NLP for Indexing and Retrieval of Captioned Photographs

Katerina Pastra, Horacio Saggion, Yorick Wilks
Department of Computer Science
University of Sheffield
England - UK
Tel: +44-114-222-1800
Fax: +44-114-222-1810
{katerina,saggion,yorick}@dcs.shef.ac.uk

Abstract

We present a text-based approach for the automatic indexing and retrieval of digital photographs taken at crime scenes. Our research prototype, SOCIS, goes beyond keyword-based approaches and methods that extract syntactic relations from captions; it relies on advanced Natural Language Processing techniques in order to extract relational facts. These relational facts consist of a "pragmatic relation" and the entities this relation connects (triples of the form: ARG1-REL- ARG2). In SOCIS, the triples are used as complex image indexing terms; however, the extraction mechanism is used not only for indexing purposes but also for image retrieval using free text queries. The retrieval mechanism computes similarity scores between query-triples and indexing-triples making use of a domain-specific ontology.

1 Indexing and Retrieval of Photographs

The normal practice in human indexing or cataloguing of photographs is to use a text-based representation of the pictorial record having recourse to a controlled vocabulary or to "free-text". On the one hand, an index using authoritative sources (e.g., thesauri) ensures consistency across human indexers, but at the same time it renders the indexing task difficult due to the size of the keyword list that is used - not to mention the cumbersome and unintuitive requirement impose to the user, to become familiar with using specific wording for the subsequent retrieval of the images. On the other hand, the use of free-text association, while natural, makes the index representation subjective and error prone. Content-based Image Processing methods are used as an alternative to the manual-annotation bottleneck (Veltkamp and Tanase, 2000). Content-based indexing and retrieval of images is based on features such as colour, texture, and shape. Yet, image understanding is not well advanced and is very difficult even in closed domains. When linguistic descriptions of the photographs are available (i.e., captions or collateral texts), they can be used as the starting point for indexing. We have focused on the development and implementation of automatic caption-based techniques for indexing and retrieval of photographs taken at scenes of crime (SOC).

Researchers in information retrieval argue that detailed linguistic analysis is usually unnecessary to improve accuracy for text indexing and retrieval; however, in the case of captioned photographs, natural language processing (NLP) techniques have proved to be particularly effective for the very same tasks (Rose et al., 2000; Guglielmo and Rowe, 1996).

Current approaches in automatic text-based image indexing fail in capturing semantic information expressed in the captions, that is important for the subsequent retrieval of the images (Pastra et al., 2002). Unlike traditional "bag of words" techniques and other methods for extracting syntactic relations from captions for indexing pur-

poses, our prototype extracts meaning representations that capture pragmatic relations between objects depicted in the photographs. Therefore, most of the complexity of the written text is eliminated, while its meaning is retained in an elegant and simple way. The relational facts that are extracted are of the form: ARG1-RELATION-ARG2 and they are used as indexing terms for the crime scene visual records. In these triples, the arguments may be simple or complex noun phrases, whereas the relations express locative arrangements, part-of associations and other relations, all coming up to 17 different relations as indicated through the analysis of a corpus of 1000 captions. The notion of extracting structres that capture semantic relations among entities originates from early theories on text representation. Our approach bears a loose connection to the "Preference Semantics" theory (Wilks, 1975; Wilks, 1978); however, in the latter, the RELATIONs captured in semantic templates were a mixture of CASE and ACT denoting relations, whereas SOCIS focuses on "static", pragmatic relations between tangible objects. The binary relational templates extracted by SOCIS allow for the indexing terms to capture semantic equivalences and differences that go beyond syntactic dependencies, bindings to specific wording or implied information such as the absence/presence of objects : "red substance on yellow table" vs. "yellow substance on red table", "knife on table" vs. "blade on bar counter", and "cable around neck" vs. "neck with cable removed" respectively.

SOCIS consists of a pipeline of processing resources that perform the following tasks: (i) pre-processing (e.g., tokenisation, POS tagging, named entity recognition and classification, etc.); (ii) parsing and naive semantic interpretation; (iii) inference; (iv) triple extraction.

The rest of this paper describes our method for indexing and retrieval using relational facts.

