Event Recognition using Information Extraction Techniques

Maria Vargas-Vera and David Celjuska
Knowledge Media Institute (KMi),
The Open University,
Walton Hall, Milton Keynes, MK7 6AA, United Kingdom
m.vargas-vera@open.ac.uk
Technical University of Kosice, Letna 9/A, 04001 Kosice, Slovakia
celjuska@neuron.tuke.sk

Abstract

This paper describes a system which recognizes events on stories. Our system classifies stories and populates a KMi Planet ontology with new instances of classes defined in it. Currently, the system recognizes events which can be classified as belonging to a single category and it also recognizes overlapping events (more than one event is recognized in the story). In each case, the system provides a confidence value associated to the suggested classification. In our event recognition system we use Information Extraction and Machine Learning technologies. We have tested this system using an archive of stories describing the academic life of our institution (these stories describe events such as an project award, publications, visits, etc.)

1 Introduction

In this paper we focus on the problem of automatically classifying documents. This is an interesting problem in Natural Language research and it has many potential applications ranging from document summarization to the semantic web. There are several approaches to text classification. In this paper we describe an approach to stories classification based on Information Extraction and Machine Learning technologies. Essentially, information extraction can be seen as the task of pulling predefined relations from texts. Efforts have been made to apply information extraction to several domains, for instance, scientific articles such as MEDLINE [Craven and Kumlien, 1999], bibliographic notices [Proux and Chenevoy, 1997] and medical records [Soderland *et al.*, 1995]. In designing an information extraction system for the KMi organization, the system should be able to extract the name of KMi projects, KMi funding organizations, awards, dates, etc, and ignore anything not clearly relevant to these pre-specified categories. Ontologies can be used in information extraction systems to help them extract relations from semi- or unstructured documents, statements or terms [Roux *et al.*, 2000]. Also, recent work on semi-automatic ontology acquisition by means of information extraction, supported by machine-learning methods, is described in [Maedche and Staab, 2000; Kietz *et al.*, 2000; Vargas-Vera *et al.*, 2001a; 2001b]. On similar lines, there is CMU's approach for extracting information from hypertext using machine learning techniques and making use of an ontology [Craven *et al.*, 1999].

Our system, as most information extraction systems, uses some form of partial parsing to recognize syntactic constructs without generating a complete parse tree for each sentence. Such partial parsing has the advantages of greater speed and robustness. High speed is necessary to apply the information extraction to a large set of documents. The robustness achieved by allowing useful work to be done from a partial parsing is essential to deal with unstructured and informal texts.

The main contribution of our paper can be summarized as follows:

- identification of events on stories by means of Information Extraction and Machine Learning technology and
- semi-automatic population of a selected ontology.

The paper is organized as follows: Section 2 shows the event topology used in our event recognition system. Section 3 describes the classification of stories. Section 4 presents the process model in our system. Section 5 presents assignation of confidence values to the rules extracted using Crystal. Section 6 describes semi-automatic population of a selected ontology with new instances of classes already defined in our selected ontology. Section 7 presents a working example. Finally, Section 8 gives conclusions and directions for future work.

2 Event topology

The KMi domain consists of events or activities happening in our Institute. Events are defined formally in our ontology as classes. Currently, in the KMi ontology we have defined 41 different types of events. As the event topology is already defined in the ontology. Then, for each event we already have defined the slots which might be instantiated by an information extraction component. Figure 1 shows a portion of the hierarchy of events as defined in the ontology.

