Experiments on Meta-data Generation of Web Business Charts

Horacio Saggion

Department of Computer Science, University of Sheffield
Regent Court, 211 Portobello Street,
Sheffield, S1 4DP
United Kingdom
{H.Saggion}@dcs.shef.ac.uk

Abstract. The availability of multimedia material on the Internet as well as on private Intranets and multimedia collections poses new challenges for information access systems. We present a number of experiments on identification of conceptual information in graphics which can be used for semantic indexing and search. We investigate the use of rule-based and machine learning systems which operate over clean and noisy textual sources associated to images and charts for the identification of concepts expressed in a domain ontology and which are used to index the images. Results show that the investigated techniques support the process of image annotation.
Keywords: Information Extraction; Ontology; Business Intelligence; Natural Language Processing; Business Images; Image Annotation

1 Introduction

The widespread availability of tools for uploading multimedia content on the Internet in the form of images, audio, video or other forms of binary encoding, has resulted in huge amounts of non-textual data available to Internet users today. While traditional search engines offer good functionalities for searching and retrieving textual data, support for similar functionalities for non-textual data is lagging behind. Methods of semantic annotation could therefore assist in the development of applications that may perform tasks such as indexing, categorisation, search or retrieval of non-textual data.

Non-textual semantic annotation can be described as the task of annotating non-textual data with semantic labels or tags that serve as meta-data of the semantic categories of the associated annotated parts. The work to be described here has been carried out in the context of the business intelligence (BI) project where multimedia content plays an important role. For example, financial news articles, web pages, and business reports tend to contain an impressive number of images/graphics depicting economic performance-related data. The Musing projects is developing tools and modules based on natural language processing (NLP) technology to mitigate the efforts involved in gathering, merging, and analysing/annotating multisource, multi-media information for BI applications.

In particular we concentrate on the application of *Ontology-based Information Extraction* (OBIE), the process of identifying in text and other sources relevant concepts, properties, and relations expressed in an ontology. Our ontology-based information extraction system [9] has been developed with the GATE platform which provides a set of tools for development of information extraction applications. In particular GATE provides support to work with ontologies. Musing works with *domain ontologies* which represent the domain of application and which capture the experts' knowledge. Here we address the problem of extracting from graphics relevant information which can be used for rich semantic indexing and search, thus allowing business analysts ontology-based access to images for decision making. An example of this would be the identification of graphics showing *oil production in African countries in a particular time period* or graphics *comparing the sales of car manufacturers*, etc. Note that using mere keywords to annotate and retrieve business graphics would not be enough to achieve the level of precision required by the investigated scenarios. It is worth noting that the emphasis of this work is on how to take advantage of textual sources for image enrichment such as collateral texts and optical character recognition (OCR) output, and the question of analysing the image itself is out of the scope of this work.

As it will be show later in the paper, the use of robust information extraction and supervised machine learning trained over clean text can be used to extract conceptual information from noisy data for meta-data generation. To the best of our knowledge, and with the exception of table processing [13] – which is also a key element in BI applications, the issue of meta-data generation in the domain of business graphics has been little investigated.

The structure of the paper is as follows: in Section 2, we describe the corpus resources acquired for experimentation and annotation for system development; in Section 3 we describe the procedure and architecture for the semantic annotation of non-textual data and, in Section 4, present details on optical character recognition, unsupervised and supervised experiments. In Section 5, we present related work and finally, in Section 6 we close with a description of current and future work.

2 Corpus Collection Methodology

We concentrate on non-textual data of two types: (i) images of business graphs and charts extracted from Internet websites in English relating to business, finance and marketing reports and announcements, and (ii) tables in PDF format in English relating to information in various location in the globe. The work reported in the paper refers to type (i), however the techniques applied for producing meta-data are also valid for type (ii).

For the collection of business graphs, an initial set of seed pages that included business graphs was created queries were issued to Google using specific keywords selected from these pages. From an initial set of about 300 business graphs in the retrieved pages, a set of selection criteria were used to filter out the

collection of images and produce a smaller set which includes 60 business graphs containing textual data within the image. The encoding/compression formats used are JPEG, GIF and PNG.

