Extracting Relational Facts for Indexing and Retrieval of Crime-Scene Photographs

Katerina Pastra, Horacio Saggion, Yorick Wilks
Department of Computer Science, University of Sheffield
211 Portobello Street, S1-4DP, Sheffield, U.K.
tel. +44 114 2221945, fax. +44 114 2221810
{katerina,saggion,yorick}@dcs.shef.ac.uk

pp. 121-134

Abstract

This paper presents work on text-based photograph indexing and retrieval for crime investigation, an application domain where efficient querying of large crime-scene photograph databases is of crucial importance. Automating this task will change current police practices considerably, by bringing "intelligence" to crime support information systems. The prototype presented, goes beyond common approaches to the automation of image indexing and retrieval, by applying a novel method that captures deep semantic relations expressed in the captions that accompany crime-scene photographs. The extraction of these semantic triples is based on advanced knowledge-based Natural Language Processing technologies and resources.

1 Introduction

Crime investigation is a task of considerable social importance that relies - among others - on both efficient multi-modal documentation of the crime scenes and effective retrieval of the information documented. The current practice in documenting a crime scene involves a Scene Of Crime Officer (SOCO) taking photographs or/and video recording the scene, gathering, packaging and labelling evidence, as well as producing an official report of his/her actions

and filling in a considerable number of other forms needed[1]. After having the photographs developed, the SOCO creates a "photo-album" for the case which includes an index page with a caption for each photograph or set of photographs [7]. All this documentation procedure is largely done manually, which results in SOCOs devoting significant time and effort for producing handwritten reports and repeating large chunks of information again and again as they fill in various forms for the same case. On top of this, retrieving available information from past cases and indicating possible similarities or "patterns" among cases relies exclusively on the investigators memory and availability for going through piles of case files.

The benefits for using Information Systems for supporting Crime Investigation are obvious and have been proved through the great success of using Information Technology for tasks such as fingerprint identification and criminal profiling. During the last two decades more and more police information technology organisations are established in countries all over the world for supporting crime investigation in all possible ways. The British "National Strategy for Police Information Systems" (NSPIS) program launched in mid-eighties an administrative support system for major crime. A decade later, the need for crime investigation support led to the creation of HOLMES 2, a crime investigation management system. HOLMES 2 allows for monitoring and control of document flow throughout the investigation, visualisation of the sequence of events and the links between them, task and exhibit management, computer-aided preparation of court documents and keyword retrieval of structured data. The first version of the system is currently under live operational pilot testing in various U.K. police forces and similar systems have been developed from software companies [7]. Clearly, these attempts to support crime investigation will change current practices by helping police officers to save time and effort in their every day work. However, "intelligent support" is still not provided; there is a demanding need to take things a step further and exploit mature Artificial Intelligence (AI) technologies, if one is to provide for substantial support for crime investigation.

Providing "intelligent" support for this domain is the general objective of SOCIS, a three-year EPSRC funded research and development project undertaken by the University of Sheffield and the University of Surrey in collaboration with four U.K. police forces[2]. The project aims at

[1]For example: "Exhibit labels" that are attached to each piece of evidence, "Fingerprint Examination Request Forms" that are sent to the corresponding laboratory along with the fingerprints collected and other.

[2]South Yorkshire Police, Surrey Police, Kent County Constabulary and Hampshire Constabulary form an advisory board for the Scene Of Crime Information System (SOCIS) project

Pastra et al. (2002) – In Proceedings of the British Computer Society Conference in Artificial Intelligence – pp. 121-134

the integration of various AI technologies that will change the current practices in documenting crime scenes and retrieving case-related information. It is envisaged that the SOCO should be able to use digital cameras for photographing the crime scene, instantly store the photographs in a central database along with verbal descriptions of their contents (captions) and other information needed for the case, provided via a hands-free microphone. The SOCIS prototype, which is still under development, will allow for:

- Automatic population of official reports, forms and photo-indexes from the information provided by the SOCO verbally and
- Intelligent indexing and retrieval of multimodal case documentation

In this paper, we will present our work onndexing and retrieval of crime-cene photographs by extracting relational facts from their corresponding captions. In particular, we will focus on the module developed for the extraction of relational triples, that will be used in SOCIS for both indexing and retrieval of images, leaving aside at the moment the description of the exact retrieval mechanism (i.e the matching of triples extracted from queries with the ones through which our images have been indexed).