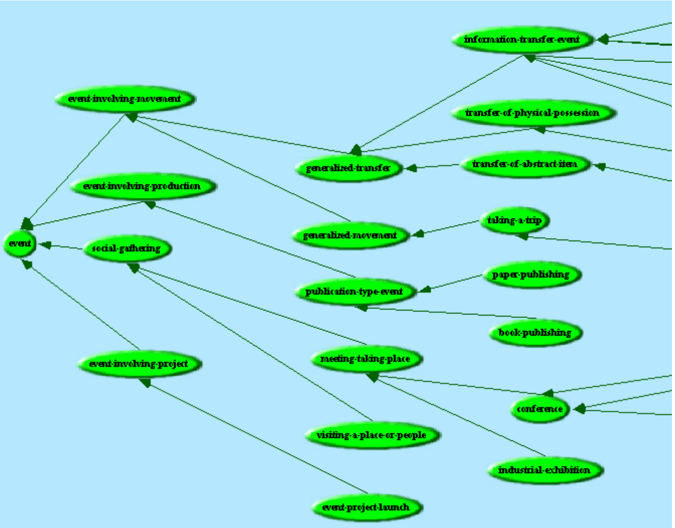


Figure 1: Event hierarchy

For the sake of space, we only present the structure of three type of events from the event hierarchy: visiting-a-place-or-people, project-award and academic-conference.

```
Class Event 1: visiting-a-place-or-people
 slots:
   visitor (list of person(s))
   people-or-organisation-being-visited
     (list of person(s) or organization)
   has-duration (duration)
   start-time (time-point)
   end-time (time-point)
   has-location (a place)
   other agents-involved (list of person(s))
   main-agent (list of person(s))
```

The structure of Event 1 (visiting-a-place-or-people) describes a set of objects which might be encountered in story describing an event visit, such as visitor, people-or-organisation-being-visited, other agents-involved, etc.

```
Class Event 2: project-award
 slots:
  has-awarding-body (an organization)
  has-award-rationale (project goals)
  object-acted-on (award)
  recipient-agents (the agents which possibly
                    receive the object-acted-on)
  start-time (time-point)
  end-time (time-point)
  location-at-start (a place)
  location-at-end (a place)
```

The structure of Event 2 describes a set of objects which might be encountered in story describing an event "project-award" such as organization, award, recipient agent, etc.

```
Class event 3: academic-conference
 slots:
   has-papers (list of papers)
   has-invited-talks (list of talk(s))
   has-demostrations (list of demonstration(s))
   has-duration (duration)
   start-time (time-point)
   end-time (time-point)
   has-location (a place)
   main-agent (list of person(s))
   other agents-involved (list of person(s))
   meeting-attendees (list of person(s))
   meeting-organizer (list of person(s))
```

Event 3 contains the structure for the event “academic-conference”. Entities that need to be recognized are papers, location, etc.

3 Classification of stories

We classify stories or documents as belonging to any of the types of events according with the objects that are found in them. For each event type we have a predefined objects that should be found in the story. For instance, for the event “visiting-a-place-or-people” the system might encounter objects of type: visitor, place and date.

In our system classification is performed in the following steps:

- pre-process the story
- find the objects in the story using partial parsing
- provide classification of a story with associated confidence value

Each event in our system has several patterns which can be used to recognize it. For instance, in case of "visiting-place-or-people" event the following patterns were encountered.

```
- X visited Y
- X visits Y
- Y visited by X
- Y visited by X at Z
- Y were visited by X
- Visit to Y
- X came
- X came to visit Y
- Y hosted X
- Y hosted a visit from X
- Y had a visit from X
- Y welcomes X
```

In all patterns shown above X is a person, Y is a place/institution and Z is a location.

Problems might occur when more than one event can be recognized in a story. Then our system decides to classify the story according with the following criteria:

it computes the confidence value as the number of slots the system was able to extract divided by total number of slots that an annotator/expert used during annotation process on any story from a given class.

$$confidence = \frac{number\ \ of\ \ items\ \ extracted}{number\ \ of\ \ slots\ \ used\ \ by\ \ annotator}$$

Then, the category which maximizes the sum of the filled slots is placed at the top of the window (i.e. the classification with the maximum confidence value). If none of the templates are able to be filled (during the extraction phase) then the story is given the status of unclassified story. The user will be presented with classification, associated confidence value and extracted objects. Once that the user agrees (rejects) one (all) of the suggested classification and extracted information, the ontology is updated with a new instance.

4 Process model

Within this work we have focused on creating a generic process model for event recognition on stories. In our system we have devised three activities: mark-up, learning and extraction. We will provide more details of each of the activities in turn.

Mark-up

The activity of semantic tagging refers to the activity of annotating text documents (written in plain ASCII or HTML) with a set of tags defined in the ontology. Our classification system provides means to browse the event hierarchy. In this hierarchy each event is a class and the annotation component extracts the set of possible tags from the slots defined in the ontology. During the markup phase as the text is selected the system inserts the relevant SGML tags into the document. Also our system offers the possibility of removing tags from a document.

