Automatic hypertext information retrieval in a corporate memory using noun phrases in context

Maria Nava\textsuperscript{1,2} and Daniela Garcia\textsuperscript{2}

\textsuperscript{1}Université de Paris-Sorbonne
Institut des Sciences Humaines Appliquées
96 boulevard Raspail,
F-75006 Paris, France
maria.nava@edf.fr

\textsuperscript{2}Electricité de France R&D
Dept. MTI/NTIC/TAIC
1, avenue du Général De Gaulle
F-92141 Clamart, France
Daniela.Garcia@edf.fr

Abstract

In this paper, we describe a method to generate an information retrieval hypertext structure on a large collection of homogeneous documents by generating links only between noun phrases that are pertinent for navigation. Noun phrases are selected by automatic extraction and filtered on the basis of the linguistic context class where they appear, also determined automatically.

1 Introduction

Hypertext generation based on automatically extracted key-words usually produces an overwhelming number of non-pertinent links. Any noun phrase (NP) can actually constitute an anchor for too large a set of heterogeneous links, a serious limitation to the effectiveness of information retrieval (IR).

The aim of our research is the development of a system capable of generating a IR hypertext structure on a large collection of homogeneous documents by selecting only those NPs that are pertinent for navigation.

Our work can be placed in the domain of IR automatic hypertext (Agosti \textit{et al.} 97, Allan 97), where paragraph (and document) linking is based on IR similarity measures, and is typed.

2 A real-world application

The specification of our IR hypertext system is based on a real-world application supplied by \textit{Electricité de France R&D} (EDF, the French electricity board)\textsuperscript{1}, that is, browsing a large document base made of customer enquiry letters, along with the associated reply letters. The aim of the navigation in the document base is to help finding consistent answers to any new incoming letter.

As our document base is liable to frequent updating, the hypertext structure has to be generated at each IR session. Therefore, the document base is dynamically indexed by a short content-sample text at the beginning of the session.

A new browsing session is booted by the content of the incoming letter, which supplies content elements to compute a thematic similarity with any enquiry letter stored in the corporate memory. Navigation allows to gather information on similar cases that have already been solved and reuse written material to compose a response to handle the problem.

Our research aims at identifying the context where interesting NPs occur in the enquiry letters, in order to enhance the selection of pertinent cross-document links. Context identification is based on spotting linguistic markers of the expression of enquiries and on the exploitation of a semantically structured terminology we can extract automatically from the overall document base.

3 A corporate memory corpus

A corpus of about 2000 customer letters, in French, has been made available by \textit{Electricité de France R&D}. The collection contains inquiries, intervention requests and complaints. Even when a complaint is not formulated explicitly, generally the writer’s intention is to point at some sort of problem that needs fixing.

The corpus is homogenous from the point of view of the general subject matter and purpose of the letters. On the other hand, the variety of speech acts performed by the writers lends a challenging

\textsuperscript{1} This PhD research is financed by \textit{Electricité de France R&D}. 
heterogeneity to the texts, interesting but problematic for automatic processing.

The corpus is introduced in the corporate memory as a case base, that can be connected to customer profile and commercial strategy databases for global information retrieval about a single customer case.

The case base is submitted to linguistic analysis and automatic lexical extraction.

4 Identifying the context of lexical expressions

Textual similarity is computed from what we call the “pragmatic profile” of an input letter. We want to identify the discursive context of NPs in order to select only the most interesting ones and create links to similar NPs appearing in the case base, in comparable discursive contexts.

Our research is based on the articulation of two principles:
• The exploitation of a lexicon structured by grammatical relations, extracted automatically from the whole text collection;
• The identification of linguistic markers indicating the expression of requests, complaints, justifications and other discourse acts that are relevant in our working context.

These two principles are implemented in two different NLP systems offering complementary functions and results, that we have integrated.