2 Ontology and Indexing Terms

We have made use of the British Police Information Technology Organisation Common Data Model and a collection of formal reports produced by scene of crime officers (SOCO) to develop OntoCrime, a concept hierarchy that structures concepts relevant to SOC investigation (e.g., physical evidence, trace evidence, weapon, cutting instrument, criminal event etc.). The ontology is used during indexing-term computations. Two types of indexing terms are obtained for each caption: (i) "lexical" terms, which are canonical representation of objects mentioned in the caption; and (ii) triples of the form *(*$Argument_1$*, Relation,* $Argument_2$*)*, where $Relation$ is the name of the relation and $Argument_i$ are its arguments. The arguments have the form $Class : String$, where $Class$ is the immediate hypernym the entity belongs to (according to OntoCrime), and $String$ is of the form $(Adj|Qual) * Head$, where $Head$ is the head of the noun phase and Adj and $Qual$ are adjectives and nominal qualifiers syntactically attached to the head. For example, the noun phrase "the left rear bedroom" is represented as *premises : left rear bedroom* and the full caption "neck with cable removed" is represented as *(body part : neck, Without, physical object : cable).*

3 NLP Processes

We have used some resources available within GATE (Cunningham et al., 2002) and have integrated a robust parser and inference mechanism implemented in Prolog. The preprocessing consists of a simple tokeniser that identifies words and spaces, a sentence segmenter, a named entity recogniser specially developed for the SOC, a POS tagger, and a morphological analyser. The NE recogniser identifies all the types of named entities that may be mentioned in the captions such as: *address*, *age*, *conveyance-make*, *date*, *drug*, *gun-type*, *identifier*, *location*, *measurement*, *money*, *offence*, *organisation*, *person*, *time*. It is a rule-based module developed through intensive corpus analysis and implemented in JAPE (Cunningham et al., 2002), a regular pattern matching formalism within GATE. Part of speech tagging is done with a transformation-based learning tagger whose lexicon has been adapted to the SOC, and lemmatisation is performed with a robust rule-based system. The lexicon of the domain was obtained from the corpus and appropriate part of speech tags were produced semi-automatically (this lexicon is used during POS tagging).

Logical forms for each caption are obtained through a bottom-up parsing component that uses a context-free syntactic-semantic grammar. Logical forms are mapped into the ontology using a lexicon attached to the ontology (implemented in XI (Gaizauskas and Humphreys, 1996)) and a number of rules. After the "explicit" semantics is mapped into the ontology, the following procedure is applied: each triple mapped onto the model is examined in the order it is asserted. For each triple X-Rel-Y, the system checks whether X and Y occur as arguments in other relations and in that case rules that account for transitive and distributive properties of the semantic relations such as AND-distribution, WITH-transitivity, WITH-distribution, etc. are fired to infer new triples (Pastra et al., 2003). Our AND-distribution rule over "On" is stated with the following rule:

If X-And-Y & Y-On-Z **Then** X-On-Z

The WITH-distribution rule is stated as follows:

If X-With-Y & Y-REL-Z **Then** X-REL-Z

So a caption such as "knife together with revolver in kitchen" is represented with the triples:

- (i) *(cutting instrument : knife, With, firearm : revolver)*
- (ii) *(firearm : revolver, In, part of dwelling : kitchen)*
- (iii) *(cutting instrument : knife, In, part of dwelling : kitchen)*

where triple (iii) was inferred using the rule.

We have evaluated the triple extraction and inference mechanism using a test corpus of 500 captions and obtained accuracy of 80%. This glass-box evaluation has indicated refinements to the extraction rules and has also enhanced the set of inferences that the system should be able to make.

4 Querying and Retrieval

The same semantic representation mechanism is also used for retrieval; SOCIS allows for free text querying. The system's interface prompts the user to think as if completing a sentence of the form "show me all the photographs in the database that depict...". This query is then processed exactly as if it was a caption (as described in the previous section 3). Relational facts are extracted from the query, if possible. These relational facts are then matched against each photograph's indexing terms and similarity scores are computed. For triples to match, their RELATION slot has to be identical. Then, a score is computed that takes into account class and argument similarity. OntoCrime is used to compute the semantic distance of the nodes needed to be transversed in order to find a class match. The formula we implement for computing the similarity between query term $T_1 = (Class_1 : Arg_1, Rel, Class_2 : Arg_2)$ and indexing term $T_2 = (Class_3 : Arg_3, Rel, Class_4 : Arg_4)$ is as follows:

$$
\begin{aligned}
&Sim(T_1, T_2) = \\
&\alpha_1 * OntoSim(Class_1, Class_3) + \\
&\alpha_2 * OntoSim(Class_2, Class_4) + \\
&\alpha_3 * ArgSim(Arg_1, Arg_3) + \\
&\alpha_4 * ArgSim(Arg_2, Arg_4)
\end{aligned}
$$

