Cross-Media Knowledge Extraction in the Car Manufacturing Industry

José Iria
The University of Sheffield
211 Portobello Street
Sheffield, S1 4DP, UK
j.iria@sheffield.ac.uk

Spiros Nikolopoulos
ITI-CERTH
6th Klm. Charilaou Thermi Rd
P.O. BOX 60361 GR - 57001
Thessaloniki, Greece
nikolopo@iti.gr

Martin Možina
University of Ljubljana
Tržaska 25
1000 Ljubljana, Slovenia
martin.mozina@fri.uni-lj.si

Abstract

In this paper, we present a novel framework for machine learning-based cross-media knowledge extraction. The framework is specifically designed to handle documents composed of three types of media – text, images and raw data – and to exploit the evidence for an extracted fact from across the media. We validate the framework by applying it in the design and development of cross-media extraction systems in the context of two real-world use cases in the car manufacturing industry. Moreover, we show that in these use cases the cross-media approach effectively improves system extraction accuracy.

1 Introduction

In large organizations the resources needed to solve challenging problems are typically dispersed over systems within and beyond the organization, and also in different media. For example, to diagnose the cause of failure of a component, engineers may need to gather images of similar components, the reports that summarize past solutions, raw numerical data obtained from experiments on the materials, and so on. The effort required to gather, analyze and share this information is considerable, consisting of up to several dozen man-months for the most complex cases.

In the EC-funded project X-Media[1], we are working together with our industrial partners, the jet engine manufacturer Rolls-Royce plc. and the car manufacturer Fiat S.p.A (FIAT), on the automatic capture of semantic metadata as an enabling step towards effective knowledge sharing and reuse solutions. This type of technology is already available for single-medium scenarios: named entity recognition and information extraction for text, scene analysis and object recognition for images, and pattern detection and time series methods for raw data. However, there is still the need for effective knowledge extraction methods that are able to combine evidence for an extracted fact from across different media. Cross-media analysis is motivated by the fact that information carried by different communication channels is important for humans to fully comprehend the intended meaning. In many cases, as shown in this paper, the whole is greater than the sum of its parts: considering the different media simultaneously can significantly improve the accuracy of derived facts, some of which are otherwise inaccessible to the knowledge worker via traditional methods which work on each single medium separately.

To address this need, we have designed a novel cross-media knowledge framework able to accomodate different approaches and systems, to fullfil the requirements of the real-world use cases provided by our industrial partners. Based on the framework, we have implemented innovative knowledge extraction systems, capable of extracting information from multimedia documents containing text, images and raw numerical data, and empirically evaluated them. We show that cross-media analysis does improve system accuracy by simultaneously exploiting the evidence extracted across media.

The rest of the paper is structured as follows. In the next section we describe our proposed cross-media knowledge extraction framework. Next, we present two use cases in the car manufacturing industry where the need for cross-media analysis was identified, and solutions based on the framework were developed. Furthermore, an evaluation of the systems is presented, which quantifies the improvement in accuracy obtained by adopting a cross-media approach. The paper ends with conclusions and future work.

[1] http://www.x-media-project.org

2 Cross-Media Knowledge Extraction Framework

The requirements for the framework were drawn from the X-Media industrial use cases, two of which are presented in detail in Section 3. The major requirement identified was the ability to exploit evidence for a fact from across several media. Other requirements, which also had important implications in the design decisions, include the ability to exploit existing (background) knowledge, portability, the ability to report uncertainty, and the ability to perform the extraction on a large scale. In this paper we focus mainly on the cross-media requirement.

The framework, depicted in Figure 1, accepts multimedia documents as input, and outputs semantic annotations about the extracted concepts and relations. It consists of the following steps:

Pre-processing. The document processing literature discusses approaches to process PDF, HTML and other structured multimedia document formats, see [5] for an overview. The goal is to extract single-medium features and cross-media features from documents to build a representation of the data suitable for learning predictive models.