4.1 The Lexter system

The acquisition and exploitation of a structured lexicon are carried out automatically by the Lexter system (Bourigault et al. 96), developed at EDF R&D Division. Lexter was designed to extract NPs from a corpus of texts (in French). NPs are then automatically organised in a structured network of head-expansion relations.

Lexter accounts for morphological variants and head-modifier relations of nouns and NPs, that are grouped into families. It also supplies simple distributional figures, such as frequency of a candidate term in the corpus or candidate term head-modifier productivity within the network. Lexter stores the whole corpus divided into paragraphs, along with a pointer to the location of each candidate term in the text.

We have taken advantage of all these features for the generation of hypertext links.

4.2 The ContextO System

The identification of context classes where candidate terms appear is based on the contextual exploration method (Desclès et al. 97) implemented by ContextO (Ben Hazez & Minel 2000).

The exploration engine deployed by ContextO can already identify markers of a large number of linguistic functions. The study of the case base corpus has helped us find a number of linguistic structures regularly associated to the expression of complaints, justifications or requests. Each text in the case base contains linguistic markers indicating a focus on certain speech acts that help the writer organize argumentative discourse.

Markers are stored in a knowledge base, which is accessed by the contextual exploration engine of ContextO.

For the first tests, the database contains about 200 markers organized into 24 functional classes.

5 Computing lexical links between texts

Our hypothesis is that the co-occurrence of a candidate term and a focusing structure selects a portion of text interesting for our similarity search in the case base.

For the moment, the search for pertinent markers is a means to refine link generation on a number of texts already selected by their lexical components, extracted by Lexter.

In order to reduce the number and, at the same time, to keep only the most pertinent links, we have decided to maintain only the links between NPs. NPs represent a form of mutual contextualisation of lexical elements and allow a more precise automatic indexing than simple nouns (Evans & Zhai 96). For example, instead of retaining the simple word electricity, we will first choose expressions like electricity bill or electricity meter (as translated from French) as content carriers, because we feel they are thematically more precise. We have then integrated this domain-specific lexical information, extracted automatically by Lexter, and semantic and pragmatic context information supplied by markers of the general language, identified by ContextO.
Links are computed by matching comparable NPs flagged with a semantic tag indicating a context class.

6 First results
The results obtained by testing the system on three sample entry letter are summarized in Table 1.

The performance of our system on sample texts shows that the simple association of NPs and their conditions of use can effectively improve retrieval precision, when compared to results obtained by generating links between NPs alone.

Table 1: First results for three sample input letters

<table>
<thead>
<tr>
<th>Samples</th>
<th>Initial number of link</th>
<th>Non pertinent</th>
<th>Non pertinent links eliminated</th>
<th>Pertinent links eliminated</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>158</td>
<td>81</td>
<td>71</td>
<td>8</td>
</tr>
<tr>
<td>2</td>
<td>93</td>
<td>23</td>
<td>16</td>
<td>11</td>
</tr>
<tr>
<td>3</td>
<td>78</td>
<td>32</td>
<td>24</td>
<td>5</td>
</tr>
</tbody>
</table>

Improved recall rates should be attained cost-effectively by refining the context exploration rules deployed by ContextO, or by adding a relatively small number of synonymic relations to the NP network.

7 Generation of a hypertext structure
The results of link computation are presented in the form of a hypertext structure generated on-the-fly.

Figure 1: Navigating in context classes
The demonstration window (Figure 1) shows the text of the input letter (left) with salient NPs highlighted and (right) a choice of links to pertinent context classes (complaint formulation, enquiry, justification, etc.).

Figure 2: From typed links to target paragraphs
Once a context class has been selected, the links to target texts appear.

8 Conclusion
The results of the first experiments are encouraging as for the precision/recall ratio. However, we feel that these traditional evaluation measures are not completely adapted to the task, as it is often a delicate matter to decide whether two letters are even loosely connected. As we are currently testing the system on a more extensive input letter set, a more flexible evaluation protocol is under study. It will possibly include an improved link type taxonomy and link weighing.

References