First we will locate text-based Image Retrieval in relation to AI and classical Information Retrieval (IR) and we will give a brief overview of related research. Then we will present our method for indexing and retrieval of captioned images; our processing and language resources will be presented in detail. Features under development and issues to be explored further form the last section of the paper along with conclusions on the benefits and lessons to be learned from our own work.

2 Text-based Image Retrieval: looking at the wider context

Text-based approaches form what is widely known as the "concept-based" stream in Image Retrieval as opposed to the "content-based" one. The corresponding research communities in this field focus on different aspects of the images. The former focuses on the message to be conveyed and which is assigned by humans in the form of titles, captions or indexing terms, while the latter focuses on image attributes such as colour, texture and shape [17]. The problems with each approach are well known; visual attributes are not sufficient for image indexing and

Pastra et al. (2002) – In Proceedings of the British Computer Society Conference in Artificial Intelligence – pp. 121-134

computer vision techniques are not that advanced yet to assist for this task. On the other hand, text-based approaches carry all the problems encountered for decades in the traditional Information Retrieval research.

While non-connectionist and non-statistical AI remains wedded to the computational tractability and explanatory power of representations of propositions in some more or less logical form, classical Information Retrieval, often characterised as "a bag of words" approach to text, consists of methods for locating document content independent of any particular explicit structure in the data. It is not committed to any notion of representation beyond what is given by a set of index terms, or set of index terms along with figures themselves computed from text that may specify clusters, vectors or other derived structures. However, the infusion of Natural Language Processing (NLP) techniques, that involve deeper representation, into Information Retrieval has been attempted several times, not only for document retrieval but for text-based image retrieval too.

2.1 Image Caption Retrieval approaches: a brief review

Testing various semantic analysis approaches on caption indexing and retrieval has been performed by many researchers, in an attempt to overcome the well known problem of the "keyword barrier". Keyterms have been widely used for image classification; these terms are either manually assigned by human annotators or automatically extracted based on statistical analyses of accompanying captions. The time and effort needed for manually creating classification schemes for image retrieval, as well as the difficulty and un-natural task of forcing the user to get familiar with specific wording in order to retrieve the images of interest successfully, have led researchers to attempts for automating the extraction of keywords for such purposes, using statistical methods. Considering the fact that currently available Crime Investigation Support systems are based on keyword retrieval of information, one realises the need for a more advanced approach for the task. It has been shown that for text-based image classification pure statistical methods (e.g TF*IDF based approaches) can be used for coarse classification into general categories such as for example "indoor - outdoor" images, with great precision [13]. Still, these methods do not perform well when accuracy and user time is of primary importance for a task.

A light-semantic approach to Image Caption Retrieval is the automatic query expansion

for encountering synonymy and polysemy phenomena [11]. This usually involves a "semantic broadening" of the query words from the synset they belong to, as provided by lexical resources such as WordNet. The similarity of the keyword and other members of its synset is usually measured in terms of the distance between the relevant nodes in the WordNet hierarchy and their information content[3]. In ANVIL [11], both semantic expansion as well as statistical methods for keyword and phrase matching are used. The combination of these techniques is claimed to achieve very high precision scores, while recall is very low. Smeaton et al.(1996)[14], compare the performance of a purely statistical method for image caption retrieval (TF*IDF) against semantic approaches. Simple term expansion from sense disambiguated synsets improves retrieval to some extent[4]. Measuring conceptual similarity in terms of class rather than word similarity, a method suggested by the researchers themselves, proved to perform much lower because of noise in computing the similarity values. What proved to perform better was the fusion of the results of two asymmetrically different metrics: one that determines for each query term the most similar caption term and computes the overall caption scores and one that computes overall scores by determining for each caption term the best-matching query term. Still, hardly does even this method achieve a 50 percent of precision and recall.