Cross-Media Feature Extraction. As mentioned, a document may contain evidence for a fact to be extracted across different media. However, it is not straightforward to know which media elements refer to the same fact. The document layout and cross-references (e.g. captions) can hint at how each text element relates to each image/raw data table [2]. The cross-media features we extract depend on the particular use case application, but generally include layout structure, distance between media elements, and cross-references.

Concept Model Learning. Once the data representation integrating all the features is ready, standard learning algorithms can be used to estimate the model of a single concept exclusively by using the concept's own examples.

Background Knowledge. Semantic metadata provide information about concepts co-occurrence and how they cooccur across different media [6], the "semantic structure" of the problem. The framework is able to exploit this type of background knowledge, to enhance the model of each individual concept and improve systems' accuracy.

3 Experimental Study

We have evaluated the proposed framework by designing, implementing and validating cross-media knowledge extraction solutions to the real-word use cases provided by our industrial partners. Two of such use cases, defined in cooperation with Centro Ricerche Fiat (CRF), the research division of FIAT, are presented in this paper. We show that the two knowledge extraction systems successfuly extend the proposed framework and that, by virtue of cross-media analysis, improve accuracy with respect to single-medium approaches.

3.1 Competitors Scenario Forecast

This use case concerns forecasting the launch of competitor car models. It comprises collecting information about the features of competitors' vehicles from various data sources and producing a calendar that illustrates the prospective launches. The collected information is used in the *set-up* stage of new FIAT vehicles (the development stage where a first assessment of the future vehicle features is carried out), and is thus of great value to the company.

In traditional competitors scenario forecast, the main role is played by someone responsible for data acquisition. Her role is to inspect a number of resources daily, such as WWW pages, car exhibitions, car magazines, etc, that publish material of potential interest. The focus of our analysis was to evaluate these documents with respect to their relevance to car components' ergonomic design. This task was selected as a pilot for demonstrating the feasibility of the proposed solution in performing cross media analysis.

3.1.1 Approach

The proposed solution can be viewed as a cross media high level concept detector. The solution adopted for this use case is based on a generative model implemented using a Bayesian Network (BN). Its full description can be found in [7]. In the following we describe in what way the solution is an instantiation of the cross media framework presented in Section 2.

With respect to multimedia document processing we employ a mechanism that dismantles a document into its visual and textual constituent parts. Concerning layout features we adopt a rather straightforward approach where all media elements of the same document page are considered to be conceptually related. Thus, layout information is incorporated only in terms of the co-occurrence relations between the media elements of the same page. For constructing the concept data models we process the features extracted from each single-medium. For image content, we employ the Viola and Jones detection framework [8] that use Haar-like features to represent visual content. For text content, we employ 18 custom analyzers that extract textual descriptions from each page based on regular expressions and a look-up table of synonyms, hyponyms and hypernyms. More details concerning the single-medium extractors can be found in [7].

According to the cross media framework of Fig. 1, we may incorporate in our solution background knowledge and cross media dependencies models. In order to do so we