A simple sub-classification of the collected images reveals that the collection consists of 27 images with graphs.

A manual inspection of the corpus of images revealed that the clearly identified information which can be categorised and targeted for semantic annotation included the following entity classes: Names of organisations including companies (e.g. OPEC, Ford); Money amounts (e.g. $5000, 46.2p); Date information (e.g. 2001, Aug-01); Percentages (e.g. 8.64%); Names of locations including those of countries (e.g. Persian Gulf, Canada); Names of people (e.g. Robert J Shifer).

Note that there is much more textual data in the image collection such as, for example, those found in the graph title or legends; however, most of this information is generally about the description of the graph or chart (e.g. *Retail Segment Profit/Loss - in millions $*) rather than individual named entities or concepts that can be annotated.

3 Meta-data Annotation System

The system architecture for the annotation of non-textual data is based around the processing pipeline shown in Figure 3.

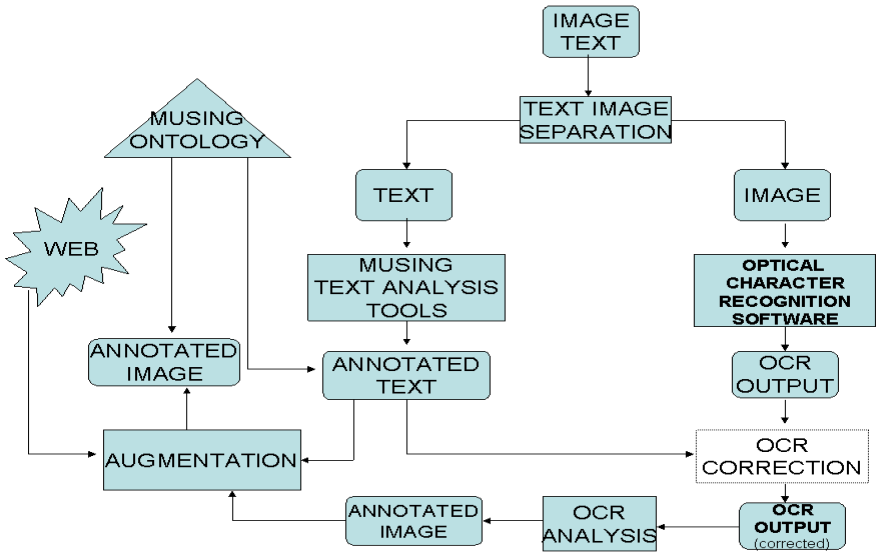

Fig. 1. Meta-data Annotation System

The steps followed in order to obtain a corpus of images annotated with semantic information relevant for meta-data creation is as follows:

- the html pages containing the images are separated into text files and graphic files using as clue the position of the image in the text;
- the text is analysed using an information extraction system adapted in order to obtain texts annotated with respect to Musing ontologies;
- the images are analysed by an off-the-shelf optical character recognition software in order to obtain recognised texts (OCR output);
- an optional correction step is attempted to correct some of the misrecognised tokens;
- the OCR output is analysed by modified versions of the text annotation tools in order to obtain a semantically enriched image;
- a further step aims at augmenting the image with more information from collateral but associated sources.

The semantic annotations produced by this process are linked to a predefined ontology a subset of which is used to describe the corpus domain. Thus, the low-level information in the image collection is mapped to high-level meta-data which can be useful for indexing, browsing, searching and retrieval applications. The annotated images are sent back to our Musing platform for storage and indexing.

3.1 Optical Character Recognition

Off-the-shelf OCR software was used to extract the textual content from non-textual data. The OCR software used for this task was ABBYY FineReader Professional (ver. 8.0) which is generally claimed to achieve very good OCR recognition accuracy with binary files. Output of the OCR process is text in plain, html or MS .doc format.