The use of syntactic dependencies in documents, their indices and queries has also been explored for Information Retrieval [15], since traditional statistical methods are unable to cope with meaning differences such as "X adjacent to Y" and "Y adjacent to X". In Guglielmo and Rowe (1996) [4], a domain lexicon and a type hierarchy is used to represent both queries and captions in a logical form and then match these representations rather than mere keywords; the logical forms are case grammar constructs structured in a slot-assertion notation. Matching takes place in two phases; first an enhanced keyword search in noun and verb indexes created for the captions is performed and then the logical forms of the captions of the most promising image identifiers are matched against the queries' logical forms. This image caption retrieval system developed within the MARIE project for navy aircraft equipment photographs has been reported to achieve 30 percent more precision and 50 percent more recall over a standard keyphrase approach.

Relying on syntactic dependencies and concept classification information, seems very promis-

[3]Approximated by estimating the probability of occurence of a term in a corpus.

[4]This is attributed to the fact that both captions and queries are short, since it has been shown that the same does not apply in text retrieval applications [18].

Pastra et al. (2002) – In Proceedings of the British Computer Society Conference in Artificial Intelligence – pp. 121-134

ing for image caption retrieval according to the thorough evaluation experiments reported [4]. Research on extraction-based text categorization point to the same direction too [10]. In Riloff et al. (1999), domain dependent extraction patterns and semantic features associated with role fillers have been proved to perform better in classifying texts. These extraction patterns are domain-dependent linguistic expressions consisting of a trigger word, conditions to be met and case roles. These patterns are considered to be dependent on the syntactic context of tokens; verb forms and prepositions are considered important indicators of the meaning of classification terms [9]. In particular, extraction patterns that contain prepositions are reported to have much higher correlation with relevant texts than their corresponding trigger words.

We have extended these approaches, by trying to extract propositional - like objects from image captions automatically. This refers to an attempt to index images by binary relational templates of the form: "blood AROUND body" and "X ON Y", so that the document is indexed by these facts which allow for more effective searches to be conducted such as "find all images that depict a body surrounded by blood" or "find all images that depict Xs on tables". These searches can bring back a more restricted set of images than the one obtained when looking for example for all the image captions that mention both "Xs and tables" or "body and blood". As it will become clear in the next section, our corpus and application driven approach integrates various NLP technologies (cf. 3.3) and copes with problems that the previously mentioned approaches cannot overcome (cf. 4). We attempt therefore, to apply an AI technique the implementation of which relies largely on advances in NLP. By constraining the whole experiment to a very specific domain, with well established practices and real needs, and to a very specific type of text that is by definition concise and short, we have tried to create the appropriate conditions for getting valuable feedback on the benefits (or lack of) of developing such an application.

3 Extracting relational facts for SOCIS

Extracting labelled relations from short texts has also been attempted within the MindNet project with success [8]. Howerver, taking advantage of the descriptive and precise nature of the caption of crime scene imagery and the fact that it is mainly "static" scenes that are described rather than events, we have relied largely on relations denoted through prepositions, spatial

Pastra et al. (2002) – In Proceedings of the British Computer Society Conference in Artificial Intelligence – pp. 121-134

verbs and other adjuncts e.g: "tie around right arm", "body lying on the floor". Among others, the description of the position of the evidence at the crime scene is of sinificant importance for the domain, i.e the exact location and position of the objects found and the relation to each other. The triples we extract are of the form: ARG1-RELATION-ARG2 and they are used as indexing terms for the crime scene imagery.

In the sections that follow, we will describe our corpus of crime scene photograph captions and the nature of this text that determined the relations to be extracted as well as the prototype system we have built itself. Building the system that extracts relational facts and uses them for indexing and retrieval of crime scene photographs involves the creation of a semantic model of the domain, as well as a pipeline of processing resources that process the captions in sequence making use of these domain modelling resources. All these will be presented in detail along with the actual extraction module that relies on the performance of the preceding modules. The extraction rules will be described and the indexing triples that can be extracted will be illustrated with examples.

