Constraining the Use of General Lexical Resources for Automatic Hyperlink Generation

Farid Cerbah
Dassault Aviation
DPR/ESA - 78, quai Marcel Dassault
92552 Saint-Cloud cedex 300
France
farid.cerbah@dassault-aviation.fr

Abstract

Turning a technical documentation into a hypertext is a tedious and time-consuming task. Unfortunately, no active help for identifying relevant hyperlinks is provided in modern editing environment. We define in this paper an NLP approach to automatic hyperlink generation based on a constrained use of general lexical resources. We propose a formal model that estimates the lexical similarity between sources and potential destinations of hyperlinks. Local expansion mechanisms are defined to precisely control the integration of external knowledge for similarity estimation. We discuss the experiments conducted on a significant technical corpus.

1 Introduction

Hypertext can be considered as one of the best structuring devices of technical writers. By means of hyperlinks, the authors can abstract away from textual linearity to provide the users with a large diversity of selective reading modes more suitable with the common use of technical documentations. However, turning a documentation into a hypertext requires tremendous manual efforts. For example, a typical maintenance documentation in aeronautics contains several hundreds of thousands hyperlinks which are mostly manually inserted. New editing softwares integrate basic assistance for manual insertion of hyperlinks, but no active help that would facilitate the identification of relevant links. The objective of this work is to define an NLP approach to automatic hyperlink identification that may provide the foundation for an assistance module to be integrated in a typical editing environment.

We define a formal method based on the estimation of lexical similarity between sources and potential destinations of hyperlinks. We pay a particular attention to the definition of *local expansion mechanisms* that allow to precisely control the integration of external lexical knowledge in the representations of link destination candidates. Another major objective of this work is to determine the extent to which general lexical resources used in a constrained way can improve hyperlink generation. We evaluate the level of performance that can be reached by a method that exploits existing general-purpose resources, such as derivational relations and synonyms, while avoiding the use of terminological resources which require considerable efforts for initial acquisition and updating.

We concentrate on the identification of *expansion links* (Allan, 1997) that relate condensed texts to longer fragments or documents on the same topics. These links are very frequent in technical documentations, more particularly to link instructions to their expanded forms in maintenance procedures.

We start by outlining in section 2 some features of hyperlink generation as applied to structured technical documentation. We show in section 3 how lexical resources can help to improve the process. In section 4, we describe the computational model, and section 5 is devoted to the experiments and evaluation. In section 6, we give some directions for future research.

2 Hyperlinking Marked-up Texts

Mark-up languages, and particularly XML, are the cornerstone of modern approaches to technical documentation production. The documents elaborated following these approaches are often highly structured through an explicit mark-up that delimits the information units. This feature of the input material has some influence on the definition of an automated approach to hyperlink generation:

- It is not necessary to invoke a prior difficult step of text segmentation (Hearst, 1997; Ferret et al., 1998) from the text

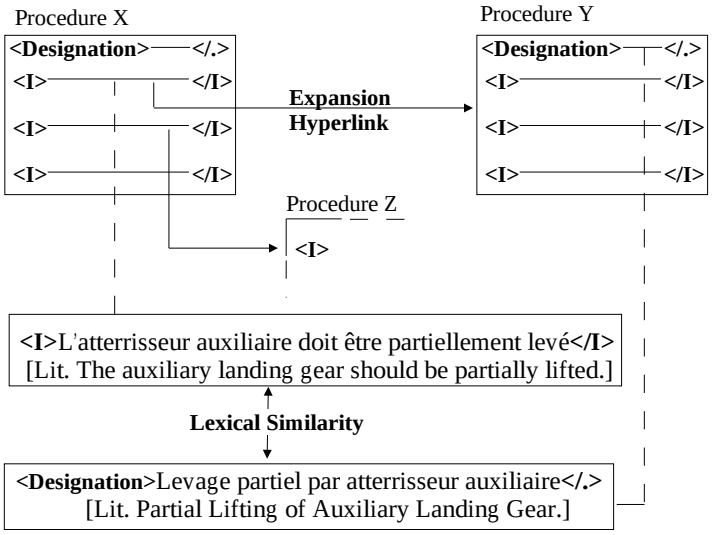

Figure 1: Expansion hyperlinks and lexical similarity between the source of a link and a subpart of the destination.

flow to identify potential anchors. It can be assumed that nodes of the hypertext to be generated are structural elements. Hence, hyperlink generation has to relate text spans whose boundaries are already marked.