In the context of extracting useful information from images, the assessment of word accuracy in the OCR output is not by itself the focus of this work since only a small proportion of the image text can be used for metadata. Although the OCR process included a spelling process by default, its effectiveness was limited because of the presence of a lot of data with reduced dictionary coverage (e.g. mentions of money, percentages, dates etc.) and the grammatical structure (or lack of it) of the remaining text (such as text in captions, axis titles, or legends). The OCR performance for this collection was found to be good for some of the images but quite poor for other images with OCR errors ranging from simple character substitutions (e.g. **$** to **S**, **O** to **D** etc.) to complete word substitutions and replacement by white space characters.

Examples of errors by the OCR software can be seen in Table 1.

Some correction of these errors might help improve semantic annotation, but the correction, if carried out, must be done with care and in particular attention must to be paid to the selection of tokens to be corrected. The question of what is a valid token in a given context is important here. For example if the word **it** in the image is recognised as **H** how one may decide that it should be corrected

OCR Output	Correct Token	OCR Output	Correct token
roup	group	w hat	what
befor	before	conputer	computer
n	on	Lil¿/.i	Lybia
BayNetworks	Bay Networks	bws	lows
$8,S94	$8,594	doHars	dollars
I960	1960	st1ouis	stlouis

Table 1. Examples of incorrect token recognition.

based on context? (if the token **H** is part of a phrase where the likelihood of finding **H** instead of **it** is very low). There are cases where the recognised tokens are clearly out-of-vocabulary items: for example **doHar** (recognised instead of **dollar**) seems incorrect (the word is not an English word) but still this depends on the context. The analysis of the errors reveals that some corrections (if correctly carried out) may help extract from the images indexing terms (**computer, group profit before tax** etc.) a well as - but in less quantity - entities such as dates and monetary values. However the correction of pseudo numeric tokens should be done very carefully and based on context.

Given the erroneous OCR output and the task of annotating named entities which may have been misrecognised, one of the questions is whether the annotation performance might be better if correction to the OCR output is performed before annotation.

To test this idea, an error correction module was developed that applies fuzzy pattern matching to the OCR text based on Levenshtein distance [2] similar to the way the egrep utility [14] works and to [11]. Given a set of patterns, each token in the OCR text is matched against each pattern using a dynamic programming algorithm. If the Levenshtein distance between the pattern and the OCR token is below a predefined threshold, the pattern is regarded as candidate for replacing the token in the text.

One of the issues with this approach is the definition of token in the OCR output. Considering the fact that string edit distance may result to substantial mistakes if applied to low level tokens (such as numerals, uppercase/lowercase sequence of characters, symbols etc.) because of the reduced string length that such tokens will have we chose to use tokens at the word level i.e. tokens separated by white spaces in the OCR text.

A second issue is the selection of a suitable pattern set. Taking into account the domain of the image text (business/scientific/statistical) and the fact that most of the annotations in it would refer to expressions that cannot be matched by using typical spelling dictionaries; we opted to use as patterns expressions taken from two sources: (a) expressions that refer to named entities in the corresponding html source for each image, and (b) all distinct tokens in the html source.

Note that the efficiency of error correction in case (a) is dependent on the likelihood that a named entity which has been misrecognised by OCR happens to be mentioned in the source html text. Also note that the applicability of using string edit distance for misrecognised mentions of date and money expressions is limited due the following reasons:

- most of the date expressions refer to years and unless a year expression that is misrecognised is mentioned in the source text literally then it can be easily confused for some other year by the string edit-distance algorithm (e.g. 2O04 is matched to 2005 with Levenshtein distance);
- most money expressions are either recognized correctly or recognized with a lot of errors;

to correct the errors, a high Levenshtein distance threshold should be used; but that might effect in a lot of incorrect matches with other money expressions found in the source text. Using all tokens in the html source texts as patterns (case (b)) is based on the logic that the extra patterns (note that this set of patterns already include the patterns in case (a)) may not be useful in correcting any misrecognised named entities in the images but they may be useful in correcting the context around the named entities. That may still be helpful to an annotation system that uses contextual information for annotation.

