Resource integration and customization for automatic hypertext information retrieval in a corporate setting

Maria Nava

Université de Paris-Sorbonne
Institut des Sciences Humaines Appliquées
96 boulevard Raspail,
F-75006 Paris, France

Electricité de France R&D
Dept. SINETICS/TAIC
1, avenue du Général de Gaulle
F-92141 Clamart, France

maria.nava@edf.fr

Abstract

In this paper, we describe our experience in reusing and customizing existing tools to meet new information retrieval needs in a corporate setting.

The problem was to supply an authoring aid to handle customers enquiry letters by exploiting a textual case base.

We decided to integrate, and go as far as possible with, a terminology extractor and a context exploration platform. They were previously developed through an academic and industrial collaborative research.

We have found a method to generate an information retrieval hypertext structure on a large collection of homogeneous documents by creating links 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.

We have tried to point out the peculiar features that made possible the reuse and integration of existing resources, to produce a relatively new solution to a fairly constrained real-world problem.

1. Introduction

Our work is motivated by a novel information retrieval (IR) need formulated in a corporate setting, at Electricité de France R&D (EDF R&D, the research and development department of the French national electricity board). The general problem was to supply an authoring aid to help EDF employees handle customers enquiry letters.

Starting from a textual case base and software available, we were invited to study a flexible and cost effective solution that would respect the employees savoir-faire and experience, and add value to existing tools.

Our approach 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 structured lexicon that we can extract automatically from the textual case base.

2. The starting point

The initial scenario presented a number of constraints to be respected, concerning both the nature of the IR solution and the technical implementation.

2.1. Two corporate memory corpora

Two corpora were used to carry out a linguistic analysis, train our system for marker identification and test processing performance.

2.1.1. A large corpus of stored letters

A corpus of about 2000 customer letters, in French, was first made available by EDF 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 heterogeneity to the texts, interesting but problematic for automatic processing.

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

Unfortunately, letters in this corpus were not associated to the answers they had actually received.

2.1.2. A smaller corpus of letters and related answers

A second corpus of about 200 question-answer pairs is used for testing and discussing evaluation issues. It is a collection of letters that were sent directly to EDF branch managers to solicit special treatment on peculiar issues.
This gives the letters a somewhat special status, which is reflected in their style, vocabulary and structure.

We have used this smaller, more personal collection to put our system to test, point out its limitations, and try to explain them.

2.2. A terminology extraction tool: Lexter

The acquisition and exploitation of a structured lexicon are carried out automatically by the Lexter system (Bourigault et al., 1996), developed at EDF R&D in the framework of a PhD research project. Lexter was designed to extract noun phrases (NPs) from a corpus of texts (in French). Extraction is based on the hypothesis that eligible NPs must exhibit the syntactic pattern of candidate terms, as established by terminology theory. For example: *definite article* + *noun* + *preposition* + *noun* is an observed candidate term pattern. NPs are not extracted by direct pattern matching, but they are isolated by spotting their syntactic boundaries, like, for instance, verbs. The "terminological hypothesis" is not without consequence for our work, as it will be pointed out in the conclusive section.

Extracted NPs are then automatically organized 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 structured network.

Lexter also stores the whole corpus divided into paragraphs, along with a pointer to the location of each candidate term in the text. This feature was initially designed to supply the terminologist with a linguistic context for validation.

Extraction and corpus-related information is stored in a relational database. We have taken advantage of all Lexter features and results for the generation of hypertext links, as described below.

2.3. A context exploration tool: ContextO

The identification of context classes where candidate terms appear is based on the contextual exploration method (Desclés et al., 1997) implemented in the ContextO platform (Ben Hazez & Minel 2000). The system was designed and is still developed at the LaLICC laboratory (Langage, Logique, Informatique, Cognition et Communication) of the Sorbonne University in Paris.