Learning

This phase was implemented by integrating two tools: Marmot and a learning component called Crystal, both from UMass (full description can be found in [Riloff, 1996]). Marmot is a natural language preprocessing tool that accepts ASCII files and produces an intermediate level of text analysis that is useful for information extraction applications. Sentences are separated and segmented into noun phrases, verb phrases prepositional phrases. Marmot has several functionalities: preprocesses abbreviations to guide sentence segmentation, resolves sentences boundaries, identifies parenthetical expressions, recognizes entries from a phrasal lexicon and replace them, recognizes dates and duration phrases, performs phrasal bracketing of noun, preposition and adverbial phrases, finally scopes conjunctions and disjunctions.

Crystal is a dictionary induction tool. It derives a dictionary of concept nodes from a training corpus. The first step in dictionary creation is the annotation of a set of training texts by a domain expert. Each phrase that contains information to be extracted is tagged (with SGML style tags).

Crystal initializes a concept nodes dictionary for each positive instance of each type of event. The initial concept node definitions are designed to extract the relevant phrases in the training instance that creates them but are too specific to apply to a unseen sentences. The main task of Crystal is to gradually relax the constraints on the initial definitions and also to merge similar definitions. Crystal finds generalizations of its initial concept node definitions by comparing definitions that are similar. This similarity is deduced by counting the number of relaxations required to unify two concept node definitions. Then a new definition is created with constraints relaxed. Finally the new definition is tested against the training corpus to insure that it does not extract phrases that were not marked with the original two definitions. This means that Crystal takes similar instances and generalizes into a more general rule by preserving the properties from each of the concept node definitions which are generalized.

The inductive concept learning in Crystal is similar to the inductive learning algorithm described in [Mitchell, 1982] a specific-to-general data-driven search to find the most specific generalization that covers all positive instances. Crystal finds the most specific generalization that covers all positive instances but uses a greedy unification of similar instances rather than breadth-first search.

Extraction

A third component called Badger (from UMass) was also integrated into our event recognition system. Badger makes the instantiation of templates. The main task of badger is to take each sentence in the text and see if it matches any of our concept node definitions. If no extraction concept node definition applies to a sentence, then no information will be extracted; this means that irrelevant text can be processed very quickly.

It might occurs that Badger obtains more than one type of event for an story[1]. Then our information extraction system decides to classify the story according with the criteria defined in section 3.

5 Confidence values associated to the extracted rules

In the automatic construction of ontologies precision is more important than recall. Reported work in [Vargas-Vera *et al.*, 2001a] has shown that in events that are not "visiting-a-place-or-people", precision was lower than 95% using the KMi domain as testbed. Therefore, our goal is to increase precision. Currently we are focusing, in associating a confidence value to the Crystal induced rules in order to increase precision. The confidence number for each rule can be computed by a three-fold cross-validation methodology on the training set. According to this methodology, the training set is split into three

[1] A first implementation of our event recognition system which only recognizes single events is described in [Vargas-Vera *et al.*, 2001b]

equally sized subsets and the learning algorithm is run three times. Each time two of the three pieces are used for training and the third is kept as unseen data (test set) for the evaluation of the induced rules. The final result is the average over the three runs. At run time each instance extracted by Crystal will be assigned the precision value of that rule. The main feature of using confidence values is that among ambiguous instantiations, we can still choose the one with the highest estimated confidence.

6 Populating the ontology

Building domain-specific ontologies often requires time-consuming expensive manual construction. Therefore, we envisage information extraction as a technology that might help us during ontology maintenance process. During the population step our information extraction system has to fill predefined slots associated with each event, as already defined the ontology. Our goal is to automatically fill as many slots as possible. However, some of the slots will probably still require manual intervention. There are several reasons for this problem:

- there is information that is not stated in the story,
- none of our patterns match with the sentence that might provide the information (incomplete library of patterns)

The extracted information could be validated using the ontology. This is possible because each slot of each class of the ontology has a type associated with it. Therefore, extracted information which does not match the type definition of the slot in the ontology can be highlighted as incorrect.

7 Example

In this section we present an example. The domain of our example is a web based news letter, KMi Planet [J.Domingue and Scott, 2000] that has been running in our institution for five years. One of the functionalities offered in KMi Planet is an editor which can be used to manually extract information and classify a story. Then this information is sent to ontology server in order to create a new instance. Whilst we are happy with the functionalities offered by KMi planet, we want to automatically extract and populate the handcrafted ontology. This is because maintenance is time consuming and error prone. Therefore, in the first instance we have selected KMi Planet (which contains 200 stories of academic life in KMi) as our first domain for our event recognition system.