3.1 Corpus

Our crime scene caption corpus can be distinguished into two sets: one that we have used for developing our extraction patterns approach and one which will be used for testing the coverage of our method. The development corpus consists of 525 captions; part of it is a small 65-caption set from a single mock crime case obtained within a speech experiment. The experiment was conducted by the University of Surrey research team and it involved a staged murder scene, attended by a real Scene of Crime Officer from the Surrey Police. The officer was asked to document the scene using a digital camera and a digital speech recorder; instead of waiting for the photographs to be developed and then provide captions for them, the officer was able to take the shots and record a short description of what he photographed at the time he took each photograph. This instant captioning of the photographs is important for the officers since it captures the rational of taking a specific photograph directly. These captions were later transcribed manually - for avoiding (at least for the moment) automatic speech recognition and transcription problems.

The rest of the development corpus as well as the evaluation corpus (the latter consisting

Pastra et al. (2002) – In Proceedings of the British Computer Society Conference in Artificial Intelligence – pp. 121-134

of about 700 captions) was taken from 300 real cases processed at the South Yorkshire police, in the form of photo-indexes in electronic format. These photo-indexes form the first page of a photo-album that is always created for a crime case. These photo-index pages have a unique reference number along with other registration details such as the type of the crime (e.g. Sudden Death) plus the name of the victim, the date the photographs were taken and the name of the photographer. Sometimes it is not the SOCO who takes the photographs, but a specialist from the Forensic Photography department. Then, there is an enumeration of the photographs that follows their sequence in the photo-album, coupled with a caption for each photograph. Sometimes a caption is given to a range of photographs e.g :"1-3 views from the bedroom towards the lounge". So, the caption - photograph is not always an 1:1 relation.

Studying both these caption collections, one realises that a "full" caption is one that indicates the 'what' and 'where' in the photographs. The location in particular, seems to be very important in these captions. In some cases even the angle from which the photograph has been taken is provided. Usually, a trick with prepositions that denote location/direction is what distinguishes one caption from another. Consider for example the following captions: "View to loft" and "View into loft". It is the preposition that holds the whole meaning of the otherwise identical phrases and of course it is the clue for distinguishing the corresponding photographs. Furthermore, the captions not only have specific linguistic characteristics as individual formations, they also appear having "textual" properties as well. What we mean is that sequencial captions, behave sometimes as a text. Phenomena such as anaphora and co-reference are not rare. For example, we came across the following captions: "View in the direction of Crookes Road" and "View in the opposite direction". In this case, what 'opposite' refers to is dependent on the caption that precedes.

Analysing the characteristics of the language used in these captions, we have seen that these characteristics are in agreement with observations that have also been made for captions elsewhere (cf. [12]). Extensive ellipsis is a main characteristic of these captions. Functional words such as determiners, articles and pronouns are omitted; in the case of possessive pronouns, this is done not for brevity, but in order to avoid drawing conclusions. Ellipsis is not restricted to these parts of speech; sometimes verbs are missing from the sentences, resulting in phrases that stand alone as if they were complete sentences. Passive expressions are preferred. Prepositions,

Pastra et al. (2002) – In Proceedings of the British Computer Society Conference in Artificial Intelligence – pp. 121-134

adjectives and nouns are heavily used. Qualifiers such as adjectives and prepositional phrases are used for describing physical objects in the crime scene and adjuncts of place, time and manner seem to prevail. No complex structures are used; in fact, co-ordination is limited and usually substituted by punctuation, whereas sub-ordination, on its turn, is avoided. They are mainly declarative, descriptive statements with quite a few named entities, such as person names, location names and addresses, organization names and dates. Even the verbally produced captions, have these characteristics to a great extend. A few repair phenomena appear, which is never found in originally writen captions, but apart from that the need to produce concise, accurate and objective captions results in the officers producing captions in which the information content of every word is extremely high. Last, most captions contain meta-information for the photograph, that is not only descriptions of what is depicted but also information on the angle from which the photograph was taken, whether it is a distant shot or a close-up and other.

We have used the development caption corpus in order to determine the relations that we will allow our system to extract. These relations are described in (3.4) and our intention is to extend them with possibly more relations that will be indicated after running the extraction module on the large evaluation caption collection. The relations that we extract, not only denote what the photograph depicts but they also express image-related meta-information (cf. 3.4).

