Indexing and Querying of Narrative Documents, a Knowledge Representation Approach

Abstract: We describe here NKRL (Narrative Knowledge Representation Language), a semantic formalism for taking into account the characteristics of narrative multimedia documents. In these documents, the information content consists in the description of ‘events’ that relate the real or intended behaviour of some ‘actors’ (characters, personages, etc.). Narrative documents of an economic interest correspond to news stories, corporate documents, normative and legal texts, intelligence messages, representation of patient’s medical records, etc. NKRL is characterised by the use of several knowledge representation principles and several high-level inference tools.

1. Introduction

Narrative documents, or ‘narratives’, are multimedia documents that describe the actual (or intended) state or behaviour of some ‘actors’ (or ‘characters’, ‘personages’ etc.). These try to attain a specific result, experience particular situations, manipulate some (concrete or abstract) materials, communicate with other people, send or receive messages, buy, sell, etc.

Leaving fiction documents aside, we can note that:

• A considerable amount of the natural language information that is relevant from an economic point of view deals with narratives. This is true for all the different sorts of news story documents, for most of corporate information, for the intelligence messages, medical records, notarised deeds, legal documents, etc.

• In narrative documents, the actors or personages are not necessarily human beings. We can have narrative documents concerning, e.g., the vicissitudes in the journey of a nuclear submarine (the ‘actor’, ‘subject’ or ‘character’) or the various avatars in the life of a commercial product.

• It is not even necessary that the narrative situations to deal with be recorded in natural language (NL) documents: they can also be represented, e.g., as Web images, video or digital audio documents.

We introduce here NKRL, the “Narrative Knowledge Representation Language” (Zarri, 1997; 1998), a language expressly designed for representing the conceptual gist of economically relevant narratives. NKRL has been used as ‘the’ modelling language for narratives in several European projects; see, e.g., the recent Concerto (Esprit P29159) and Euforbia (IAP P26505) projects.

2. The main knowledge representation tools used in NKRL

Each ‘component’ of the set of knowledge engineering representation tools of NKRL takes into account a specific aspect of the global narrative domain.

The “definitional component” supplies tools for representing the ‘concepts’, intended here as a formal representation of the ‘important notions’ of a given application domain. A concept is rendered, substantially, as a frame-like data
structure associated with a symbolic label like \textit{human\_being}, \textit{taxi\_}, \textit{city\_}, \textit{chair\_}, \textit{gold\_}, etc. Concepts are inserted into a generalisation/specialisation hierarchy that is called \textsc{H\_CLASS}(es), and which corresponds well to the usual ontologies of terms; see (Zarri, 1997) for additional details.

The “enumerative component” of \textsc{NKRL} concerns the formal representation, as instantiated frames, of the concrete realisations (\textit{lucy\_}, \textit{taxi\_53}, \textit{paris\_}) of the concepts inserted in the \textsc{H\_CLASS} ontology. In \textsc{NKRL}, these instances take the name of \textit{individuals}. Throughout this paper, we will use the italic type style to represent a \textit{concept}, the roman style to represent an \textit{individual}.

The “descriptive component” concerns the tools used to produce the formal representations (called “templates”) of \textit{general classes of narrative events}, like “moving a generic object”, “formulate a need”, “be present somewhere”, etc. In contrast to the traditional \textit{ternary} (name-attribute-value) frame-like structures used for concepts and individuals, see \textsc{H\_CLASS}, templates are characterised by a \textit{quaternary} format connecting together, essentially, the \textit{symbolic name} of the template, a \textit{predicate} and several \textit{arguments} of the predicate. These last are, in turn, differentiated through the use of a set of named relations, the \textit{roles}. If we denote with $L_t$ the generic symbolic label identifying a given template, with $P_j$ the predicate, with $R_i$ the generic role and with $a_k$ the corresponding argument, the template data structures have then the following format:

$$\left(L_t, \left(P_j (R_1, a_1) \ (R_2, a_2) \ldots (R_n, a_n)\right)\right).$$

The instances (called “predicative occurrences”) of the templates, i.e., the representation of specific elementary events like “Tomorrow, I will move the wardrobe” or “Lucy was looking for a taxi” are in the domain of the last component, the factual one.