4 Information Extraction Experiments

We manually annotated the texts produced by OCR and the textual forms of the associated HTML documents for the following common IE targets: Date, Person, Organization, Location, Money and Percent which are incorporated into the Musing Ontology. Figure 2 shows the collateral text of an image annotated with respect to the Musing ontology while Figure 3 shows the corresponding OCR of the image annotated with respect to the ontology.

	Collateral Text			OCR'ed Text		
Type	Precision	Recall	F-Measure	Precision	Recall	F-Measure
Date	83%	96%	89%	84%	87%	85%
Person	30%	89%	45%	-	-	-
Organization	67%	55%	60%	56%	32%	40%
Location	85%	69%	76%	82%	82%	82%
Percent	94%	94%	94%	93%	99%	96%
Money	78%	75%	76%	68%	34%	45%
Overall	74%	87%	73%	84%	87%	78%

Table 2. Information Extraction over OCR'ed and Collateral Texts

We then processed the OCR results and the collateral texts with our semantic annotation tools. In order to measure extraction accuracy, we use information

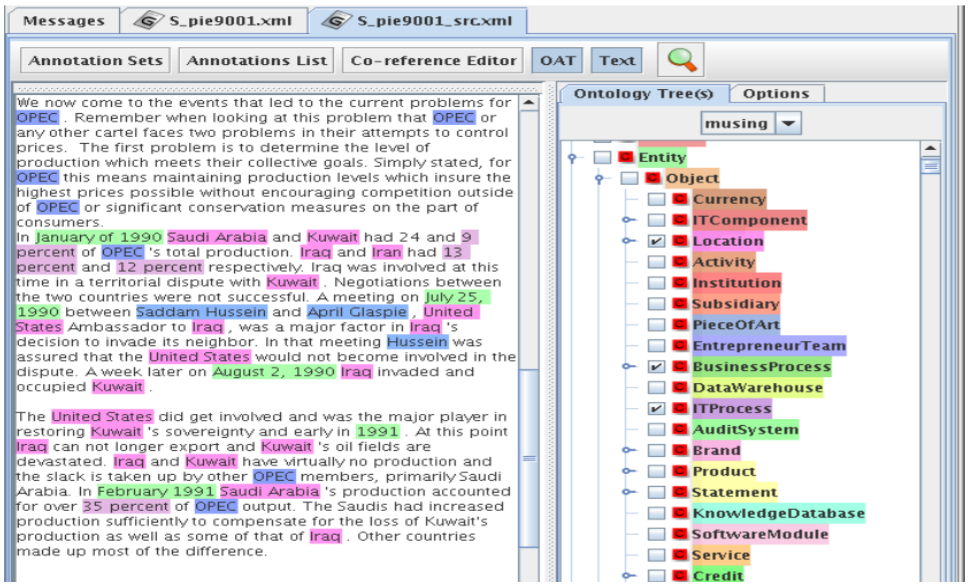

Fig. 2. Text annotated with respect to the Musing ontology

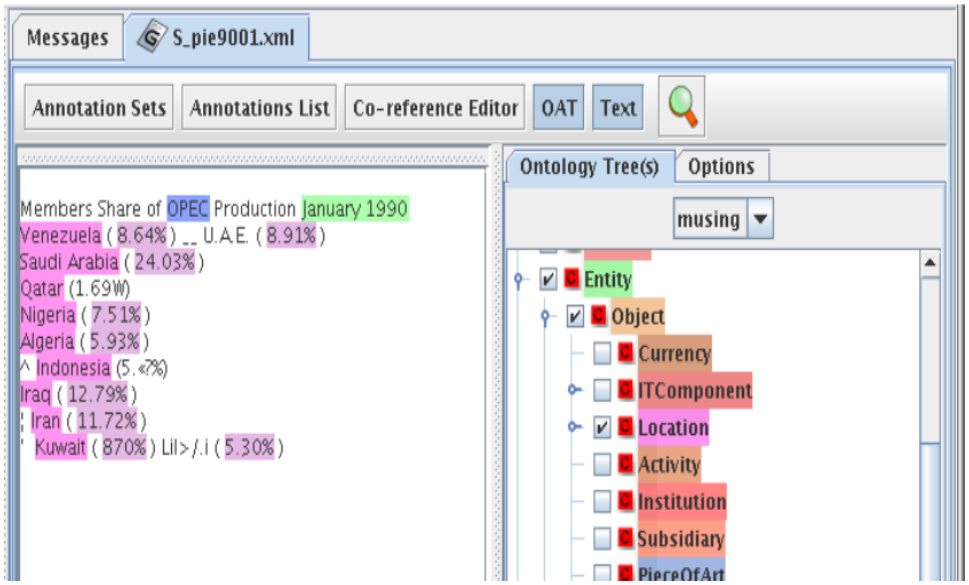

Fig. 3. Text of image annotated with respect to the Musing ontology

extraction evaluation metrics [1]: precision, recall, and F-measure. Results are reported in Table 2 (note that there are no Person annotations on the OCR output.)

In summary, we obtain reasonable and similar results over both OCR data and the associated texts. The poor performance on identification of Organizations in the noisy documents can be explained by the fact that the extraction system, in this particular case, is relying on lexical knowledge (e.g. gazetteer) which might be incomplete for dealing with these texts. In normal unstructured documents the lack of lexical information can be palliated by the use of contextual information in the grammar rules. The supervised experiments in the next section show a way to solve this.

4.1 Supervised Learning Experiments

We have also carried out experimentation with the software TIES [3]. TIES uses a variety of parameters for learning and testing the input data including parameters for feature extraction, validation strategy (cross-validation, split validation etc.), configuration of the weak learner (look ahead, whether non-labelled tokens should be considered negative examples etc.). The number of combinations of these parameters can be vast but in practice the results for the different strategies are very close (at least for the Musing data tested here). The features used for these experiments are the defaults offered by TIES' tokeniser.

Based on them, rules for extraction are learnt from annotated examples. The images annotated with semantic information described above where used to learn a system for annotation of semantic information. A number of experiments where carried out, here we concentrate on a configuration which gave us the best results which consist on training the software with all the available data (minus one document) and testing over the leave-out-document. The results are shown in the Table 3. In these experiments, we observed a considerable improvement in performance with respect to the generic extraction system. This might be attributed to the fact that the system was trained on the same type of data it was tested on.