The exploration engine deployed by ContextO is based on the identification of markers of a large number of linguistic functions, as observed in the general language. Markers are acquired through a "manual" linguistic analysis of a corpus of texts, to model the expressions of linguistic functions, depending on the application. They are subsequently organized in semantic classes with object-model relations. Markers are stored in a knowledge base (a relational database, the same as Lexter's), which is accessed by the contextual exploration engine of ContextO, a Java application. Markers are exploited by specialized agents, performing specific tasks. A number of tasks were already available; for example, the identification of static relations (*is-a*, *has*, etc.), causal relations, thematic focus, citations, definitions, etc. For the time being, ContextO exploits markers from French and Spanish, but could easily be adapted to other languages.

The study of our own corpus of letters has helped us find a number of linguistic structures regularly associated to the expression of complaints, justifications or requests. Each letter contains linguistic markers indicating a focus on certain speech acts that help the writer organize argumentative discourse.

For the first tests, the database contained about 200 markers organized into 24 functional classes ("complaint", "demand", "justification", etc.).

3. HyTEC, a new tool born from customization

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

By exploiting the features of the existing tools, we have designed a system, HyTEC (*Hypertext from Terms in Context*), 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 et al., 1997; Allan, 1997), where paragraph (and document) linking is based on IR similarity measures, and is typed.

The specification of our IR hypertext system is based on a real-world application, that is, browsing a large textual case 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, we found it interesting that the hypertext structure 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 enquiry letters 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.

3.1. 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:

1. The exploitation of a lexicon structured by grammatical relations, extracted automatically from the whole text collection;
2. 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 the two different NLP systems, that offer complementary functions and results, that we have integrated.

3.2. 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.

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 contextualization of lexical elements and allow a more precise automatic indexing than simple nouns (Evans & Zhai, 1996). 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. The lexical information triggers context analysis to create a “signature”, a context-tag / NP relation, that is used for indexing and filtering.

Even if the actual language we use is French, the same principle may as well be illustrated with an example in English, like

Due to temporary money problems, I’d be happy if I could pay the bill by installments.

Context analysis is triggered by the phrases in italics (pay the bill would be a nominal form in French). As markers like I’d be happy if (demand) or Due to (justification) would be found in a particular context (by context exploration rules), the sentences would be tagged as belonging to a pertinent context class.

Links between portions of texts are computed by matching signatures formed by NPs that are flagged with a semantic tag indicating a context class.

4. Similarity search 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>
<thead>
<tr>
<th>Samples</th>
<th>Initial number of links</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>

Table 1: Results for three sample input letters on the main corpus

For instance, consider the following input text (as translated from French), where extracted NPs are in italics and context markers are in bold:

Dear Sirs,

Earlier this month, I have received an invoice from you, concerning the use of gas and electricity, whose amount I do not agree with. As the big amount you are asking for apparently concerns only a 2-month period, I have taken down the numbers shown on my gas meter. The meter indicated 00613, but your invoice reports 00878. I know this number represents an estimation.

On the other hand, if we consider the huge difference between your estimation and my actual gas consumption, I refuse to pay the amount you are asking for and invite you to send me a new invoice, reporting figures closer to reality.

Context identification allowed to retain a target text like:

Dear Sir,

I am the tenant of the apartment located X Street in TheTown, belonging to Mr and Ms Y. Since I have taken up the place in 1996, I have only received invoices reporting estimations of my electricity consumption.

Before I was here, the place was unoccupied. I’ll take the liberty to tell you that at present, the electricity meter indicates 36,637.

I’d be grateful if you could send me an invoice corresponding to the actual consumption.

Notice that the NP real ... consumption was also included in a focused sentence in the input letter (On the other hand ..., I refuse ...). Notice also that the target text focuses on the NP electricity meter (I’ll take the liberty
to…), that could also be found in the input letter under the form gas meter.

We have found that it is not necessary to carry out context indexing for the input letter to improve precision: it is enough to search the context of NPs in the candidate target texts. On the other hand, to execute a relational indexing (NP + context tag) both for the input and the target texts allows precise link typing, which makes navigation easier.