3.2 Concept hierarchy: OntoCrime

Apart from analysing our corpus and deciding on the relational facts to be extracted from the captions, the semantic enhancement of the argument terms of these relations had to be considered before any processing of the captions had taken place. A domain-specific ontology was thought to be the best way for both allowing our system to automatically expand the queries that would be submitted and to take advantage of semantic information in the relational fact extraction rules. We have build our domain ontology by priming with lists obtained from the Police Information Technology Organisation (PITO) [16], and in particular their Common Data Model.

In an attempt to standardise the wording used in all tasks that involve Police, PITO formed a team for the creation of a Common Data Model. This model contains words and phrases

Pastra et al. (2002) – In Proceedings of the British Computer Society Conference in Artificial Intelligence – pp. 121-134

clustered semantically and hierarchically. In order to deal with the complexity of real world domain modelling, we only extracted from the data model semantic categories of interest for Crime Investigation applications. OntoCrime is our concept hierarchy, implemented in the form of a direct acyclic graph (DAG) whose top node is "entity". OntoCrime has a number of *object*, *event* and *property* classes.

The *object hierarchy* consists of a disjunction of classes that denote tangible and intangible objects. Currently, we have sixteen object classes each with its own subclasses and subsubclasses down the word level . The *event hierarchy* contains classes that denote a state or an action such as "criminal events", "police actions", "spatial events" (e.g "surround") and "meta-information events" (e.g "show" as in "photograph showing X"). Last, the *property hierarchy* has a number of *functional* and *relational* properties/attributes that can be assigned to the object and event classes. Functional properties are the single-value ones e.g the 'age', whereas the 'colour' is a relational property since an object may have more than one colours. The functional properties are further distinguished into *mutable* and *immutable* ones, according to whether the value of the property can change at some point in time, or it is forever fixed e.g the 'age' is a mutable property, whereas the 'sex' is -normally- immutable. The leaf nodes of the ontology are lemmas themselves.

For implementing OntoCrime, we have adopted the XI Knowledge Representation Language [2], a formalism that allows the user to code and operate with symbolic knowledge. XI is compiled into Prolog, making it possible to mix procedural knowledge with the basic declarative formalism's constructs. We have chosen XI because it has been proved successful in other NLP appplications [6]. XI provides basic language constructs to specify hierarchical relations and as such OntoCrime. Individuals, classes of individuals, inclusion relations between classes of individuals and multiple inheritance hierarchies can be defined and attribute-values may be associated with classes or with individuals. In XI, classes are represented as unary predicates and individuals as atoms. An attribute or property is a binary predicate, the first argument of which is the class/individual the attribute is assigned to and the second being the value of this attribute. The value can be a fixed term or a term that became instantiated in appropriate situations when new knowledge is deduced during reasoning. We have implemented a visualization component that allows the user to easily explore the ontology.

Pastra et al. (2002) – In Proceedings of the British Computer Society Conference in Artificial Intelligence – pp. 121-134

3.3 The SOCIS Indexing and Retrieval Prototype

Our prototype has been developed using GATE components [1] enriched with full syntactic and semantic analysis implemented in Prolog. The system is composed of a simple tokeniser that identifies words and spaces, a sentence segmenter, and a named entity recognizer specially developed for crime scene applications (photographs usually mention persons and locations that need to be identified before full parsing) [7]. Part of speech tagging is done with an implementation of the "independence and commitment" approach to POS tagging [5], based on the Brill tagger. We have tunned the default lexicon and rule set produced by the learning step because of many errors found during development.

We also use a rule-based lemmatiser that produces an affix and root for each noun and verb in the input text. The lemmatiser program is implemented as a set of regular expressions specified in flex and translated into C code. We are using an implementation of the Bottom-up chart parsing described in [3], enriched with semantic rules that construct a naive semantic of each sentence in first order logical form. The parser is complete in the sense that every analysis licensed by the grammar is produced, though there is a mechanism to control this. On completion a "best parse" algorithm is run to select a single analysis of the sentence, which may be partial if no tree spanning the whole sentence can be constructed. We use a context-free phrasal grammar of English enriched with features and values; it consists of a sequence of sub-grammars for: noun phrases (NP), verb phrases (VP), prepositional phrases (PP), relative phrases(R) and sentences (S). The parser is quite accurate in the treatment of noun phrases. The semantic rules produce unary predicates for entities and events (e.g., *body(e1)*, *lay(e2)*, *floor(e3)*) and binary predicates for properties (e.g. *lsubj(e1,e2)*, *on(e2,e3)*). Constants (e.g., $e1$, $e2$, $e3$) are used to represent entity and event identifiers. The semantics produced by the parser is not enough to produce the propositions we need for indexing. An additional step of discourse interpretation is required in order to correctly produce semantic predicates and associated arguments.

