (2 persons)
    - Extraction from pdf / input from transLectures (transcription) & text enrichment
    - Topic modelling level, scope, quality, didactic, applications, topics and categories, use modes, user contextdetection modules
    - Context retrieval module, pointers to external topics
  - User Interfaces: (2 persons)
    - Visualisation module (basic version) + extended with existing visualisation (JSI) techniques
    - Interactive input module
  - Evaluation (1 person)
    - Implementation of data collection methods
    - Analysis and interpretation after 6 months

Provisional names of personnel: The management team listed above would be involved for the initial week and available for video conferencing on a weekly basis. They would cover their costs from other sources though some costs might be charged funds permitting. Provisional names of those more actively involved in the project include:

Matjaž Rihtar will be the lead manager of the project and oversee the execution of the different developments. See above for more information about Matjaž.

Blaž Fortuna is a senior research assistant and a PhD student at JSI in the area of kernel methods, statistical learning and semantic Web with strong focus on text analysis. In the recent years he had several publications at international conferences and developed several software modules for scalable machine learning, cross-lingual information retrieval and classification, ontology learning and active learning which are part of Text Garden software environment. He is also major author to the OntoGen (http://ontogen.ijs.si) system for ontology learning and Document Atlas (http://docatlas.ijs.si) text visualization software. Bla was at internships in Microsoft Research and in Bloomberg. He is a co-inventor on a patent on root cause analysis developed within several projects with British Telecom and New York Times. In tems of project experience, Blaz was working on FP6 machine translation project SMART-STREP, FP6 semantic technology integrated project SEKT. Recently he is one of the major contributors for FP7 Integrated Project LarKC (Large Knowledge Collider) working on integrating Cyc system and knowledge base in the core project platform. Bla is co-organizer of Workshop on Inductive Reasoning and Machine Learning for the Semantic Web at ESWC 2011 and contributed to text mining and machine learning tutorials at ACM KDD and ISWC conferences.

Mitja Trampuž is a PhD student at the Faculty of Computer Science and Informatics, University of Ljubljana, working at the Artificial Intelligence Laboratory at Jožef Stefan Institute. Mitja's research field is data mining with the particular interest in text mining and structured data mining (graphs). He have also worked on visualizations of complex data in association with text mining. Currently, he is involved in research on semantic parsing and extraction of events from text in the two projects VIDI (visalization of discussion forums) and RENDER (mining for diversity in text). Mitja has been involved in many industry applications with IBM Slovenia, Facebook and Cosylab Slovenia.

Tadej Štajner is a research assistant at JSI in the area of applying machine learning and semantic knowledge management methods to natural language processing tasks involving many types of sources, ranging from enterprise document repositories, news outlets and social media. Recently, he has published several papers on named entity disambiguation, enterprise knowledge process analysis and prediction and is currently focusing on extending cross-lingual integration of text and structured knowledge. He has participated on several FP7 projects, being a significant contributor to ACTIVE-IP, RENDER-STREP, MultilingualWeb-CA and XLIKE-STREP.

Janez Brank is a final year PhD student who has submitted his thesis for assessment. He has worked in text mining, ontology learning and related topics that are directly relevant to the La Vie project.

William Martin is currently a final year undergraduate student at UCL. He has been accepted onto a PhD in the medical department at Royal Free Hospital in London. He is undertaking a final year project in probabilistic topic modeling under the supervision of John Shawe-Taylor and has expressed interest in participating in the La Vie project.

Milestones

Milestones are specified in work days – the total number of working days is 60.

D5:	Specification of main module datatypes and interfaces with individual module functionality at a high level.
D10:	Specification of first phase modules in more detail with empty implementations plus implementation of overall framework
D15:	Samples of data inputs to enable module developers to test their modules
D25:	Initial implementations of modules
D30:	First test of complete system, review of modules and identification of weaknesses in the design and modules that should be improved/added
D50:	Final version of the demonstrator assembled
D55:	Completion of correctness testing and bug fixing
D60:	Completion of initial on-line testing and correction of any bugs identified
9 months:	Completion of analysis of user access data and user trials

Requested funding

The infrastructure for the project will be provided by videolectures and IJS. Separate negotiations are underway to ensure continuity of this infrastructure beyond the end of the PASCAL Network. For this reason the main request is for travel and per diems. As we are not sure who the implementors will be these amounts are estimates. We estimate that there will be four non-local participants (apart from the managers – note that in the original proposal we estimated seven non-local participants, but we have not identified three local participants). The per diems are calculated as 4×60 days for the four non-local participants.

	Qty	Unit cost	Cost
per diems	240	90	21,600
travel expenses	7	200	1,400
Total			23,000