URUK, a platform for causal text retrieval


Dirk VERVENNE
Hogeschool Gent,
Cel voor Onderzoek en Dienstverlening
Voskenslaan 270 B-9000 Gent, CTL-building
Gent, Belgium, 9000
Dirk.Vervenne@hogent.be

R. Godijns (**), S. Vandepitte (**),
H. Verplaetse (**), C. De Groote (**),
K. Denturck (**) P. Kaczmarski (**),
S. Gierts (**), F. Vandamme (**)
(**) Hogeschool Gent, Departement Vertaalkunde
(*** ) Labo Toegep. Epistemology, University Gent

Abstract

In this paper we present the Uruk-platform: a system that indexes multilingual documents on the basis of causal patterns. The basic idea is twofold: (1) to develop a search engine that is specialised in causal retrieval of documents and (2) to connect documents automatically when their causal indexes have common causes or results, taking into account the 3 standard relation types of a thesaurus. The paper presents the global architecture, identifying the basic components. The multilingual approach is explained and we end the paper by describing the test environment. The project is currently funded by IWT through the HOBU-funding programme under contract 10136.

Introduction

Text indexing platforms that incorporate the semantic richness of a thesaurus are usually imbedding the latter as a manual support during the indexing and/or the search process (Callan, 1993; Church, 1988; Croft, 1991; Crouch, 1990). The automatic thesaurus-based text indexing have been introduced in, e.g., Khan et al, 1997, Ginsberg A. 1993, Gao et al, 1995 and Vervenne, 1999. The advantage is this approach is the automatic extension of the index with (1) broader terms through the hierarchic BT/NT-thesaurus relations, with (2) equivalent terms based on the USE/USE FOR-thesaurus relation and (3) related terms through the RT-thesaurus-associations.

However, all these approaches are limited to the true nature of a thesaurus, being the exclusive focus on nouns and their semantic interrelations.

Since the standardisation rules that are contained in ISO-5964, 1985 and ISO-2788, 1986, both express the requirement not to use verbs while constructing and maintaining thesauri, it means that action-related dependencies in the indexed documents, such as causal relations, are not captured during the thesaurus-based indexing process.

Since “causality is one of the most intriguing concepts which human being acquires in the earliest stages of life” (Vandepitte, 2000), it is astonishing that little research effort has been established in the area of automatic causal text indexing. Furthermore, scientific research in nearly every domain is becoming more and more highly specialised; this evolution therefore requires tools which link automatically the small pieces of causal reasoning chains that might be expressed in scientific texts. We therefore decided to combine the strength of our thesaurus-based indexing algorithms (IKEM®) with the complexity of the English and French causal text patterns: this combination provides the cornerstone of the URUK-architecture.

1 The URUK-architecture

The Uruk-architecture, illustrated in figure 1, is basically developed on top of a set of a pattern-matching algorithms. The work by, e.g., Hearst (1998) indicated that this approach is a powerful paradigm for knowledge-based document indexing.

The Uruk-platform is developed as a compatible module to the existing IKEM®-system, which uses a thesaurus to generate automatically keywords and concepts as meta-information (Vervenne, 1999). The latter is then used in the
Uruk-platform as values for variables that are expressed in the causal patterns. These patterns are stored in the Uruk-pattern database, for which a dedicated editor is provided to support the maintenance of the language-sensitive causal expressions.

The Uruk-indexing engine includes a fuzzy algorithm for scanning all sentences in order to map the causal patterns. When a causal expression is found, the thesaurus is scanned for relevant descriptors. Such descriptors are then recognized as ‘causes’ and ‘results’, which are stored in the Uruk-index.

2. Causal document linking

The causal index of our Uruk-platform can be queried through the dedicated search module which both provides the causal query-facility and following advanced search options that are currently developed:

(1) the ‘multi-doclink’ option gives the user the possibility to search for linked documents on the basis of common causal indexes. If, e.g., a document D1 contains the causal link
   ‘A → B’
and an other document D2 is indexed by
   ‘B → C’, then the query
   ‘Is C influenced by A?’
will automatically yield the link between the documents D1 and D2, through the ‘multi-doclink’ search option. This multi-doclink option could function also in case
   ‘Bbis → C’
is given, where Bbis is, e.g., a synonym for B.

(2) the ‘concept-link’ option gives the user the facility to search causal links on the basis of the broader-term descriptors of the indexed causes and results stored in the index-records. If, e.g., a document D1 contains the causal link
   ‘AA → BB’
and if A and B are thesaurus-descriptors that are not present in D1 and if AA is a narrow term of A and BB is a narrow term of B in the thesaurus, then the query
   ‘Is B influenced by A?’
will deliver document D1; we call this the ‘concept-link’ option since the IKEM® platform considers A and B as concepts for document D1.

(3) The combination of the two mentioned query-options leads to the facility of identifying ‘knowledge gaps’ within a collection of documents that have been indexed with the Uruk-platform. If, e.g., a document D1 contains the causal link
   ‘AA → BB’
and an other document D2 is indexed by
   ‘B → DD’,
and if term BB is narrow term of B then the Uruk-platform will propose the hypothesis that a causal relation might exist between AA and DD.

3. Multilingual causal patterns

Causal patterns have been edited manually by the research members of the project: as experts in English and French language, they developed
a classification of causal expressions and stored the results in a formalised way within a pattern lexicon (Vandepitte, S. 1993 and Godijns, et al., 2002).

Causes and results in such patterns are represented as variables that can be searched in the respective domain thesaurus. Two example-patterns are included in figure 2 (for English) and figure 3 (for French).

\[
< C \text{ produces } R > = \{\text{KEYS1}; \text{produces}; \text{KEYS2};
\]

\[
< R \text{ être le produit de } C > = \{\text{KEYS2}; \text{être le produit de}; \text{KEYS1};
\]

The variables KEYS1 and KEYS2 represent cause en result-terms which are captured from the pre-installed thesaurus. In Kaczmarski P., Gierts S., 2002, a prototype called PatViews, is presented. A screendump of the PatViews-viewer is given in figure 4.

As a side effect of the project, causal pattern structures are compared between the English and the French languages. Also, the developed thesauri in the field of the stock exchange will be studied: e.g., the way the hierarchic broader/narrow relations between terms are expressed in both languages and how a merging process could be supported by the ThesEdit-software.

4. Evaluation tests

The Uruk platform is currently tested within the domain of stock exchanges: the hypothesis is that in this field, many causal relations within socio-economic processes can be detected (Hoover, K. D. 2001). A collection of documents in that domain, both in French and in English were selected. These documents have been used to construct the two domain thesauri.

The tests for the automatic recognition of the causal patterns, are planned in winter 2002. A final prototype will be demonstrated mid 2003. Current tests are limited to ‘unambiguous causal patterns’ that operate within one single sentence. Since many patterns operate on a paragraph-level, we plan to tackle unambiguous anaphoric expressions in the near future.

Tests have already revealed that some causal relations only hold under certain conditions. The latter should be mentioned when query results are presented to the enduser.
Conclusion

The Uruk-platform combines automatic thesaurus-based indexing algorithms with causal pattern analyses. This platform aims at indexing automatically multilingual texts from a causal point-of-view. The search facilities combine thesaurus-based concept-search as well as document-linking features.

Current tests are running within the domain of stock exchanges for unambiguous patterns that operate on a single sentence level. Since our Uruk-project is half way, more complex patterns will be treated in the next phase of our project.

Acknowledgements

Our thanks go to IWT for providing financial support to the project HOBU-10136.

References