Discourse interpretation is carried out by mapping the information produced by the parsing and semantic interpretation into an evolving Discourse Model of the input text using the ontology. The mapping is done using the dictionary and disambiguation rules. In our application the discourse interpreter is responsible for completing the partial semantic representation produced

Pastra et al. (2002) – In Proceedings of the British Computer Society Conference in Artificial Intelligence – pp. 121-134

by the parser. As entities and properties are added into the model, rules are checked and fired according to the local context. In cases where syntactic relations (either prepositional attachment or verb-argument relations) are correctly identified by the parser the processing consists of mapping the superficial form into its semantic representation. When the parse is partial, the ontology is used to search for evidence to attach the correct arguments to the semantic relations using properties of entities in the current sentence.

All these processing resources run in sequence over the photograph captions and each one dependes on the output of another. After the results of the discource interpreter have been obtained, our relational fact extractor is activated.

3.4 The SOCIS Extractor

The SOCIS extractor attempts to extract relational triples from each caption, that will be used for indexing the corresponding photograph. The module extracts as many relations as can be found in the caption, but it also infers relations that are not explicitly mentioned in the text. If no relations are found, only single keyterm extraction is performed. Each triple consists of a relation name and two arguments; the arguments are only entities that are classified as "objects" in the ontology. It is not necessary that these entities are recorded in OntoCrime; if they are not classified there, they are added under a general class according to their grammatical information (e.g under the 'object' class if they are nouns). The relation in the triple denotes the relation that holds between the two entities. Currently we extract the following relations:

- ABOVE: e.g. view of roof above seat... = view ABOVE seat
- AND: the grouping relation. It is mainly used for inferring other relations that hold for all the entities linked with the AND relation
- AROUND: e.g. body surrounded by blood = blood AROUND body
- BEHIND: e.g bottles BEHIND bar
- BETWEEN: e.g. photograph of deceased between vehicle and garage wall = deceased BETWEEN vehicle - garage wall
- DESTINATION: usually denoted with the preposition 'to' e.g entrance DEST public+house

Pastra et al. (2002) – In Proceedings of the British Computer Society Conference in Artificial Intelligence – pp. 121-134

- IN: for the literal meaning of 'in' (inside) e.g blood IN bathroom
- MADE-OF: e.g footwear impression in blood = footwear+impression MADE-OF blood
- NEAR: e.g body NEAR table (denoted via 'near', 'adjacent' etc)
- OF: only for cases when a part-of relation is denoted e.g rear OF machine
- ON: e.g table showing bottles = bottles ON table
- SOURCE: e.g rear garden from Lancing Street = rear garden SOURCE Lancing Street
- UNDER: e.g chair leg found underneath table = chair leg UNDER table
- WITH: shot of male showing open shirt = male WITH open+shirt
- WITHOUT: it captures negation/absence of something e.g table knife with no blood = table+knife WITHOUT blood

As can be seen from the above examples according to the extractor rules, relation extraction goes beyond the actual presence of prepositions in the captions. The "+" symbol denotes noun phrases. The arguments of the triples are not only single words; they can be noun phrases of any length. The images in these cases are indexed not only with triples containing the compound arguments but also with triples that have only the headword of the noun phrase as argument.

Apart from these relations, we also extract unary relations for capturing meta-information, such as:

- META-POSITION: e.g shot of bar with tables on the foreground = bar WITH tables, tables META-POSITION foreground.
- SOURCE-BEHIND: it denotes the viewpoint from which the photograph was taken e.g shot of floor from behind the bar = floor SOURCE-BEHIND bar

We have also identified some clue anaphora expressions that denote that triples extracted from the previous caption have to be used for the current caption too e.g: "same shot".