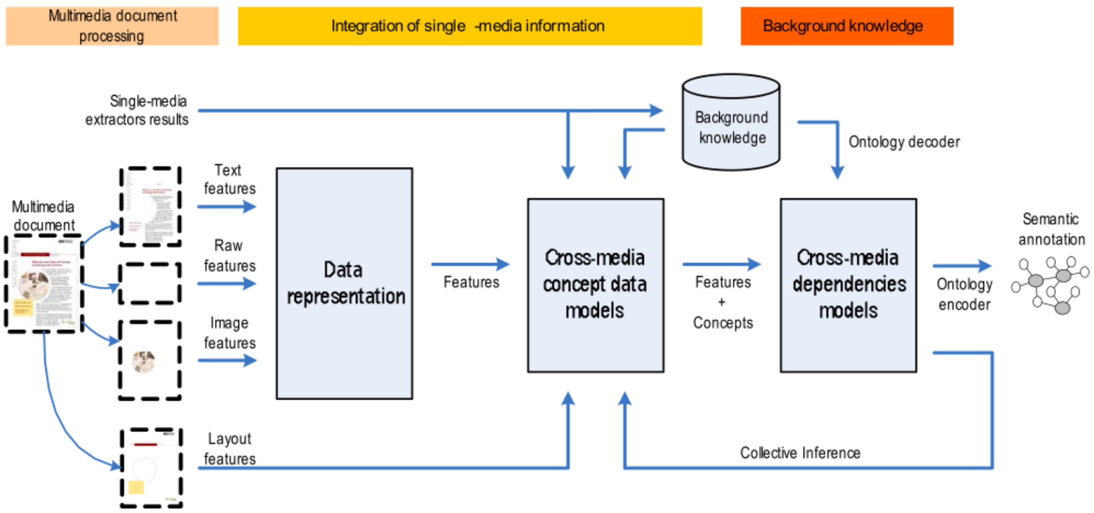

Figure 1. The proposed cross-media extraction framework.

use the methodology presented in [3]. By following this methodology we integrate ontologies and conditional probabilities into a Bayesian Network (BN) that is able to perform probabilistic inference. The ontologies are used to express the background knowledge, which in this case is the type of knowledge stating whether two domain concepts are related and what type of relation associates these concepts. The conditional probabilities, on the other hand, are useful in quantifying the dependencies between concepts by approximating their strength using frequency information implicit in the training samples. In this case the training samples are concept labels that are used to train the BN using the Expectation Maximization algorithm, see [7] for details. Thus, our solution implements the cross-media framework of Fig. 1 by using a BN that derives its structure from an ontology, learns the dependencies between concepts using a set of training samples and performs probabilistic inference by fusing the output of concept data models applied on single-medium information.

3.1.2 Experiments

For evaluating the effectiveness of our approach, we have used a dataset of 54 annotated pdf documents (altogether 200 pages). This dataset was provided by CRF and consists of advertising brochures describing the characteristics of new car models. The analysis process involves applying all aforementioned textual and visual analyzers to the constituent parts of a document page and, according to their output, update the value of the corresponding BN nodes. Then an inference process is triggered in the BN using message passing belief propagation. Eventually, the posterior probability of the BN root node is compared against against an empirical threshold that determines the decision made by our system.

The implemented solution is capable of producing an output independently of the amount and origin of evidence injected into the network. When no evidence is injected, the confidence degree of the fact that the analyzed page is concerned with car components ergonomic design is equal to the frequency of appearance of such pages in the training set. As evidence is injected into the network, this degree changes according to the dependencies that have been learned from the BN. This property allows us to evaluate the performance of the cross-media classifier using evidence extracted only from text, only from images, or both.

The threshold value was uniformly scaled between [0,1] for drawing the evaluation curves in Fig. 2. Out of the 200 annotated pages, 150 were used for training the BN while the remaining 50 were used for testing. From Fig. 2 one can verify that the configuration using cross-media evidence outperforms the cases where evidence originates exclusively from one media type.

3.2 Vehicle Noise Analysis

The goal of the Vehicle Noise Analysis use case, also defined in cooperation with CRF, is to help analyse wind noise in a vehicle and provide solutions to reduce it. A particular task is to identify the source of noise, i.e., which car component is generating the noise. The data was gathered through several tests of competitors' vehicles in a wind tunnel. A report compiled by experts describes the results of a single testing session, which consists of a set of tests, each test on a different vehicle configuration. A configuration is a set of vehicle components relevant to noise reduction like:

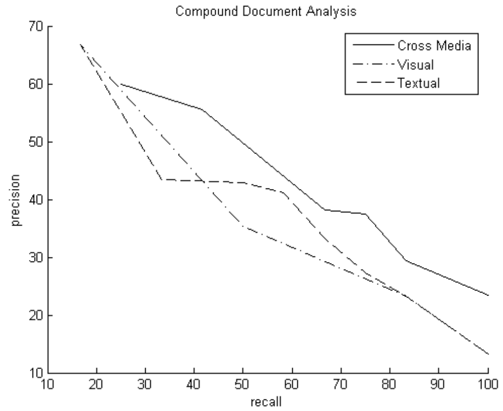

Figure 2. Comparison between the accuracy obtained by the cross-media approach against the single-medium approaches in the Competitors Scenario Forecast use case.

mirrors, antenna, windscreen wipers, etc.