Type	Precision	Recall	F-Measure
Date	100%	100%	100%
Organization	87%	98%	92%
Location	100%	100%	100%
Percent	98%	98%	98%
Money	94%	94%	94%
Overall	95%	98%	96%

Table 3. Trained text system over OCR data

To test the efficiency of the learned system on the OCR results we split the document set in two parts: the (annotated) source html files were used for training and the learnt system was used for testing over the OCR data. The results were very encouraging. Locations were identified with 100% accuracy, Organizations with almost perfect accuracy, amounts of money with recall of 65% but very low precision (similar to Percentages). Finally Dates have accuracy of 100% (note that no Person names were present in the images). So it seems that with enough data, a learning system can correctly extract some types of information such as dates, locations and organizations from noisy data (i.e. OCR'ed images) as long as training has been performed on their sources.

5 Related Work

Images and other multimedia material are usually annotated with metadata by humans who follow strict guidelines and use specially developed controlled vocabularies. Work on annotation of images and other multimedia artifacts with respect to an ontology is not new [4]. However, relying on fully manual annotation is impractical in many situations. Annotation of non-textual sources with semantic indices derived from clean and noisy data has been carried out in specific domains [10]. The most common method of automatically enriching images with semantics is by relying on the images' associated captions. Keywords can be extracted from the captions in order to produce indexing terms. Deeper analysis of the captions can also be used in order to identify entities of interest (e.g. persons) [12]. Statistical methods relying on term frequency and inverted term frequency have also been applied for image classification in categories such as in-door or out-door environment, the idea is that terms such as sky, blue, etc. are statistically associated with outdoor images while other terms (e.g. furniture) might be associated with in-door images [8]. Full syntactic analysis of texts for the purpose of semantic indexing has also been attempted in order to create complex logical terms for indexing [6, 7]. Most of the above approaches rely on clean well formed text to carry out extraction where linguistic analysis such as parsing and semantic representation can be applied, we differ in the application of information extraction techniques to noisy unstructured text. Work on processing text in images has mainly concentrated on optical character recognition [5], we are more interested in the creation of semantics out of textual material.

