Inducing Hyperlinking Rules in Text Collections

Roberto Basili, Maria Teresa Pazienza, Fabio Massimo Zanzotto
University of Rome Tor Vergata,
Department of Computer Science, Systems and Production,
00133 Roma (Italy),
{basili, pazienza, zanzotto}@info.uniroma2.it

Abstract

Automatic hyperlinking methods based on Information Extraction techniques and on linking rules firing on salient facts have been proposed to connect documents with "typed" relations. However, the activity of defining link types and writing linking rules may be cumbersome due to the large number of possibilities. In this paper, we tackle this issue proposing a model for automatically extracting link types and, as a consequence, linking rules from large text collections. The novel idea is to exploit relations among facts expressed within documents and to use them in deciding the hyperlink types. The viability of our approach has been investigated using a collection of financial documents.

1 Introduction

Hyperlinked text collections are often seen as an added value. For instance, on-line news agencies and newspapers tend to offer news items enriched with links to the so-called "related articles" in order to better serve their customers. A journalist can write more rapidly an article for the current breaking news if he or she can easily access related facts. Similarly, a market analyst could better understand the sudden rise of a share if he or she is provided with the news items related to the acquisition activities of the involved company.

Tracing hyperlinks between documents involves the ability of finding relations among concepts or facts. This is the same ability used when writing. It is therefore an inherently difficult task. As pointed out in (Ellis *et al.* 94) the inter-agreement among the linking annotators may be very low even if they are only asked to produce links between "related texts". The disagreement may be even bigger if the relevant link types are more than one. For instance the "cause-effect" relation may be used to better help to decide if it is worthy to traverse the provided link. This would help the final users to filter out the information they are not interest in.

Computational models able to suggest automatic procedures for linking documents (e.g. (Green 97)) and for typing the drawn links (e.g. (Allan 96)) have been proposed. However, the notion of "relatedness" provided by automatic approaches such as the ones based on the bag-of-word model (e.g. (Allan 96)) or the ones based on more "semantic" model (as the lexical chains (Morris & Hirst 91) used in (Green 97)) is not sufficient to classify links in types as "cause-effect". The "cause-effect" link type is considered a "manual" link in (Allan 96) where a computational model for typing links in 6 classes is described (i.e. *revision*, *summary/expansion*, *equivalence*, *comparison/contrast*, *tangent*, and *aggregate* there called the "automatic" links). Moreover, this dichotomy ("automatic" vs. "manual" link types) suggests the perceived inherent limitations of the above automatic approaches. These boundaries may be pushed forward, as also suggested in (Allan 96), using deeper text understanding models.

In (Basili *et al.* 01) a hyper-linking method based on Information Extraction techniques conceived to connect documents with typed relations is proposed. Linking among documents is based on an intermediate representation of the documents, called the "objective representation". The objective representation is a surrogate of the document containing only events judged relevant according to an underlying knowledge-base that models the given domain. On the basis of rules firing on the event classes, a link is justified according to events (and the involved entities) appearing in the two documents. The model offers a language in which specific linking rules may be manually written building on the supported event classes. However, the activity of defining link types and writing the related linking rules may not be an easy task mainly when a large number of fact classes is foreseen.

In this paper, we want to tackle this last issue by proposing a model for the automatic definition of link types and the related linking rules in the context of a rule-based hyper-linking method.

The basic assumption we make is that the activity of building hypertexts is very similar to the process of writing. Therefore, the novel idea is to exploit the relations among facts as they appear inside the domain documents for inducing hyperlinking rules. In our opinion, the discourse structures of the domain documents are valuable resources for defining the types of relevant relations and the related linking rules.

Trusting in such an assumption we propose a light discourse model able to infer regularities in documents belonging to a collection. Co-occurrences of event classes will be used to derive link types and linking rules. As the exploration of the proposed model demands for a linguistic analysis of the texts applied over large amount of data, we refer to assessed language processing technologies. Therefore, in line with the previously referred approach we build on robust methods for processing natural language (i.e. a robust syntactic parser (Basili & Zanzotto 02) and a shallow knowledge-based method for a semantic analysis in line with approaches such as (Humphreys *et al.* 98)) and on well assessed statistical measures (i.e. mutual information). It is worth noticing that insights borrowed from the Discourse Analysis theory have also fashioned the definition of "generic" link types such as *revision*, *association*, and *aggregation* (see (Van Dyke Parunak 91)).

To better explain our model for deriving hyperlinking rules, we will firstly summarize the rule-based approach to hyper-linking (Sec. 2). Then we will introduce our light model for inspecting the discourse (based on a robust syntactic parser and a knowledge-based shallow semantic analyser) and describe how the proposed model enables the extraction of the relevant and stable relations among event types using statistical measures (Sec. 3). Finally, the viability of the approach is inspected in the financial domain represented by a document collection of news articles (Sec. 4).

2 Rule-based Hyperlinking with Information Extraction techniques

The identification of links among texts in document collections may be a very subjective matter according to the "perception" of what is useful. Therefore, a hypertext should take into account that the final users may ask for very different supports in reasoning. As it is suggested in (Glushko 89), some hypertextual links may even be misleading for a user not interested in the information they are pointing to. Assigning types to hyper-links is then suggested to overcome such a problem. From the point of view of the possible services, a "typed" text network is useful for supporting a personalised access to the linking information according to dinamically changing specific needs. The extent of the personalisation depends on the used link type system T and on the availability of automatic methods able to assign types in T to the retrieved links.