When only one relation is extracted from each caption, things are quite straight forward. However, most captions consist of more than one relations and in these cases logical relation rules are needed in order to extract the right triples. Assigning the appropriate arguments to

the triples and inferring more relations that are implicit is performed in our application with a finite set of rules. We have identified exception cases, in which our default rules need to behave differently; these exceptions have been dictated by our small caption corpus and their validity needs to be tested in our larger collection. In order to illustrate how these rules work, consider the following caption: "Photograph of writing in blood on wall". The relational facts to be extracted in this case, should be: writing MADE-OF blood[5], blood ON wall (and: blood ON wall). However, the two triples share an argument (i.e 'blood') and the extractor needs specific rules that will allow for this. Our default rule for instructing the extractor on cases when two relations need to be extracted could be expressed as follows:

when X REL1 Y REL2 Z then extract the triples: X REL1 Y, Y REL2 Z.

A special case of this rule concerns captions such as "footwear impression on rear of games machine". If we apply the default rule, we will get two triples one of which is not meaningful: footwear impression ON rear, rear OF games machine. The former triple has an underspecified second argument, which in fact denotes "part of" the real argument. We treat these cases with specific rules that make use of OntoCrime and the "part-denoted" entity category, instructing the system to use the main entity and not its part as the second argument of the triple. Therefore we extract the following relational facts: footwear impression ON games machine, rear OF games machine.

We have similar rules for captions with three and four relations. When AND relations have been identified in captions that contain more relations, then rules are also needed for infering relations that are linguistically implied. For example in "bottles and ashtray on table", three triples should be extracted: bottles AND ashtray, bottles ON table, ashtray ON table. In order to cope with such cases we have written a rule which determines that:

when [X AND Y] REL Z then extract: X AND Y, X REL Z, Y REL Z.

The appropriate rules have been writen to cover similar AND cases.

[5]We should note, that in the above example the MADE-OF relation is extracted because of a rule that says that when the preposition 'in' is followed by an entity classified in the ontology as a 'substance' then it denotes that something is made-of this substance.

4 Issues to be explored - Benefits

The SOCIS project is due to finish in the end of 2002 and therefore our application prototype has reached the phase of extensive evaluation and testing. The language and processing resources for the application have all been developed as described in the previous sections and what now remains is to expand the coverage of our relation extraction rules. Evaluating the SOCIS extractor module by running it on a large evaluation caption corpus (cf. 3.1) will help us enrich and refine our rules. What is more important is that both our development and evaluation data is real data, captions produced by Scene of Crime Officers themselves, mostly for real cases. This means that the real needs and problems for indexing and retrieving captions for crime imagery are not only taken into consideration while developing the application, but they also determine the course of our work.

This testing and expansion procedure that is currently under way will be followed by a "real time" test of the application. The SOCIS prototype will be given to our police advisory board for testing from their own setting, in their every day working environment. This will be a proper usability evaluation phase for our application; the system will be accessed through the web, and police officers will use a very simple interface for querying a database with case records[6]. The officers will be able to search for images by providing image registration data (e.g name of photographer, case id, etc.), meta-information data (e.g camera position from which a photograph was taken, the visual focus of the photograph or its background etc), natural language queries enhanced with semantic relations and/or combinations of these. From our field work for the system and discussions with our advisory board, we expect that being able to use semantic relations similar to the ones recorded in the caption text will be of great help to the police, since it is these relations that express the important facts for the domain. If this proves to be so, our prototype is expected to perform very well and assist in retrieving the images of interest to the crime investigators. In any case, testing the system with the people that it is intended for, will indicate its weaknesses and strenghts. Our application does not require any particular training on the part of the police officers, apart from some basic familiarity with information technology systems.

By completion of the SOCIS project, a research prototype for a real world application

[6]The interface has already been implemented by the University of Surrey research team and web access to the full SOCIS system will be available shortly at `http://www.dcs.shef.ac.uk/research/group/nlp/socis`.

tested by potential users will be available. Our collaboration with the interested parties for using such an application (i.e U.K. police forces) has indicated that such a system will be more than welcome for the police, since it will assist greatly throughout the "life-cycle" of a crime investigation and beyond this. Extracting relational facts from captions provided by the police officers themselves results in indexing images in a "meaningful" - especially for the officers- way, that emphasises the information they are interested in and they are more likely to search for. From the attendance of crime scenes by a SOCO to taking a case at court and going through very old cases, "intelligent" indexing and retrieval of images is an application very much needed from the police, that if taken further will end up in a complete -rather than prototype- system ready for use with very low maintenance costs and great usability.

