A Semantic-driven Approach to Hypertextual Authoring

R. Basili, A. Moschitti, M.T. Pazienza, F.M. Zanzotto

University of Rome Tor Vergata,
Department of Computer Science, Systems and Production,
00133 Roma (Italy),
{basili, moschitti, pazienza, zanzotto}@info.uniroma2.it

Abstract

Even if the use of the hypertext paradigm is nowadays very diffused, its potential benefits are not completely exploited by the community of the users. This is particularly evident in the case of the news agencies. The major reasons for the above limitation are the high costs for manually creating and maintaining the sets of complete links of a large-scale hypertext. This is especially true for news agencies. Therefore, in this paper we propose a method to address the problem of the automatic construction of the hyper-links based on Information Extraction techniques that enable documents (mainly news items) to be represented in a canonical form, hereafter called *objective representation* (OR). Our hyper-linking method is presented after the analysis of the traditional approaches to the same problem. We will describe the notion of objective representation and the formalism to express the linking constraints. Finally, we will sketch our future research work in the area.

1. Introduction

Even if the use of the hypertext paradigm is nowadays very diffused, its potential benefits are not completely exploited by the community of the users. This is particularly evident in the case of the news agencies. A survey, reported in (Outing, 1996), found that there were 1,115 commercial newspaper online services world-wide, 94% of which used a simplified version of hypertext which does not provide the full use of the hypertext capabilities of the WWW. The users may be able to navigate to a particular article in the current edition of an online paper by using hypertext links, but they must then read the entire article to find the information that interests them. The documents are dead ends in the hypertext, rather than offering starting points for explorations. In order to truly reflect the hypertext nature of the Web, links should to be placed within and between the documents.

The major reasons for the above limitation is, as (Westland, 1991) has pointed out, the high costs for manually creating and maintaining the sets of complete links of a large-scale hypertext. This is especially true for news agencies, given the volume of articles produced every day. Aside from the time-and-money aspects of building such large hypertexts manually, humans are inconsistent in assigning hypertext links between the paragraphs of documents (Ellis et al., April 1994; Green, 1997). That is, different linkers disagree with each other as to where to insert hypertext links into a document.

The cost and inconsistency of manually constructed hypertexts does not necessarily mean that large-scale hypertexts can never be built. It is well known in the IR community that humans are inconsistent in assigning index terms to documents, but this has not hindered the construction of automatic indexing systems intended to be used for very large collections of documents.

The taxonomy of link types given in (Allan, 1995) is very useful to understand the problem of the automatic construction of hyperlinks since it classifies links according to the abilities required for an eventual manual construction. Links are classified according the following classes:

- *Pattern Matching links*, which are easy link to discovered as they can be found through a pattern-matching algorithm. An example of these is glossary links or links between proposition.
- *Automatic links*, which can be in part captured by traditional Information Retrieval techniques. For example links among documents discussing about the same topics.
- *Manual links*, which require text analysis at level of Natural Language Understanding.

While the first two types of links have been approached successfully the third one is judged by Allan (Allan, 1995) to be inaccessible to automatic hypertext construction.

In this paper we propose a method to address the problem of the automatic construction of the "manual" links as defined in (Allan, 1995). The proposed method is based on Information Extraction techniques that enable documents (mainly news items) to be represented in a canonical form, hereafter called *objective representation* (OR). This latter describes some of the important information contained in the documents, mainly the named entities and the domain events found in the target document. Therefore, this document representation allows to draw more motivated inter-document hyper-links since a declarative language for describing linking constraints can be settled over it. Linking rules, i.e. the rules that justify a link between two documents, are in fact written as constraints over the related ORs. The detection of the domain events and of the named entities relies on a knowledge-based IE system composed by a robust parser (Basili et al., 2000b) and a discourse interpreter (Gaizauskas and Humphreys, 1997). As any IE system, this linking methodology requires a large domain knowledge base. The overall approach foresees the methods for the automatic extraction of this knowledge in an unsupervised fashion (Basili et al., 2000a; Basili et al., 2002).