In this section we show two examples in the first story only one single event was recognized and the second example two events were recognized in a story. Figure 2 shows an story classified as "project-award" with a confidence of 75% whilst in Figure 3 our system suggest two classifications for the story "visiting-a-place-or-people" and "project-award" with confidence values of 33% and 25% respectively.

In Figure 2 we can see in the top right window that the user has selected the project-award class from the KMi ontology. Also in the same figure the number 3/4 means that three out of four slots values were able to be extracted from story. The number in brackets is the confidence value associated to the suggested classification. For this particular story titled "The AKT begins" our system was able to extract the object-acted-on (thing that was given as an award), recipient-agent (man/organization/project that was given an award) and has-awarding-body (man/organization that gave award).

In Figure 3 our system was able to extract visitor for classification "visiting-a-place-or-people" and recipient-agent for classification project-award. The reason why this story is classified as project-award is that during the learning phase the system learn a rule:

```
Cn-type project-award       ID:18
    Status: GENERALIZED

Constraints:
VB::
    mode active
    root: bring
    terms: BRINGS
    mod terms:  <null>
    head terms:BRINGS
    classes:     ws_Root_Class
    mod class:   ws_Root_Class
PP::        ==> recipient-agents
    terms: TO
    mod terms: TO
    classes:     ws_Root_Class
    mod class:   ws_Root_Class
    head class:  ws_Root_Class
```

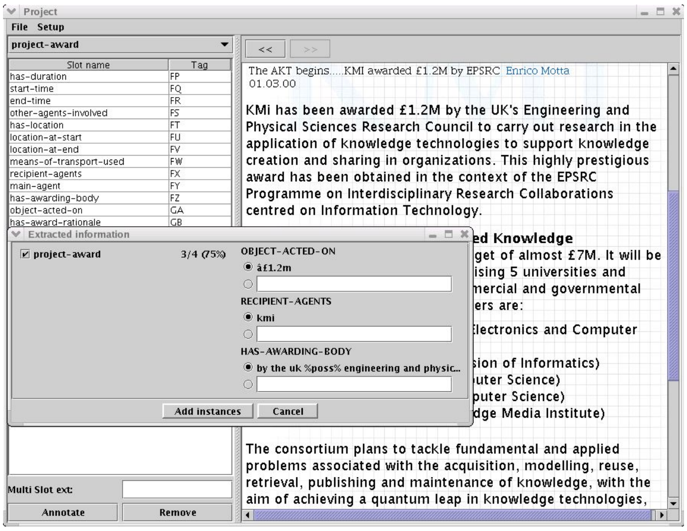
As we can see this rule is very general and the system from any sentence with verb=BRING and preposition=TO extracts Prepositional phrase(PP) as recipient agent. This problem could be solved by constraining slots to the correct type given in an ontology.

Finally, after information is extracted the classification and extracted values are presented to the user. If the user accept these values then the ontology will be updated with a new instance.

8 Conclusions and future work

We have built a tool which recognizes events on stories and extracts knowledge using an ontology. Currently, our system have been trained using the archive of 200 stories that we have collected in KMi. [2] The training step was performed using typical examples of stories belonging to each of the different type events defined in the ontology. Our system recognizes single events and overlapping events. Then it is able to suggest possible classification for a story. Currently, the population of the selected ontology is performed at the level of instances.

[2] URL:http://kmi.open.ac.uk/planet/



Slot name	Tag
has-duration	FP
start-time	FQ
end-time	FR
other-agents-involved	FS
has-location	FT
location-at-start	FU
location-at-end	FV
means-of-transport-used	FW
recipient-agents	FX
main-agent	FY
has-awarding-body	FZ
object-acted-on	GA
has-award-rationale	GB

☑ project-award 3/4 (75%)

OBJECT-ACTED-ON
● â£1.2m
○

RECIPIENT-AGENTS
● kmi
○

HAS-AWARDING-BODY
● by the uk %poss% engineering and physic...
○

The AKT begins.....KMI awarded £1.2M by EPSRC Enrico Motta
01.03.00

KMi has been awarded £1.2M by the UK's Engineering and Physical Sciences Research Council to carry out research in the application of knowledge technologies to support knowledge creation and sharing in organizations. This highly prestigious award has been obtained in the context of the EPSRC Programme on Interdisciplinary Research Collaborations centred on Information Technology.