The same search session as above allowed to eliminate a number of target texts, that had been retrieved be found on the basis of simple NP matching, but were not pertinent, like:

Sirs:

I take the liberty to draw your attention to the dangerous situation menacing all the families living in our building.

We experienced important damages due to water overflow last summer. To day, the leakage, which has not been stopped yet, affects the wall bearing our electricity meters and wiring.

If you consider that there actually is no danger, I'd be grateful if you could send us an official written declaration about this, etc.

In this target text, there is no co-occurrence of context markers and extracted NPs, as compared to the input letter. Therefore it was not retained by HyTEC, which improves retrieval precision.

Eventually, we are left with 1) non-pertinent targets that have been retained, but also 2) pertinent candidates that have not been retrieved.

In the first case, co-occurrence of markers and NP has been identified in a sentence or paragraph, yet the rest of the letter relates events or circumstances that are different from those found in the input letter. However, the noise caused by uninteresting letters is very low, considering the number of searched texts. In this case, retrieval precision would probably improve if we could rely on a global text model, accounting for lexical and discursive chaining.

In the second case, in spite of global similarity between the input letter and a possible candidate target, content proximity has not been identified. The most frequent cause of this kind of failure is that pertinent markers focus on synonyms of extracted NPs. Improved recall rates should be attained cost-effectively by adding a relatively small number (the corpus is homogeneous) of synonymic relations to the NP network. We are planning to test the integration of a tool (SynoTerm) that automatically supplies Lexter with candidate synonyms from general language resources (digital dictionaries) (Hamon, 2000).

5. Generation of a hypertext structure

The results of link computation are presented in the form of a hypertext structure generated on-the-fly, directly exploiting the data structure in Lexter’s relational database.

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 (Figure 2).

6. Evaluating task performance

The results of the first experiments are encouraging in terms of precision/recall ratio (Nava & Garcia, 2001). However, we feel that 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.

7. Methodological issues about customization

In this paper, we have shown how we have reused, customized and integrated two different NLP tools.
Lexter and ContextO belong to two different paradigms, which are, we believe, complementary. Lexter and ContextO were not designed to be integrated. Lexter is a corpus-based extraction tool, ContextO is a knowledge-rich filtering system. However, we found that their results are complementary, and their coupling has provided benefits that reach beyond the simple application of cascading NLP processes. In our case, we have observed that facility of integration and customization are related to a number of features, ranging from modularity and separation of linguistic resources and procedures, to implementation by off-the-shelf technology.

7.1. Lexter

We have taken advantage of the following:

1. Corpus-based extraction without domain-specific resources (dictionary, thesaurus, etc.);
2. Shallow morpho-syntactic structuring;
3. Access to the full-text source;
4. Simple distributional data (frequency, head-modifier productivity in the morpho-syntactic network);
5. Extraction results stored and organized in an off-the-shelf relational database (Microsoft Access).

On the other hand, considering our particular application, we have experienced an important limitation due to the fact that Lexter is basically an extractor of candidate terms. This is certainly well suited for technical, domain-specific text processing; but for our collection (customers letters), this constraint is rather restrictive. Given the source, purpose and style of the texts, we would have been happier with additional lexical information, like, for example, verbal phrases (which are generally ignored by the classical terminology theory and applications). In informal writing, an expression like pay the bill is often preferred to bill payment.

7.2. ContextO

ContextO was designed to facilitate the acquisition and reuse of linguistic knowledge, based on the following:

1. Separation of linguistic knowledge and search engine;
2. State-of-the-art object model of text, linguistic data and tasks;
3. Independent specialized agents exploiting the knowledge base;
4. Portable Java engine implementation accessing an off-the-shelf relational database (Microsoft Access);
5. Exploitation of markers related to structures of the general language.

It must be noted, however, that if the marker collection is largely domain-independent, it is sensitive to style and textual genre. Moreover, certain semantic classes (for example, thematic markers) are more generally reusable than others (for example, static relation markers).

Prospective work includes the adaptation of our approach to automatic, corpus-based terminology structuring.

8. Acknowledgements

Our PhD research is financed by EDF R&D.

9. References