Our hyper-linking method is presented in Sec. 3. after the analysis of the traditional approaches to the same problem (Sec. 2.). We will describe the notion of objective representation and the formalism to express the linking constraints. Finally, we will sketch the future work (Sec. 4.).

2. Traditional Approaches

In literature the automatic construction of hypertext is based on classical IR techniques to measure the relatedness of document couples. Only a *bag of words* are used for expressing the document contents. This results in a poor set of link type manageable in automatic way. In (Allan, 1995) is presented a reformulated taxonomy of links (Trigg, 1983) in order to identify the link type achievable with an automatic approaches. The set of link type has been divided into three major categories based upon whether or not their identification can be carried out automatically (with the IR current technology). The three categories are *Pattern-matching*, *Automatic* and *Manual*. Unfortunately, some types of links straddle the boundaries of the taxonomy, depending upon the document collection being linked.

Pattern-matching Links is a large class of link types. They can be found easily using simple pattern-matching techniques. An obvious example of such a link type is *definition* that can be found by matching words in a document to entries in a dictionary. In almost cases, these links are from a word or phrase to a small documents. They do not take into account the context of the definition so the destination document may be the same for the word or phrase searched for; no matter where the word or phrase occurs. Structural links belong to the pattern-matching category. They are those that represent layout or possibly logical structure of a document. For example, links between chapters or sections, links from a reference to a figure to the figure itself, and links from a bibliographic citation to the cited work, are all structural links. They can be discovered by mark-up codes embedded in the text. Pattern-matching links form a class that is computationally simple for automatic detection.

Automatic Links are links which cannot typically be located trivially using patterns, but which the automatic IR techniques can identify with marked success. Typical automatic links that can be identified are:

- *Revision links* are a fairly straightforward class of relationship between texts, including both ancestor and descendent relationships.
- *Summary and expansion links* are inverses of one another. A summary link type is attached to a link that starts at a discussion of a topic and has as its destination a more condensed discussion of the same topic. Equivalence links represent strongly related discussions of the same topic.
- Tangent links are equivalence links that relate topics in an unusual or tangential manner (often by comparison with other links). For example, a link from a document about *Sivlio Berlusconi* as Italy Prime Minister to one about Milan football club (whose Berlusconi is the president) would be a tangential link.
- Aggregate links are those that group together several related documents. An aggregate link may in fact have several destinations, allowing the destination documents to be treated as a whole when desirable.

Manual links are those which are judged by the IR community unable to be located without human intervention. The natural language understanding researchers have had some significant success within constrained subject areas, so some manual links could be automatically described within those limited domains. Unfortunately, those techniques are not yet extensible to a general setting, so this class of link types seems to remain inaccessible to automatic approaches. Manual links include those which connect documents which describe circumstances under which one document occurred, those which collect the various components of a debate or argument, and those that describe forms of logical implication (caused-by, purpose, warning, and so on).

2.1. A more semantic based approach

An attempt to extend the boundaries of automatic links towards the manual links has been done in (Green, 1997). In this work an automatic method for the construction of hypertext links via *lexical chains* has been carried out. Lexical chains capture the semantic relations between words that occur throughout a text. Each chain is a set of related words that captures a portion of the cohesive structure of a text. By considering the distribution of chains within an article it is possible to build links between the paragraphs. A link is activated if the similarity score of the chains contained in two different articles overcome a predefined threshold. The method comprises three steps: determining the lexical chains in a text, building links between the paragraphs of articles, and building links between articles. A comparison of this methodology with the traditional IR techniques resulted in higher user satisfaction. Lexical chains allow to retrieve a wider set of link type. As an example let us consider two documents that speak about the same fact with different words. The scalar product (a wide used IR metrics in the Vector Space Model) between the two documents would be very low as the documents have different *bag of words*. This prevents the activation of a relatedness link. On the contrary lexical chains refer to the meaning of words. They use synonyms of words in texts so their similarity between documents will be higher.

Lexical chains seems to solve some of IR problem in discovering links but some problems remain unsolved:

- The link type of two documents, which have similar lexical chains, is unknown. We could claim as an explanation that the documents contain some related semantic information. However this explanation is too generic as it is valid for each generated link.
- *Consequence links* remain unsolved. It is not possible specify the consequence relation between two documents for two main reasons: a) The lexical chains of the premised tend to be very different from the consequence. b) These links are directional while the similarity between chains is symmetric.