- An estimation of the linguistic similarity between sources and potential destinations can be used to identify relevant hyperlinks. Additionally, with structured information as input, similarity estimation can be focused on clearly localized parts of the elements to be linked. If subparts that are potentially more relevant are identified in sources or destinations, it could be more efficient to use them for similarity estimation, instead of exploiting the entire textual content of the elements.

These proposals can briefly be exemplified with the corpus we used in our experiments. This corpus is composed of about 2000 XML documents extracted from the maintenance documentation of a civil aircraft that contains more than 20 000 documents. Most of these documents are procedural. Various types of hyperlinks can be found in this documentation. Expansion links are the most frequent, and much of them are established between instructions and procedures.

When an expansion link exist, a lexical similarity relationship is observed between the source instruction and a subelement of the destination procedure named *Designation* (which takes the form of a title at presentational level). This is illustrated in figure 1.

A method for hyperlink generation which is based on the estimation of lexical similarity can focus on these subparts, and thus avoid an integral exploration of the potential destinations.

3 Which Lexical resources ?

A key problem tackled in this work is the identification of the lexical resources needed to efficiently estimate semantic relatedness between text spans. It could be argued that the technical nature of the texts to be produced considerably reduces the lexical choices that could be made by the technical writers. Consequently, the difference between two linguistic formulations of the same idea is so small that an acceptable level of performance could be reached by simply identifying strict word overlaps between potential link ends. But even a quick look at our corpus shows a lexical variability that excludes such a minimalist approach to the problem. We will see in section 5 that our experiments do confirm the inadequacy of this approach.

Figure 2 gives some examples that illustrate the role of various types of lexical resources in similarity estimation. Example 1 is an "ideal" case where several strict lexical overlaps exist between the source (in italics) and the correct destination (in boldface). It is not necessary to involve other information than lemmas to discriminate the correct destination from the other candidates that also share elements with the source.

1. *Effectuer un essai de détection incendie.* [Lit. Do a test for fire detection] **[L]**
   - **Détection incendie. Essai** [Lit. Fire detection. Test]
   - Détection de fuite. Essai [Lit. Leak detection. Test]

2. *Déconnecter la batterie de servitudes.* [Lit. The ancillary battery should be disconnected] **[L+D]**
   - **Batterie de servitudes. Déconnexion** [Lit. Ancillary battery. Disconnection.]
   - Batterie de servitudes. Contrôle de la tension [Lit. Ancillary battery. Voltage check.]

3. *Vérifier la tension de la batterie essentielle.* [Lit. Check the voltage of main battery] **[L+D+S]**
   - **Batterie essentielle. Contrôle de la tension** [Lit. Main battery. Voltage Control]
   - Batterie essentielle. Dépose [Lit. Main battery. Removal]

4. *Déposer l'ensemble télescopique* [Lit. Remove the telescopic assembly.] **[L+D+T]**
   - **Echelle intégrée. Dépose** [Lit. Integrated Ladder. Removal]
   - ensemble frein. Dépose [Lit. Brake Assembly. Removal]

- **L:** Lemmatization, **D:** Derivation, **S:** Synonymy, **T:** Terminology

Figure 2: Some examples that illustrate the various types of lexical relationships between sources and destinations of hyperlinks. We give the source in italics, the correct destination in boldface, and in next line, a candidate that is also lexically related to the source instruction. Letters **L, D, S**, and **T** indicate the types of lexical resources required to discriminate the correct destination from other candidates.