The textual parts of a document (expert's opinions and table captions) contain relevant information that may, if used together with raw data, improve the prediction accuracy of learned models. The captions contain textual information necessary to associate a specific experiment with the concepts (possible configurations) present in a domain ontology. The information found in expert's opinions can be used to spot the main findings in the experiments. A more detailed description of the use case can be found in [4].

3.2.1 Approach

We aim to predict the complete audio spectrum of wind noise (vector of 110 sound pressure values) for a given vehicle configuration. The learning data is a combination of text and raw-data features. Our solution is an instantiation of the framework presented in Fig. 2. We implemented a document parser extracting text and raw data features from the reports, and learn a cross-media data model from raw data and classified text descriptions. Further, we use some of the text features as additional input features (after classification) along with the raw data features, while the rest of the text features were used as background knowledge.

Each cross-media learning example contains a set of following attributes: (a) a vehicle shape wind noise spectrum, where all component-related noise is removed by fully taping critical parts of a vehicle (the "optimal" configuration), (b) an original vehicle noise spectrum (noise measured in the original vehicle), (c) configuration description (in this experiment, a configuration is specified by a single component, e.g., front door cut), (d) the resulting vehicle wind noise spectrum with the selected configuration, and (e) experts' comments on the obtained spectrum for a given configuration.

The "raw" type of data are the three noise spectra (a,b,d) parsed from tables and graphs in the reports. The parsing of raw data features was done entirely by regular expressions.

A configuration (c) is specified by the name of the component being tested (one of approximately 30 different components). The configuration is obtained by performing text categorization with a *k*-nearest neighbor algorithm on the graph captions and associating the spectra with the corresponding captions (and with the experiment performed). The list of all possible components was extracted from the domain ontology. The experts' comments (e) written in natural language contain the salient aspects of a car test, in particular, the car components that have been tested and their influence on the car noise. The influences are characterized as either *critical* or *non-critical*, which was estimated from text using a strategy similar to sentiment analysis.

The task in this setting is to estimate the level of noise change for each component proposed by an expert, in other words, predict the attribute denoted by letter (d). Such a model, which could estimate noise produced by a component prior to testing, would significantly decrease the required time needed to test new vehicles, by recommending which of the available components should be tested first.

As mentioned, the criticality attribute contains comments of a particular experiment result. As the goal is to predict the result for a new (not yet tested) case, the criticality attribute will not be provided yet, and therefore should not be used as a feature in learning. However, we can use the values of this feature as background knowledge in the process of learning. We applied a principle similar to the one used in QFilter [9], where quantitative predictions need to be consistent with a provided qualitative model. In our case, the prediction of the model needs to be consistent with experts' comments: if component A was marked as critical for a specific vehicle and component B as not critical, then the model should assign a higher influence to component A. This constraint was used in the algorithm for optimizing the weights of the distance function.

3.2.2 Experiments

Initially, we used only the raw data features (a and b) for predicting noise influence (d). The root mean squared error (RMSE), a measure that quantifies the prediction error, of our kNN predictor was 10.73. Afterwards, we added the name of the component (c) extracted from text to the feature set, and the RMSE decreased significantly to 9.49. It should be noted that the extracted attribute value does not always correspond to the true component name, as the accuracy of text extraction tool was 85%. As mentioned, the other text feature (criticality) could not be used as a feature in learning, because it is not available for the new vehicles, and was rather used as background knowledge. The RMSE of the method slightly improved to 9.42, however, it should be noted that we used the true values of criticality attribute (manually extracted), as the performance of the text extraction tool turned out to be correct only in 51% of the cases.