- Ambiguity and data sparseness affect the precision in discovering valid chains. So we can expect a lot of wrong links.

In next Section it is presented a different approach that solve the two first problems. It provides a methodology for capturing the unsolved link as well as the explanation for them. The third problem has been bound using domain knowledge for conceptualise the information.

3. A "semantic-driven" hyper-linking method

The above approaches mainly relate documents if they are enough similar according to the chosen document representation space, i.e. the bag-of-word abstraction or the lexical chain model. Therefore, according to these approaches "relatedness" is the only reason why two documents may be hyper-linked together. However, this notion of relatedness does not give the possibility of defining user-oriented hypertexts. Each user has to be exposed to the same hypertext regardless his information needs. For instance, the above approaches may relate the two news items in Fig. 1 because of the fact that in the two documents the *Intel* stem increases the relatedness of the two documents. However, the link user is not aware of the reason why the two documents are related and, while reading the first news item, he has not hints that may suggest if the related news article is of any interest to him. The justification of the link may be more easily highlighted if the domain relevant information is captured, i.e. the fact that both the first item and the second one describe an *Intel acquisition activity*.

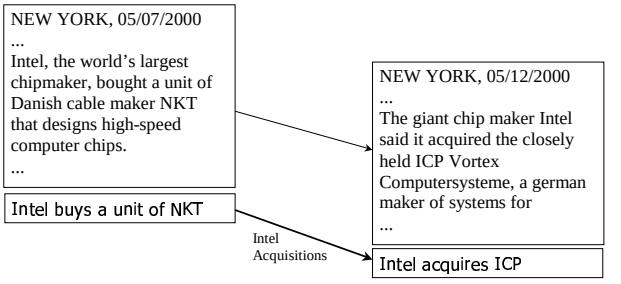

Figure 1: An example of justified link

The facts justifying the hyper-link between the two documents are respectively:

- Intel buys a unit of NKT
- Intel acquires ICP

It is worth noticing that a very precise information is needed for linking the two documents, i.e. the "equivalence" of *buy* and *acquire*. This information may be also used in an IR based hyper-linker using a query expansion technique but the justification of the link is still very difficult.

Furthermore, this notion of relatedness limits the possibility of linking documents. For instance in Fig. 2,

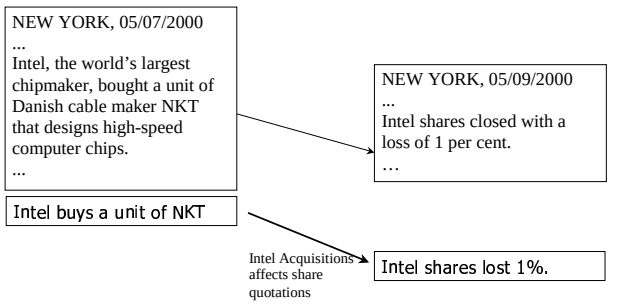

Figure 2: A complex justified link

the link between the two documents is justified by the fact that an *Intel acquisition* affects the *share prices* of a particular period of time. The facts justifying such a kind of document relation are respectively:

- Intel buys a unit of NKT
- Intel shares lost 1%

Such a kind of link is very difficult to capture if the analysis is not based on the more structured document representation.

The automatic hyper-linking method we propose is then based on an abstraction of the document, the objective representation (OR) that describes in a canonical form the salient information carried by the document. This objective representation, due to its nature, may be also considered language independent. Therefore, it enables the automatic hyper-linking between documents of different languages. Both his canonical representation, i.e. the OR, and the language for defining the linking constraints are described in the following sections.

3.1. The objective representation

The quality of the hyper-links that may be drawn in such a method strictly depends on the assumed representation of the document content. Furthermore, it is crucial that the intended information is actually captured by the IE system.

The objective representation we have defined is not too far from the actual document content and aims to represent the relevant document information with respect to a given knowledge domain. In particular, given a document *D*, its OR contains the named entities and the main events of the document *D*. These latter mainly represent particular domain relevant verb phrases that appear in the document. Both the named entities and the events are classified according to a knowledge representation scheme related to a target domain.