6 Conclusions

The availability of multimedia material on the Internet as well as on private Intranets and multimedia collections poses new challenges for information access systems. The automatic generation of metadata from textual sources for business images has been little investigated in the literature. We have presented a number of experiments to identify conceptual information in graphics which can be used for semantic indexing and search in the context of the Musing project. Our experiments show that a system developed for identification of concepts in clean

text can achieve comparable performance over noisy (OCR) output. We have also shown that a learning system – trained on clean text associated to the images – can be used to reliably annotate the images.

7 Acknowledgements

We would like to thank the comments and suggestions of two anonymous reviewers. This work is partially supported by the EU-funded Musing project (IST-2004-027097).

References

1. Nancy Chinchor. MUC-4 Evaluation Metrics. In *Proceedings of the Fourth Message Understanding Conference*, pages 22–29, 1992.
2. Maxime Crochemore and Wojciech Rytter. *Text algorithms*. Oxford University Press, Inc., New York, NY, USA, 1994.
3. D. Freitag and N. Kushmerick. Boosted wrapper induction. In *AAAI/IAAI*, pages 577–583, 2000.
4. L. Hollink, G. Schreibe, J. Wielemaker, and B. Wielinga. Semantic annotation of image collections, 2003.
5. K. Jung and A.K. In Kim, K. Jain. Text information extraction in images and video: a survey. *Pattern Recognitio*, 37(5):977–997, May 2004.
6. K. Pastra, H. Saggion, and Y. Wilks. Nlp for indexing and retrieval of captioned photographs. In *EACL '03: Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics*, pages 143–146, Morristown, NJ, USA, 2003. Association for Computational Linguistics.
7. T. Rose, D. Elworthy, A. Kotcheff, A. Clare, and P. Tsonis. Anvil: a system for the retrieval of captioned images using nlp techniques, 2000.
8. C.L. Sable and V. Hatzivassiloglou. Text-based approaches for the categorization of images. In Serge Abiteboul and Anne-Marie Vercoustre, editors, *Proceedings of ECDL-99, 3rd European Conference on Research and Advanced Technology for Digital Libraries*, number 1696, pages 19–38, Paris, FR, 1999. Springer Verlag, Heidelberg, DE.
9. H. Saggion, A. Funk, D. Maynard, and K. Bontcheva. Ontology-based information extraction for business applications. In *Proceedings of the 6th International Semantic Web Conference (ISWC 2007)*, Busan, Korea, November 2007.
10. H. Saggion, J. Kuper, T. Declerck, D. Reidsma, and H. Cunningham. Intelligent Multimedia Indexing and Retrieval through Multi-source Information Extraction and Merging. In *IJCAI 2003*, Acapulco, Mexico, 2003.
11. R. Schneider and I. Renz. The Relevance of Frequency Lists for Error Correction and Robust Lemmatization. In *Proceedings of the 5mes Journes Internationales d'Analyse Statistique des Donnes Textuelles (JADT 2000)*, 2000.
12. R.K. Srihari. Automatic indexing and content-based retrieval of captioned images. *IEEE Computer*, 28(9):49–56, 1995.
13. M. Vilain, J. Gibson, and R. Quimby. Table classification: an application of machine learning to web-hosted financial texts. In *Proceedings of Recent Advances in Natural Language Processing*, 2007.
14. S. Wu and U. Manber. Fast text searching with errors. Technical report, University of Arizona, 1991.