However, in example 2, the derivational relation between *déconnectée* (*disconnected*) and *déconnexion* (*disconnection*) should be identified. In example 3, the discrimination of the relevant destination is achieved through the combination of the derivational and synonymy information required to associate the verb *vérifier* (*check*) with the noun *contrôle* (*control*). Finally, the correct destination in example 4 cannot be identified without considering the domain-specific synonymy relation between multi-word terms *ensemble télescopique* and *échelle intégrée*.

Only general purpose resources are involved in our analysis of examples 1, 2 and 3, while example 4 requires specific terminological resources. Our study is focused on the exploitation of general resources. However, we briefly discuss in section 6 the possible integration of terminological resources in the proposed model.

We used a network of synonyms extracted from the Crisco dictionary[1]. This highly relational lexicon has been elaborated by merging seven dictionaries into a unique coherent source of more than 50 000 entries and 400 000 relations. The term synonymy has in this lexical network a broad meaning, including hyperonymy and hyponymy.

In Information Retrieval (IR), there is no real agreement about the utility of semantic lexical relations for improving search results (Voorhees, 1993; J.Gonzalo et al., 1998; de Loupy and El-Bèze, 2002). Several studies have shown that involving such external resources in similarity estimation tends to improve recall but at the cost of significantly decreasing precision. In the hyperlink generation framework, our experiments tend to confirm that a systematic use of synonyms heavily degrades the performance. However, we show that a small but yet positive effect can be obtained through a constrained use of these semantic resources.

4 A Formal Model of Hypertext Generation

It is quite common to draw an analogy between the classical IR problem and hypertext generation (Allan, 1997; Nakagawa et al., 1998). In both issues, an input text (request or hyperlink source) has to be associated with the most semantically related documents or text fragments extracted from a large text collection, and it is assumed that textual similarity can reasonably mirror semantic relatedness. Following previous work, we use the standard vector space model (VSM) of IR as a reference framework for similarity estimation. For sake of clarity, we give in next section a short description of this well known model. The key issue for us in the definition of such similarity models is to find suitable ways for incorporating external knowledge. We define in section 4.2 expansion mechanisms well suited for link generation.

[1] http://elsap1.unicaen.fr

4.1 The Basic Similarity Model

In the VSM framework, textual similarity is identified with spatial proximity between vectors in a high-dimensional space. Each dimension corresponds to a lexical unit (or indexing term).

Every manipulated text fragment — sources and potential destinations of hyperlinks in our context — is mapped to a vector representation in an N dimensional space composed with the set of lexical units $U = (u_1, \ldots, u_N)$.
A potential destination d is represented by a vector $D = (d_1, \ldots, d_N)$ where the component d_i is the estimated weight of unit u_i in d. Similarly, a source s is represented by a vector $S = (s_1, \ldots, s_N)$.
Different types of weighting functions can be used for assigning values to components of S and D. We use the following variant of *TF.IDF*:

$$d_i \ (or\ s_i) = tf_i \ \cdot \ log_{10}\left(\frac{M}{df_i}\right)$$

tf_i is the frequency of unit u_i in the destination (or source), M is the total number of documents, and df_i is the number of documents where u_i occurs at least once.

The *cosine* measure can be used as an estimation of the similarity between a source s and a potential destination d:

$$M_{cos}(S, D) = \frac{\sum_{i=1}^{N} s_i d_i}{\sqrt{\sum_{i=1}^{N} s_i^2}\sqrt{\sum_{i=1}^{N} d_i^2}} \qquad (1)$$

Vector normalizations in the denominator reduce the bias in favor of long fragments.

Finally, the link generation procedure calculates this measure for all potential destinations, and selects the candidate that holds the maximum score.

The lexical units are in practice simple or complex canonical forms that result from morphological parsing of wordforms attested in the text collection. The resulting units can correspond either to stems or lemmas. However, since our goal is to introduce semantic lexical relations in the process, it is clearly more convenient to manipulate lemmas that could be directly mapped to entries in general or specialized lexicons.

4.2 Local Expansion Mechanisms

The integration of external knowledge is performed through reweighting mechanisms that are applied on potential destinations. A richer vector representation is assigned to each candidate by taking into account lexical relations that hold between its units and the units of the source. Two successive reweighting steps are applied corresponding respectively to derivational and synonymy relations.