3. Using templates and occurrences to represent concrete events

To represent a narrative like “British Telecom will offer its customers a pay-as-you-go (payg) Internet service in autumn 1998”, we must select firstly the template (descriptive component) corresponding to “supply a service to someone”, see the upper part of Table 1. This template is a specialisation of the particular \textsc{MOVE} template of \textsc{H\_TEMP} corresponding to “transfer of resources to someone”. In a template, the arguments of the predicate (the $a_i$ terms in (1)) are represented by variables with associated constraints — that are expressed as concepts or combinations of concepts, i.e., using the terms of the \textsc{H\_CLASS} hierarchy. The constituents (as \textit{SOURCE} in Table 1) included in square brackets are optional.

When deriving a predicative occurrence, like $c_1$ in Table 1, the role fillers in this occurrence must conform to the constraints of the father-template. For example, in occurrence $c_1$, \textit{british\_telecom} is an individual instance of the concept \textit{company\_} that is, in turn, a specialisation of \textit{human\_being\_or\_social\_body}; etc. The meaning of the expression “\textit{BENF (SPECIF customer\_\_british\_telecom)}” in $c_1$ is self-evident: the beneficiaries (role \textit{BENF}) of the service are the customers of — \textit{SPECIF(ication)} — British Telecom.

The (about 200) templates that make up actually the \textsc{H\_TEMP} hierarchy —
the 'catalogue' of NKRL templates — are permanent and fully defined. We can say that these templates are part and parcel of the definition of the language; an (extremely simplified) image of H_TEMP is given in Figure 1. This approach is particularly advantageous for practical applications because it implies that: i) a system-builder does not have to create himself the structures needed to describe the events proper to a (sufficiently) large class of narratives; ii) it becomes easier to secure the reproduction or the sharing of previous results.

---

**name**: Move:TransferOfService  
**father**: Move:TransferOfResource  
**position**: 4.231  
**NL description**: 'Transfer or Supply a Service to Someone'

```
OWN SUBJ var1: [var2]  
OBJ var3  
    [SOURCE var4: [var5]]  
    BENF var6: [var7]  
    [MODAL var8]  
    [TOPIC var9]  
    [CONTEXT var10]  
    {[modulators]}

var1 = <human_being_or_social_body>  
var3 = <service>  
var4 = <human_being_or_social_body>  
var6 = <human_being_or_social_body>  
var8 = <action_name>  
var9 = <sortal_concept>  
var10 = <event>  
var2, var3, var7 = <physical_location>
```

c1) **MOVE SUBJ** british_telecom  
**OBJ** payg_internet_service  
**BENF** (SPECIF customer_british_telecom)  
**date-1**: after-l-september-1998  
**date-2**:  

---

Table 1. Deriving a predicative occurrence from a template.

The basic NKRL tools are enhanced by the use of two additional mechanisms:
- the AECS 'sub-language' that allows the construction of complex (structured) predicate arguments called 'expansions';
- the second order tools (binding structures and completive construction) used to code the 'connectivity phenomena' (logico-semantic links) that, in a narrative situation, can exist between single narrative fragments.

Because of the space restrictions, it is now impossible to supply detailed information about these two mechanisms, see (Zarri, 1997; 1998) for further details.
4. Some remarks on the queries and inference procedures

Search patterns are NKRL data that supply the general framework of information to be searched for, by filtering or unification, within an NKRL knowledge base. The upper part of Table 2 is the representation of a very simple narrative fragment: “On June 12, 1997, John and Peter were admitted (together = COORDination) to hospital”. The “temporal modulator” begin asserts that the date associated with date-l corresponds to the beginning of the state of being at the hospital. Modulators (deontic, modal, and temporal modulators) are special codes that are added to the basic core of a predicative occurrence to better specify its conceptual meaning, see (Zarri, 1998).