The objective representation is then a couple $OR(D) = (NEs, Events)$ where NEs is the set of the categorised named entities of D while the $Events$ is the set of the categorised events. Each event in $Events$ has the following form:

$$EventType(Verb, Arguments) \quad (1)$$

where $EventType$ is the type of the event, $Verb$ is the actual verb that appears in the document and $Args$ are the arguments of the verb according to the event type. Each argument representation carries its syntactic/semantic relation, the actual lexical of its semantic governor, and the type of this latter. For instance, the documents in Fig. 1 should contains respectively in their ORs the following events:

- `buy_event( agent(company,Intel), patient(object,a_unit_of_NKT))`
- `buy_event( agent(company,Intel), patient(company,ICP))`

Naturally, the efficacy of the OR strictly depends on the nature of the information that is contained in the knowledge base. The method for extracting such a knowledge and for the definition of the equivalence between different surface forms is described in (Basili et al., 2002).

3.2. Typing links using events: a declarative formalism

Once an objective representations of documents are available it is possible to write down a set of rules that can activate several links that traditional IR techniques (see Section 2.) cannot capture. However it is not possible to define general linking rules valid for each domains and for each user needs. As an example consider two documents: d_0 that speaks about Ferrari race in the grand prix of Imola and d_1 in which it is stated that FIAT market shares increase their quotation. If a user wants know all the facts which cause the event in d_1 (e.g. the document d_0) some knowledge about the correlation between FIAT and Ferrari have to be draw (i.e Ferrari is a part of FIAT and winning a race increases the share value of a Company).

Thus a systems that really wants to afford hypertext construction including links of third type (see Section 2.) should provide both a set of general rules and a set of specific rules. Moreover, the specific rules should be customisable to satisfy a wide range of user needs. These rules will be then used by the linking algorithm to draw links among documents.

3.2.1. The linking rule formalism

We have adopted a declarative formalism in which the rules and the knowledge required are easy to be written by the final user. The rules are expressed by a logical formalism.

The events in the OR are coded by means of Prolog predicates of the following type:

```
ev(EVENT_CATEGORY, EVENT_LEX,[
  arg(AGENT, AGENT_CATEGORY,
        AGENT_LEX),
  arg(DIROBJ, DIROBJ_CATEGORY,
         DIROBJ_LEX),
  arg(MODIFIER1, HANDLE1, LEX1),
  ...,
  arg(MODIFIERm, HANDLEm, LEXm)
]).
```

The first two arguments of the predicate `ev` are the category and the lexical of the *event* (i.e. the category and the lexical of the action accomplished by the object versus the direct object). The third argument is a set of participants (agent and direct object and modifiers), expressed as list of Prolog predicates. The category of the agent (AGENT_CATEGORY), the category of direct object (DIROBJ_CATEGORY) as well as their lexical form (AGENT_LEX and DIROBJ_LEX) are included in the predicative description of the event argument (`arg`).

Linking rules should therefore describe when two news items have to be linked together. These are written over the objective representation of the investigated documents. In particular, they exploit the notion of event. Linking rules are then Prolog predicates defining a linking criteria that motivates the existence of an link among the source and the target news items from which events are derived. Linking rules define all the constraints that the participants of two events must satisfied for generating a link between them. Each generated link has therefore a *LINK_TYPE* that is determined by the application of a specific rule. In order to compile a linking rule a list of pre-defined constraints, expressed as predicates, needs to be defined. The constraints act over the basic constitutes of an event (i.e. event lexical/category, subject, object and modifiers). In particular as the event category and lexical are supposed to have a different semantic from subject, object and modifier, two type of constraints have been defined. More precisely a linking rule is a Prolog predicate of the form:

```
lrule( LINK_TYPE,
       SOURCE_EVENT_CATEGORY,
       TARGET_EVENT_CATEGORY,
       SET_OF_EVENT_CONSTRAINTS,
       SET_OF_ARGUMENT_CONSTRAINTS)
```

where:

- *LINK_TYPE*, is the type of the link that is generated by such a rule.
- *SOURCE_EV._CATEGORY* and *TARGET_EV._CATEGORY* are the category of events involved in the linking rule. For example in case of an event that relates to a meeting and another event that relates to an acquisition of stocks in that meeting, it would be useful to have a linking rule characterised by MEETING_EVENT as category of source event and BUY_EVENT as category of target event.
- *SET_OF_EVENT_CONSTRAINTS* is the set of constraints to be activated on the event category/lexical information of the source and target events.
- *SET_OF_ARGUMENT_CONSTRAINTS* is the set of constraints to be activated over the arguments of the source and target events.

Given the above description a linking rules which expresses correlation between the participants of a meeting and a company acquisition in the meeting could be:

```
lrule('Acquisition during a meeting',
       MEETING_EVENT, BUY_EVENT,
```

```
        SET_OF_EVENT_CONSTRAINTS,
        SET_OF_ARGUMENT_CONSTRAINTS)
```

The *SET_OF_ARG._CONSTRAINTS* specify relation between the participants of the meeting and those that acquire something. The *SET_OF_EVENT_CONSTRAINTS* specify the relation between MEETING_EVENT and BUY_EVENT as well as the lexicals associated to them.

3.2.2. Expressing constraints in the linking rules

The aims of the constraints are to select the properties of the participants and the properties of the event categories. These constraints compositionally build linking rules. A simple set of constraints is:

- *Category Identity*: two participants must be of the same category. This implies that two entity must belong to the same class. For example IBM and INTEL are both companies so they belong to the company category. If a Category identity constraint is included inside a SET_OF_EVENT_CONSTRAINTS, it casts different events to be in the same category. If we use this constraint leaving unspecified the event category we are grouping together event of the same category.
- *Lexical Identity*: the participants must have the same lexical e.g. the participant *Bill Gates* is the same lexical in *Bill Gates buy IBM* and in *Bill Gates get married*. A rule based on the category identity constraint would not be useful in the above case as a lot person get married. The Lexical Identity for the set of event constraint is less meaningful. However it can be used to specify the relation involved in a couple of events more precisely. For example if we have the event *Bill gates sell IBM*, its category will be *acquisition*. This information would not useful if we want build a rule for capturing document about company selling.
- *Conceptual Similarity*, it is an extension of the category identity type. In this case categories are grouped in a hierarchical structures. It is possible to express a relation of parents among participants.

Given the above constraints the following events:

```
ev(MEETING_EVENT, invite,[
  arg(AGENT, Company, Intel),
  arg(DIROBJ, person, Bill Gates),
  arg(MODIFIER, in, Seattle)
]).
ev(BUY_EVENT, acquire,[
   arg(AGENT, person, Bill Gates),
   arg(DIROBJ, company, Intel)
]).
```

two sample rules for capturing the link type *Acquisition during a meeting* are:

```
lrule('Acquisition during a meeting',
       MEETING_EVENT, BUY_EVENT,
       [],
       [lex_id(AGENT,DIROBJ)]
).
```

```
lrule('Acquisition during a meeting',
       MEETING_EVENT, BUY_EVENT,
       [],
       [lex_id(DIROBJ,AGENT)]
).
```

It is worth noticing that the above rule involves general events so the information about participants has to be more specific (i.e. lexical information about participants is needed). This pushes for the use of *lex_id* constraint.

Another generic rule is that groups document speaking about a target agent doing whatever action. For example the events in which Bill Gates buy something could be captured by the following rule:

```
lrule('Same participants rule',
       _,_,
       [cat_id()],
       [lex_id(AGENT,AGENT)]
).
```

In the above rule the only requirement is the same agent in the linking documents. The agents in the target events have to do an action of the same category type (e.g. Acquisition event, Announce event, Market strategy events,...).