We start with an informal description of these reweighting principles. Suppose that a unit of the source does not occur in the potential destination at hand, but one of its morphological derivatives (or synonyms) does. Then, the derivative (or synonym) is replaced in the representation of the destination by the related unit found in the source with an adjusted weight. This means that external knowledge is exploited to get the representation of the candidate as close as possible to the representation assigned to the source.

This reweighting scheme can be defined formally in the VSM framework.

Before any comparison with a given source s, a potential destination d is mapped to a new representation D' defined by the following vector combination:

$$D' = \alpha \cdot D + \beta \cdot M + \gamma \cdot E$$

- $M = (m_1, \ldots, m_N)$ is a vector that results from the identification of derivational relations between elements of s and d.
- $E = (e_1, \ldots, e_N)$ results from the identification of synonymy relations.
- $\alpha, \beta, \gamma \in [0, 1]$ are parameters which are used to adjust the importance given to the different types of information involved in this combined representation.

It follows that the components of the new vector are defined by:

$$d'_i = \alpha \cdot d_i + \beta \cdot m_i + \gamma \cdot e_i$$

The components of m_i and e_i are obtained through iterative invocation of the reweighting rules given in figure 3. These rules provide a precise account of the reweighting principles in terms of vector manipulation operations.

The similarity between the source and the potential destination is estimated by $M_{cos}(S, D')$ as defined by expression (1).

This formal description bears some relations with IR methods of request expansion based on

- Derivational Reweighting

- $\left.\begin{array}{l} m_i = s_i \\ m_j = -d_j \end{array}\right\}$ for each couple $(i,j), 1 \leq i,j \leq N$ such that $\left\{\begin{array}{l} d_i = 0 \\ s_i \neq 0 \\ mderiv(u_i, u_j) \end{array}\right.$
- All other components of M are set to zero.

- Synonymy Reweighting

- $\left.\begin{array}{l} e_i = s_i \\ e_j = -d_j \end{array}\right\}$ for each couple $(i,j), 1 \leq i,j \leq N$ such that $\left\{\begin{array}{l} d_i = 0 \\ s_i \neq 0 \\ m_i = 0 \\ syn(u_i, u_j) \end{array}\right.$
- All other components of E are set to zero.

mderiv is satisfied if a derivational relation holds between the lexical units u_i and u_j.
syn is satisfied if u_i and u_j are synonyms (this relation is transcategorial in the sense that the related units might have distinct parts of speech).

Figure 3: The Reweighting Rules

relevance feedback (Hust et al., 2002; Harman, 1992). In these methods, reweighting schemes are used to feed back user judgments of relevance on documents provided by the system as a response to an initial query. The representation of the request is enriched with additional search terms extracted from the documents marked as relevant. However, we should note some basic differences with these approaches. Firstly, in our context, the expansion is not user-mediated since the additional units are obtained through an automated extraction process from relational lexical resources. Secondly, reweighting does not modify the source (equivalent of the request), but the potential destinations. This is a prominent feature of our approach which deserves to be highlighted in more technical terms with respect to the common idea of request expansion in IR. In traditional expansion methods, the request representation is the steady reference point to be compared with elements of the document collection. The expansion is necessarily global, and it aims at integrating beforehand in the request representation all units that might help to identify relevant documents. As a consequence, a short request is often turned into a long one. This has undoubtedly negative side effects on relevance estimation. In the VSM framework, there is no theoretical obstacle to applying expansion on documents (i.e. destinations). Each document could be modified by a specific expansion that might get it closer to the unchanged input text (request or link source). A single global expansion is then replaced by several local expansions. This is the basic idea behind our reweighting scheme. In practice, such a scheme would probably be inefficient in a classical IR context because of its computational complexity. However, as confirmed by our experiments, it can be successfully applied to structured document collections that fall in the quite large category defined in section 2.