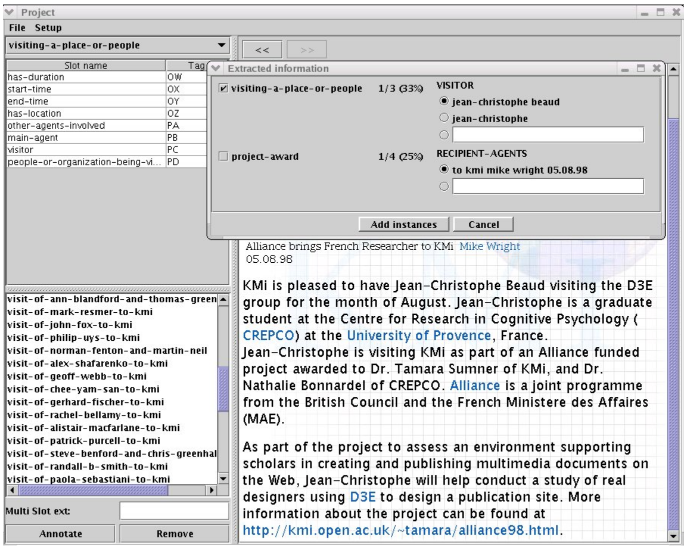
ed Knowledge
get of almost £7M. It will be
ising 5 universities and
mercial and governmental
ers are:

lectronics and Computer

sion of Informatics)
uter Science)
puter Science)
dge Media Institute)

The consortium plans to tackle fundamental and applied problems associated with the acquisition, modelling, reuse, retrieval, publishing and maintenance of knowledge, with the aim of achieving a quantum leap in knowledge technologies,

Figure 2: One suggested classification for story titled "The AKT begins"



Slot name	Tag
has-duration	OW
start-time	OX
end-time	OY
has-location	OZ
other-agents-involved	PA
main-agent	PB
visitor	PC
people-or-organization-being-vi...	PD

visit-of-ann-blandford-and-thomas-green
visit-of-mark-resmer-to-kmi
visit-of-john-fox-to-kmi
visit-of-philip-uys-to-kmi
visit-of-norman-fenton-and-martin-neil
visit-of-alex-shafarenko-to-kmi
visit-of-geoff-webb-to-kmi
visit-of-chee-yam-san-to-kmi
visit-of-gerhard-fischer-to-kmi
visit-of-rachel-bellamy-to-kmi
visit-of-alistair-macfarlane-to-kmi
visit-of-patrick-purcell-to-kmi
visit-of-steve-benford-and-chris-greenhal
visit-of-randall-b-smith-to-kmi
visit-of-paola-sebastiani-to-kmi

☑ visiting-a-place-or-people 1/3 (33%)

VISITOR
● jean-christophe beaud
○ jean-christophe
○

□ project-award 1/4 (25%)

RECIPIENT-AGENTS
● to kmi mike wright 05.08.98
○

Alliance brings French Researcher to KMi Mike Wright
05.08.98

KMi is pleased to have Jean–Christophe Beaud visiting the D3E group for the month of August. Jean–Christophe is a graduate student at the Centre for Research in Cognitive Psychology (CREPCO) at the University of Provence, France.
Jean–Christophe is visiting KMi as part of an Alliance funded project awarded to Dr. Tamara Sumner of KMi, and Dr. Nathalie Bonnardel of CREPCO. Alliance is a joint programme from the British Council and the French Ministere des Affaires (MAE).

As part of the project to assess an environment supporting scholars in creating and publishing multimedia documents on the Web, Jean–Christophe will help conduct a study of real designers using D3E to design a publication site. More information about the project can be found at http://kmi.open.ac.uk/~tamara/alliance98.html.

Figure 3: Two suggested classifications for a story titled "Alliance brings French researcher to KMi"

Our system extracts instances of classes defined in the event ontology. However, in future we will explore the possibility of use the extracted information with concep-tool [Compatangelo and Meisel, To appear 2003] in order to create new classes in a selected ontology. This will allow us to refine our ontology with a finer granularity.