When is needed grouping together documents in which a target action is carried out, it is possible to use the category constraints for the agent and object (i.e. the cat_id constraint). For example a linking rule in which agents of the same category make acquisitions of object of the same category is the following:

```
lrule('Person acquire Company',
       BUY_EVENT, BUY_EVENT,
       [],
       [cat_id(AGENT,AGENT),
        cat_id(DIROBJ,DIROBJ)]
).
```

3.3. The linking algorithm

Once the linking rules formalism has been developed it is possible to design the linking algorithm. This should takes as input the ORs of two documents: the source and the target. For each couple of events in the source and in the target, the linking rule database $LRDB$ is considered. If some rule is matched a link is generated and it is stored in a link DB. The rules are composed of some basic constraints that act on the constituents of an event. In this way, if an extended list of basic constraints is available it is possible for the user to define several linking rules. The rule can be described in an external data file so new rules can be added to the similarity model without re-designing the entire architecture.

The linking algorithm takes as input two documents, one is the source S and the second is the target T, and given their sets of events, respectively $Ev(S)$ and $Ev(T)$, check if any couple $< ES, ET >$, where $ES \in Ev(S)$ and $ET \in Ev(T)$, satisfy any of the linking rules contained in $LRDB$.

The algorithm is composed of the following steps:

```
function Link(text S, text T) returns Linkset
begin
   Linkset L = ∅ ;
   Ev(S) = BuildEv(S) ;
   Ev(T) = BuildEv(T) ;
   for each (ES, ET) ∈ Ev(S) × Ev(T)
      begin
      while ((R = SelectNextRuleFor(ES, ET)) != NULL)
         begin
         R = (RuleType, SEvCat, TEvCat, CatConstr, ArgConstr);
         if ( ApplyCatConstr(CatConstr, ES, ET) == true )
         begin
         boolean sat = true;
         while (SArg, TArg) ∈ nextArg(ES, ET) AND sat )
            sat = ApplyArgConstr(ArgConstr, SArg, TArg) ;
            if (sat)
               AddLink(ES, ET, RuleType, L)
         end
      end
   end
return L ;
end
```

4. Conclusions and future work

In this paper, we have presented a methodology for the automatic hyper-linking among news items. The presented approach is based on Information Extraction techniques that give the possibility of building semantically motivated links among documents. This approach is more expressive that the traditional approaches to the problem that allows the automatic construction of links only between related documents. The approach has been used to build the Namic prototype (EU-founded project NAMIC, News Agencies Multilingual Information Categorization, IST-99 12392).

As the approach is rather different from the pre-existing the comparison is hard. We will therefore compile, according to our definition of the task, a large test set that should enable the validation of the methodology and of the implemented system.

5. References

James Allan. 1995. Automatic hypertext construction. Technical Report TR95-1484, 13.

Roberto Basili, Maria Teresa Pazienza, and Michele Vindigni. 2000a. Corpus-driven learning of event recognition rules. In *Proc. of the Workshop on Machine Learning for Information Extraction , held jointly with ECAI2000*, Berlin, Germany.

Roberto Basili, Maria Teresa Pazienza, and Fabio Massimo Zanzotto. 2000b. Customizable modular lexicalized parsing. In *Proc. of the 6th International Workshop on Parsing Technology, IWPT2000*, Trento, Italy.

Roberto Basili, Maria Teresa Pazienza, and Fabio Massimo Zanzotto. 2002. Lerning ie patterns: a terminological perspective. In *Proc. of the Workshop on Event Modelling for Multilingual Document Linking, held jointly with 3rd LREC*, Canary Islands, Spain.

David Ellis, Jonathan FurnerHines, and Peter Willett. April 1994. The creation of hypertext linkages in fulltext documents: Parts i and ii. Technical Report RDD/G/142.

Robert Gaizauskas and Kevin Humphreys. 1997. Using a semantic network for information extraction. *Natural Language Engineering*, 3, Parts 2 & 3:147–169.

S. Green. 1997. *Automatically generating hypertext by computing semantic similarity*. Ph.D. thesis, Department of Computer Science, University of Toronto.

Steve Outing. 1996. Newspapers online: The latest statistics. *AEditor and Publisher Interactive [Online]*.

Randall H. Trigg. 1983. *A networkbased approach to text handling for the online scientific community*. Ph.D. thesis, University of Maryland.

J. Christopher Westland. 1991. Economic constraints in hypertext. *Journal of the American Society for Information Science*.