5 Experiments

We used a corpus of 1927 document modules (450 000 words) which is a significant basis for experiments in a specific technical domain. Since this documentation is already in a hypertext form, the system is evaluated in its ability to follow the steps of the technical writers, and make the same hyperlink choices.

These experiments are focused on expansion links between instructions and procedures. As discussed in section 2, it could be more efficient to concentrate similarity detection on a potentially more relevant subpart of a document rather than exploiting the entire textual content of the document. More particularly, to identify expansion links in our corpus, higher performances are obtained by using only subelements named *Designation* for detection of similarities with instructions[2].

5.1 Implementation

We have developed a comprehensive processing chain that can support large-scale experiments.

[2]We also tried several runs with the entire textual content. The results were really poor compared with the results reached using only highly relevant subelements.

α / β / γ	(1) 1 / 0 / 0	(2) 1 / 1 / 0	(3) 1 / 1 / 1	(4) 1 / 1 / 0.5
$k = 1$	24,40%	56,10%	49,10%	57,11%
$k = 5$	54,58%	73,33%	70,44%	73,51%

Table 1: Results of experiments on the generation of 7721 hyperlinks.

In the overall generation process, the XML documents are parsed in order to extract the source instructions (`<i>` elements) with their hyperlinks, and the subelements representing the potential destinations (`<Designation>` elements). The extracted textual elements are analyzed with MultAna, a POS tagger built upon the MMORPH morphological analyzer (Petitpierre and Russell, 1995). After this stage, the manipulated units are in a lemmatized form. Then, the tagged texts are explored to build the space of lexical units, and to compute the frequencies required for assigning the weights. Standard representations of the potential destinations are constructed before link identification. However, the representations that are actually used are obtained after local expansion when processing a particular source. The input sources are mapped to their representations in order to be compared with the expanded potential destinations. Finally, hyperlink identification is applied to assign a ranked list of destinations to each source.

5.2 The Exploited Resources

We used a network of morphological derivations that covers the lexical units of this corpus. The noun-verb derivations are the most exploited in similarity detection, but the adverb-adjective derivations are also frequently involved.

The synonyms extracted from the Crisco dictionary are used without any pre-filtering for excluding words not pertaining to the lexical register of the aeronautical domain.

We started by experimenting a massive use of synonyms following the expansion scheme defined in 4.2. We observed a drastic drop of performance that tends to confirm the inadequacy of a systematic integration of external semantic information in similarity estimation (see section 3). These bad results are mainly due to reformulations that are imposed to terminological units which are much less affected by synonymy variation. To reduce this negative effect, an even more constrained use of these semantic resources can be defined by taking advantage of stylistic regularities pertaining to the specific text genres of the analyzed corpus. A lexical corpus analysis reveals that synonymy variations in procedural texts primarily concern the predicative units, and particularly the verbs expressing the main actions in instructions[3].

We restricted the general relation *syn* defined in figure 3 to the relation syn_{PP} which can only be satisfied if the synonymy relation is established between predicative units[4]. We should stress that no manual customization of the resources is required to apply such a global constraint.

5.3 Evaluation

The hyperlink identification process has been evaluated on 7721 input sources. We experimented several configurations characterized by the values of parameters α, β, and γ. Setting one of these parameters to zero amounts to ignore the corresponding resources.

In our aligned corpus, most of the input sources have only one known relevant destination (as assigned by the technical writers). Hence, the traditional precision/recall distinction is not really informative in this context. We adopted the following evaluation scheme. A destination is considered as correctly assigned if it is among the first k elements of the ranked list provided by the system. Table 1 shows the performance of four representative configurations ($k = 1$, for strict evaluation, and $k = 5$).

Firstly, we should notice the bad performance of the baseline configuration (1) restricted to the use of lemmas (α set to 1, β and γ to 0). This is due to the large amount of discriminative morphological derivations between units of the input source and destinations that could not be identified. Consequently, a slight improvement is obtained with configuration (2) that exploits

[3]For example, *vider* $\leftrightarrow$ *vidanger/vidange* [Lit. to empty $\leftrightarrow$ to drain/draining], *vérifier* $\leftrightarrow$ *inspecter/inspection* [Lit. to check $\leftrightarrow$ to inspect/inspection]

[4]More specifically, verbs, and nouns which are morphological derivations of verbs.

the derivational resources (α and β set to 1, γ to 0).