As a medium term goal, we plan to link our event recognition system with KMi research profiles. Then we will filter stories and send them by e-mail to the people who might be interested in reading them. In this way, KMi researchers will be informed about events in our institution without having to browse the archive of news.

Currently, our event recognition system works with the KMi Planet ontology. But, in future, we plan to offer a selection of ontologies.

Acknowledgments

This work was funded by the Advanced Knowledge Technologies (AKT) Interdisciplinary Research Collaboration (IRC), which is sponsored by the UK Engineering and Physical Sciences Research Council under grant number GR/N157764/01. The AKT IRC comprises the Universities of Aberdeen, Edinburgh, Sheffield, Southampton and the Open University. The authors would like to thank Mark Gaved for his invaluable help in reviewing the first draft of this paper.

References

[Compatangelo and Meisel, To appear 2003] E. Compatangelo and H. Meisel. Reasonable support to knowledge sharing through schema analysis and articulation. *Journal of Engineering Intelligent Systems*, To appear 2003.

[Craven and Kumlien, 1999] M. Craven and J. Kumlien. Constructing Biological Knowledge Bases by Extracting Information from Text Sources. In *Proceedings of The 7th International Conference on Intelligent Systems for Molecular Biology (ISMB-99)*, 1999.

[Craven *et al.*, 1999] M. Craven, D. DiPasquo, D. Freitag, A. McCallum, K. Nigam T. Mitchell, and S. Slattery. Learning to Construct Knowledge Bases from the World Wide Web. *Artificial Intelligence*, 1999.

[J.Domingue and Scott, 2000] J.Domingue and P. Scott. Kmi planet: A Web Based News Server. In *In proceedings of the Asia Pacific Computer Human Interaction Conference (APCHI-98)*, 2000.

[Kietz *et al.*, 2000] J-U. Kietz, A. Maedche, and R. Volz. A method for semi-automatic ontology acquisition from a corporate intranet. In *Proceedings of the EKAW'00 Workshop on Ontologies and Text, Juan-Les-Pins, France*, oct 2000.

[Maedche and Staab, 2000] A. Maedche and S. Staab. Semi-automatic engineering of ontologies from texts. In *Proceedings of the 12th International Conference on Software Engineering and Knowledge Engineering, SEKE2000, Chicago, IL, USA*, pages 231–239, july 2000.

[Mitchell, 1982] T. Mitchell. Generalization as search. *Artificial Intelligence*, 18:203–226, 1982.

[Proux and Chenevoy, 1997] D. Proux and Y. Chenevoy. Natural Language Processing for Book Storage: Automatic Extraction of Information from Bibliographic Notices. In *Proceedings of The Natural Language Processing Pacific Rim Symposium (NLPRS'97)*, pages 229–234, 1997.

[Riloff, 1996] E. Riloff. An Empirical Study of Automated Dictionary Construction for Information Extraction in Three Domains. *AI Journal*, 85:101–134, 1996.

[Roux *et al.*, 2000] C. Roux, D. Proux, F. Rechenmann, and L. Julliard. An Ontology Enrichment Method for a Pragmatic Information Extraction System gathering Data on Genetic Interactions. In *Proceedings of The 14th European Conference on Artificial Intelligence (Workshop on Ontology Learning ECAI-2000)*, 2000.

[Soderland *et al.*, 1995] S. Soderland, D. Aronow, D. Fisher, J. Aseltine, and W. Lehnert. Machine Learning of Text Analysis Rules for Clinical Records. Tr 39, Center for Intelligent Information Retrieval, 1995.

[Vargas-Vera *et al.*, 2001a] M. Vargas-Vera, J. Domingue, Y. Kalfoglou, E. Motta, and S. Buckingham Shum. Template-driven information extraction for populating ontologies. In *In proceedings of the Workshop Ontology Learning IJCAI-2001*, 2001.

[Vargas-Vera *et al.*, 2001b] M. Vargas-Vera, J.Domingue, E. Motta, S. Buckingham Shum, and M. Lanzoni. Knowledge extraction by using an ontology-based annotation tool. In *In proceedings of the Workshop Knowledge Markup & Semantic Annotation, held in conjuction with the First International Conference on Knowledge Capture (K-CAP 2001), Victoria Canada*, pages 5–12, 2001.