where $OntoSim(X, Y)$ is the inverse of the length between X and Y in OntoCrime, and $ArgSim(A, B)$ is computed using the formula:

$$
\begin{aligned}
&ArgSim(A, B) = \\
&\beta_1 * Match(A_{Head}, B_{Head}) + \\
&\beta_2 * Match(A_{Qual}, B_{Qual}) + \\
&\beta_3 * Match(A_{Adj}, B_{Adj})
\end{aligned}
$$

where $Match(X, Y)$ is 1 when $X = Y$ and 0 when $X \neq X$. The weighs α_i and β_i have to be experimentally identified. When more than one relational fact is extracted from the query, the system attempts to match each query triple with each indexing term of each photograph and a sum of the scores that each photograph receives is calculated and used for the final selection of the most appropriate images to be returned to the user. In cases when no relational facts can be extracted from the query, simple keyword extraction (following the rules for argument extraction for the triples) and matching takes place, using the ontology for se-

mantic expansion.

5 Related Work

The use of conceptual structures as a means to capture the essential content of a text has a long history in Artificial Intelligence. For SOCIS, we have attempted a pragmatic, corpus-based approach, where the set of primitives emerge from the data. MARIE (Guglielmo and Rowe, 1996) is a system that uses a domain lexicon and a type hierarchy to represent both queries and captions in a logical form and then matches these representations instead of mere keywords; the logical forms are case grammar constructs structured in a slot-assertion notation. Our approach is similar in the use of an ontology for the domain and in the fact that transformations are applied to the "superficial" forms produced by the parser to obtain a semantic representation, but we differ in that our method does not extract full logical forms from the semantic representation, but a finite set of possible relations. Also related to SOCIS is the ANVIL system (Rose et al., 2000) that parses captions in order to extract dependency relations (e.g., head-modifier) that are recursively compared with dependency relations produced from user queries. Unlike SOCIS, in ANVIL no logical form is produced nor any inference to enrich the indexes.

6 Work in Progress

The SOCIS prototype is a web-based application that allows SOC officers to upload digital photographs and their descriptions in a central database, index the photographs automatically according to these textual descriptions and retrieve them using free text queries. The retrieval mechanism is currently being implemented. Once the retrieval will have been fully implemented, proper usability testing of the whole system by real users will take place and a comparison of our free-text retrieval approach to other approaches that allow for unrestricted natural language queries will be undertaken. During the system's development cycle usability evaluation through constant user assessment has been carried out with the help of the project's advisory board consisting of scene of crime officers and investigators. This preliminary feedback has indicated that making use of relational facts in order to make a digital image collection accessible with high precision and recall, since expressing such relations in both captions and queries is intuitive for the target users of SOCIS.

References

H. Cunningham, D. Maynard, K. Bontcheva, and V. Tablan. 2002. GATE: A framework and graphical development environment for robust NLP tools and applications. In *Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics*, Philadelphia, PA.

R. Gaizauskas and K. Humphreys. 1996. XI: A Simple Prolog-based Language for Cross-Classification and Inhetotance. In *Proceedings of the 7th International Conference in Artificial Intelligence: Methodology, Systems, Applications*, pages 86–95, Sozopol, Bulgaria.

E. Guglielmo and N. Rowe. 1996. Natural language retrieval of images based on descriptive captions. *ACM Transactions on Information Systems*, 14(3):237–267.

K. Pastra, H. Saggion, and Y. Wilks. 2002. Extracting Relational Facts for Indexing and Retrieval of Crime-Scene Photographhs. In A. Macintosh, R. Ellis, and F. Coenen, editors, *Applications and Innovations in Intelligent Systems X*, British Computer Society Conference Series, pages 121–134. Springer Verlag.

K. Pastra, H. Saggion, and Y. Wilks. 2003. Intelligent Indexing of Crime-Scene Photographs. *IEEE Intelligent Systems, Special Issue in Advances in Natural Language Processing*, 18(1):55–61.

T. Rose, D. Elworthy, A. Kotcheff, and A. Clare. 2000. ANVIL: a System for Retrieval of Captioned Images using NLP Techniques. In *Proceedings of Challenge of Image Retrieval*, Brighton, UK.

R. Veltkamp and M. Tanase. 2000. Content-based image retrieval systems: a survey. Technical Report UU-CS-2000-34, Utrecht University.

Y. Wilks. 1975. A Preferential, Pattern-Seeking, Semantics for Natural Language Inference. *Artificial Intelligence*, 6:53–74.

Y. Wilks. 1978. Making Preferences More Active . *Artificial Intelligence*, 11:197–223.