This positive effect of morphological derivations was predictable, as *stemming* is widely recognized as a useful (though approximate) technique in IR. Such morphological reduction operation is an indirect integration of derivational information for similarity estimation. However, the importance of the discrepancy with respect to the baseline configuration (1) is more surprising. This can be explained by the prominence of the relations between the action verbs and their nominalized forms in these highly procedural texts.

The synonyms used with parsimony in configurations (3) and (4) have a minor effect. Nevertheless, the fact that no heavy loss of performance is observed should be considered as a significant and reliable result because of the scope of these experiments.

The best performance is achieved by configuration (4) with a small improvement over configuration (2). The synonyms are discriminative in few cases, but a thorough analysis of the results shows that the good ranking of many correct destinations is systematically consolidated by the identified synonymy relations. Besides, they are rarely the direct cause of errors (correct destinations found with configuration (2) are also found with configuration (4)). Finally, the loss of effectiveness observed with configuration (3) reveals that synonyms are more likely to be used as an adjustment external source ($\gamma = 0.5$): strict overlaps and derivational relations should have a stronger influence on the final ranking decision than synonymy relations.

6 Conclusion and Further Work

We proposed a model that shows how to carefully control the integration of external resources for similarity estimation in the context of hyperlink generation. Our local expansion method is restricted to derivational and synonymy resources, but its formal definition can be extended to cover other types of resources. An extended model would provide the basis for the integration of terminological resources that might greatly improve the effectiveness of similarity estimation. However, building such domain-specific resources requires huge manual efforts, even though existing NLP techniques can be involved to partially automate the acquisition process (Jacquemin, 2001; Hamon and Nazarenko, 2001). The range of terminological resources that can be exploited is quite large (from multi-word terms to semantic variations), and the magnitude of the required manual effort highly varies from one type to another. Hence, we plan to evaluate the impact of each type, both individually and in combination.

References

J. Allan. 1997. Building hypertext using information retrieval. *Information Processing and Management*, 33(2):145–159.

C. de Loupy and M. El-Bèze. 2002. Managing synonymy and polysemy in a document retrieval system using wordnet. In *LREC '02*.

O. Ferret, B. Grau, and N. Masson. 1998. Thematic segmentation of texts: two methods for two kinds of texts. In *Proceedings of ACL-COLING'98*, Montréal.

T. Hamon and A. Nazarenko. 2001. Detection of synonymy links between terms: Experiments and results. In *Recent Advances in Computational Terminology*. John Benjamins.

D. Harman. 1992. Relevance feedback revisited. In *Proceedings of 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*.

M. A. Hearst. 1997. TextTiling: Segmenting text into multi-paragraph subtopic passages. *Computational Linguistics*, 23(1):33–64.

A. Hust, S. Klink, M. Junker, and A. Dengel. 2002. Query reformulation in collaborative information retrieval. In *Proceedings of the International Conference on Information and Knowledge Sharing, IKS 2002*. ACTA Press.

C. Jacquemin. 2001. *Spotting and discovering terms through NLP*. MIT Press, Cambridge.

J.Gonzalo, A. Penas, and F. Verdego. 1998. Indexing with wordnet synsets can improve text retrieval. In *Proceedings of Workshop on Usage of WordNet for NLP*.

H. Nakagawa, T. Mori, N. Omori, and J. Okamura. 1998. Hypertext authoring for linking relevant segments of related instruction manuals. In *Proceedings of COLING 98*.

D. Petitpierre and G. Russell. 1995. MMORPH – The Multext Morphology Program. Technical report, Multext Deliverable 2.3.1.

E. M. Voorhees. 1993. Using wordnet to disambiguate word senses for text retrieval. In ACM Press, editor, *Proceedings of 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 171–180.