Apart from the obvious benefits to all the crime investigation related organisations, our text-based image indexing and retrieval application will contribute to the "Intelligent Image Retrieval" research field. Our approach of automatically extracting relational facts for image caption retrieval is tested in the crime investigation domain and needs to be put on the test in order to judge its coverage and adaptability to other domains. Going through related literature for image caption retrieval in specific domains (cf.2.1) we have seen that the relations expressed in other domain captions [12] are quite similar to the ones we encode in our approach and using these relational facts for indexing and retrieval of the corresponding images may result in overcoming problems that cannot be dealt with other methods. In the more general IR research, our work would be expected to be only complementary to other approaches and at least indicate the need for deeper semantic representations in indexing and retrieval.

4.1 Aknowledgements

We would like to thank the Surrey-SOCIS research team for all the extensive and fruitful discussions on the project and the work they have done for building the SOCIS system: Khursid Ahmad, Bodgan Vrusias, Mariam Tariq and Chris Handy. We would also like to express our gratitude to our police-advisory group and in particular to Mr Andrew Hawley, a Scene of Crime Officer at the Rotherham Police Station, South Yorkshire, for all his help in providing as with information on the crime scene documentation practices and in collecting all our caption corpus.

References

[1] H. Cunningham, D. Maynard, K. Bontcheva, and V. Tablan. 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*, 2002.

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

[3] G. Gazdar and C. Mellish. *Natural Language Processing in Prolog.* Addison-Wesley, Reading, MA, 1989.

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

[5] Mark Hepple. Independence and commitment: Assumptions for rapid training and execution of rule-based POS taggers. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics (ACL-2000)*, Hong Kong, October 2000.

[6] K. Humphreys, R. Gaizauskas, S. Azzam, C. Huyck, B. Mitchell, H. Cunningham, and Y. Wilks. Description of the LaSIE system as used for MUC-7. In *Proceedings of the Seventh Message Understanding Conference (MUC-7).* `http://www.itl.nist.gov/iaui/894.02/-related_projects/muc/index.html`, 1998.

[7] K. Pastra, H. Saggion, and Y. Wilks. Socis: Scene of crime information system. Technical Report CS-01-19, University of Sheffield, 2001.

[8] St. Richardson, W. Dollan, and L. Vanderwende. Mindnet: acquiring and structuring semantic information from text. In *Proceedings of COLING*, 1998.

[9] E. Riloff. Little words can make a big difference for text classification. In *Proceedings of the 18th ACM SIGIR Conference*, pages 130–136, 1995.

[10] E. Riloff and J. Lorenzen. Extraction-based text categorization:generating domain-specific role relationships automatically. *Natural Language Information Retrieval*, 1999.

[11] T. Rose, D. Elworthy, A. Kotcheff, A. Clare, and P. Tsonis. Anvil: a system for the retrieval of captioned images using nlp techniques. In *Proceedings of CIR2000, 3rd UK Conference in Image Retrieval*, 2000.

[12] N. Rowe. Precise and efficient retrieval of captioned images: the marie project. *Library Trends*, 1999.

[13] C. Sable and V. Hatzivassiloglou. Text-based approaches for the categorization of images. In *Proceedings of ECDL*, 1999.

[14] A. Smeaton and I. Quigley. Experiments on using semantic distances between words in image caption retrieval. In *Proceedings of the 19th International Conference on Research and Development in Information Retrieval*, 1996.

[15] A. Smeaton and C. van Rijsbergen. Experiments in incorporating syntactic processing of user queries into a document retrieval strategy. In *Proceedings of the 11th ACM SIGIR Conference*, 1988.

[16] Police Information Technology Organisation Data Standards Team. Common data model v8.1. CD Version 1.1 - PITO, 1999.

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

[18] E. Voorhees. Using wordnet to disambiguate word senses for text retrieval. In *Proceedings of the 16th ACM SIGIR Conference*, pages 171–180, 1993.