A simple example of search pattern, translating the query: “Was John at the hospital in July/August 2001?” is represented in the lower part of Table 2. The two timestamps associated with the pattern constitute the “search interval” that is used to limit the search for unification to the slice of time that it is considered appropriate to explore.

In the Java XML/RDF-compatible version of NKRL (see Zarri, 2000), the FUM (Filtering Unification Module) deals with the unification of search patterns. Its inference level is only a first step towards more complex reasoning strategies, see (Zarri and Azzam, 1997). A powerful class of NKRL inference rules concerns the ‘hypotheses’, see the example of Table 3. We suppose here to have directly retrieved, thanks to FUM, a given information within a base of NKRL occurrences, e.g., the information: “Pharmacopeia, an USA biotechnology company, has received 64,000,000 dollars from the German company Schering in connection with an R&D activity”.

We suppose, moreover, that this occurrence is not already explicitly related with other occurrences of the base. Under these conditions, we can activate a specific Java module, Inference Engine that, using a rule like that of Table 3, will try to rely automatically the information found originally by FUM with other information in the base. If this is possible, this last information will represent a sort of ‘causal explanation’ of the information originally retrieved — in our example, an ‘explanation’ of the money paid by Schering. We will find, e.g., that “Pharmacopeia and Schering have signed an agreement concerning the production by Pharmacopeia of a new compound” and that “In the framework of the agreement previously mentioned, Pharmacopeia has actually produced the new compound”.

c2)  EXIST  SUBJ  (COORD john_ peter_): (hospital_1)
{ begin }
date-1: 2-june-1997
date-2:

(\?w  IS-PRED-OCCURRENCE
 :predicate  EXIST
 :SUBJ  john_
 :location of SUBJ  hospital_
 (1-july-1997, 31-august-1997))
5. Conclusion

In a 'traditional' ontology, concepts are defined (synthetically) as frames. It is now evident that an organization in terms of frames is largely sufficient to provide a static definition of the concepts — i.e., a definition a priori of each concept considered in itself. It is, however, difficult to admit that this type of organization could be sufficient to define the dynamic behaviour of the concepts, i.e., to describe their mutual relationships when they take part in some concrete action, situation, etc. (the NKRL "events").

Premise:

\[
\text{RECEIVE SUBJ } x \\
\text{OBJ } money_ y \\
\text{SOURCE } y
\]

\[x = \text{company}_\]
\[y = \text{human_being} | \text{company}_\]

"A company has received some money from another company or a physical person"

First condition schema:

\[
\text{PRODUCE SUBJ } (\text{COORD x y}) \\
\text{OBJ } z \\
\text{BENF } (\text{COORD x y}) \\
\text{TOPIC } (\text{SPECIF process}_ v)
\]

\[z = \text{mutual_commitment} | \text{business_agreement} \]
\[v = \text{tool/product}_\]

"A general or business-oriented agreement about the creation of a new product has been concluded by the two parties mentioned in the premise"

Second condition schema:

\[
\text{PRODUCE SUBJ } x \\
\text{OBJ } v \\
\text{MODAL w} \\
\text{CONTEXT } z
\]

\[w = \text{industrial_process} | \text{technological_process} \]

"The company that received the money has actually created the product mentioned in the first condition schema"

---

Table 3. An example of high-level inference rule in NKRL.

Reducing the description of events to the description of concepts — as has been sometimes proposed — is nothing more than a further manifestation of that well known 'uniqueness syndrome' proper to some Artificial Intelligence and Knowledge Representation milieus. In NKRL, we make use in an integrated way of several sorts of representational principles (more details in the final and complete...
paper) and several years of successful experimentation with the most different narrative situations are there to testify that this seems not to be a totally unreasonable approach.

References