4 Conclusions

The technology focus in Knowledge Management has moved from simple keyword-based search towards more advanced solutions for extraction and sharing of knowledge [1]. The focus is still very much on providing more advanced text-based solutions, though image and video are considered by some industry players. Recently many projects dealing with knowledge extraction and sharing have been funded at European level. Most address the problem of knowledge extraction over a single medium, but some do address extraction over multimedia data, e.g., Reveal-This[2], MUSCLE[3], and MESH[4]. However, most of the research is themed around video retrieval applications, which typically consider video, caption and speech analysis, differing quite substantially from X-Media's need to analyze and mine documents comprising text, static images and raw data. In fact, X-Media's knowledge-rich real-world environments such as those presented in Section 3 set it apart from other projects in the area.

The contributions of this paper are threefold: (i) the design of a novel machine learning-based cross-media knowledge extraction framework; (ii) the validation of the suitability of the framework to accomodate knowledge extraction systems operating in diverse use cases; (iii) the empirical demonstration that cross-media analysis successfully improves systems' accuracy with respect to the single-medium approaches in real-world technical domains. Future work concerns the further improvement of the accuracy and scalability characteristics of both our single-medium and cross-media methods.

5 Additional authors

Alberto Lavelli, FBK, lavelli@fbk.eu
Claudio Giuliano, FBK, giuliano@fbk.eu
Lorenza Romano, FBK, romano@fbk.eu
Damjan Kužnar, University of Ljubljana, damjan.kuznar@fri.uni-lj.si
Ioannis Kompatsiaris, ITI-CERTH, ikomp@iti.gr

6 Acknowledgments

This work was funded by the X-Media project (www.x-media-project.org) sponsored by the European Commission as part of the Information Society Technologies (IST) programme under EC grant number IST-FP6-026978.

References

[1] W. Andrews and R. E. Knox. Magic quadrant for information access technology. Technical report, Gartner Research (G00131678), October 2005.

[2] A. Arasu and A. H. Garcia-Molina. Extracting structured data from web pages. In *ACM SIGMOD International Conference on Management of Data*, San Diego, California, USA, 2003.

[3] Z. Ding, Y. Peng, and R. Pan. A bayesian approach to uncertainty modeling in OWL ontology. In *Proc. of International Conference on Advances in Intelligent Systems - Theory and Applications*, Nov. 2004.

[4] M. Giordanino, C. Giuliano, D. Kužnar, A. Lavelli, M. Možina, and L. Romano. Cross-media knowledge acquisition: A case study. In J. Magalhaes and S. Nikolopoulos, editors, *Proceedings of the SAMT-2008 Workshop on Cross-Media Information Analysis, Extraction and Management*, volume 437 of *CEUR-WS*, 2008.

[5] A. Laender, B. Ribeiro-Neto, A. Silva, and J. Teixeira. A brief survey of web data extraction tools. *SIGMOD Record*, 31(2), June 2002.

[6] J. Magalhães and S. Rüger. Information-theoretic semantic multimedia indexing. In *ACM Conference on Image and Video Retrieval (CIVR)*, Amsterdam, Holland, 2007.

[7] S. Nikolopoulos, C. Lakka, I. Kompatsiaris, C. Varytimidis, K. Rapantzikos, and Y. Avrithis. Compound document analysis by fusing evidence across media. In *Proceedings of the 7th International Workshop on Content-Based Multimedia Indexing (CBMI 2009)*, Chania-Crete, Greece, 2009.

[8] P. A. Viola and M. J. Jones. Rapid object detection using a boosted cascade of simple features. In *CVPR (1)*, pages 511–518, 2001.

[9] D. Šuc, D. Vladušič, and I. Bratko. Qualitatively faithful quantitative prediction. *Artificial Intelligence*, 158(2):189–214, 2004.

[2] http://www.reveal-this.org/

[3] http://www.muscle-noe.org/

[4] http://www.mesh-ip.eu/