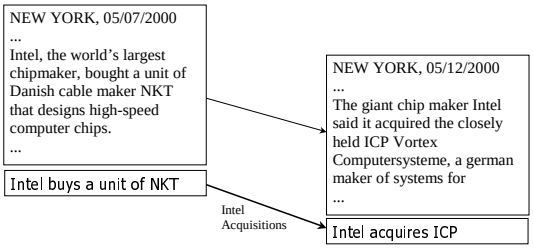

Figure 1: An example of justified link

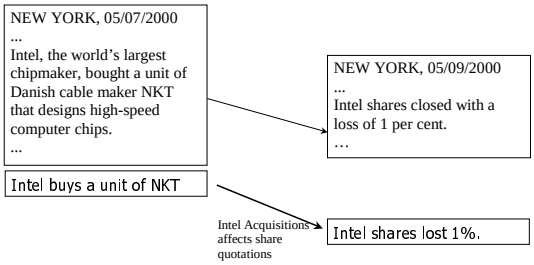

Figure 2: A complex justified link

In this context, neither the bag-of-word model (Allan 96) nor the more complex model based over the lexical chaining idea (Green 97) may not be very well suited since it supports only the notion of "relatedness" or the 6 link types expressed in (Allan 96). In fact, for example, while these approaches may relate the two news items in Fig. 1 or in Fig. 2 using the `Intel` stem, they cannot offer any further linking motivation. On the other hand, it may be relevant to clarify that the link between the two news items in Fig. 1 is motivated by the fact thar both deal with the `Intel acquisition activity` or that the link shown in Fig. 2 describes a kind of "cause-effect" relation between the two articles.

The approach to which we adhere (Basili *et al.* 01) supports the extraction of typed links. The main ideas are that: (1) the justification of a link may be found using the relevant facts carried by the two involved documents and (2) the identification of this salient information may be achieved using knowledge-based Information Extraction techniques. As a consequence, a "same fact" type can be identified in Fig. 1 by considering the two facts `Intel buys a unit of NKT` and `Intel acquires ICP`. Similarly, in Fig. 2 a "cause-effect" link can be derived using the two facts `Intel buys a unit of NKT` and `Intel shares lost 1%`. Documents may be therefore seen as sequences of relevant (and classified) facts while linking is based on rules triggered by the events identified in the two analysed documents.

The resulting document model, called "objective representation" (OR), specifies that a document d is represented as a sequence of typed events, i.e.

$$d = \{e_1, ..., e_n\} \tag{1}$$

where each event $e_i = (t_i, \{a_i^1, a_i^2, ..., a_i^n\})$ is represented by a type $t(e_i) = t_i$ and a set of arguments $p(e_i) = \{a_i^1, a_i^2, ..., a_i^n\}$ related to the participants. For instance, given the event class *buy_event*, the two news articles in Fig. 1 would have the following "objective representation":

$$\begin{aligned} d_1 = \{(buy\_event, \quad & \{agent(Intel), \\ & patient(a\_unit\_of\_NKT)\})\} \\ d_2 = \{(buy\_event, \quad & \{agent(Intel), \\ & patient(ICP)\})\} \end{aligned} \tag{2}$$

The construction of such a kind of representation from the raw text requires information extraction techniques able to exploit very precise model of the knowledge domain, i.e. a model based on extraction rules. These rules must describe the "equivalence" between events of the same type with different surface linguistic forms. For instance, Fig. 1 shows two similar events (that may be called *buy_event*) expressed with two different surface forms: the first is governed by the verb *buy* whilst the second by the verb *acquire*.

Building on this document model, a powerful language for the description of complex linking rules has been provided in (Basili *et al.* 01). Hyper-linking rules may be written as logical forms in which triggers justify firings. The prototypical rule linking two documents, d_1 and d_2, according to the link type $lt(et_1, et_2)$ (where et_1 and et_2 are two event types) is then expressed by:

$$\begin{aligned} link(d_1, d_2, lt(et_1, et_2)) \\ & e_1 \in d_1, \\ & e_2 \in d_2, \\ & link\_ev(e_1, e_2, lt(et_1, et_2)). \end{aligned} \tag{3}$$

The rule suggests that two documents can be linked according to the given linking type if there are, respectively, two events, e_1 and e_2, justifying a typed link $lt(et_1, et_2)$. The related linking rule prototype is expressed by the following:

$$\begin{aligned} link\_ev(e_1, e_2, lt(et_1, et_2)) \leftarrow \\ & t(e_1) = et_1, \\ & t(e_2) = et_2, \\ & constr(p(e_1), p(e_2)). \end{aligned} \tag{4}$$

The rule predicates that the link of the type $lt(et_1, et_2)$ between the two events e_1 and e_2 can be drawn if event types are respectively $t(e_1) = et_1$ and $t(e_2) = et_2$ and cross-constraints between the participants of the events, i.e. $constr(p(e_1), p(e_2))$, are met. Note that it is generally sufficient that the involved events share at least one participant.

Link types, $lt(et_1, et_2)$, and linking rules are strictly correlated: both are completely defined when the types of the involved events have been identified. For instance, a relevant and completely defined link type able to explain the relation in Fig. 2 is $lt(buy\_event, share\_trend)$ if *buy_event* and *share_trend* are foreseen as event types.

The rule-based hyper-linking paradigm may support automatic construction of a typed text networks. Different types of links may be defined according to the foreseen information needs. However, the definition of the hyper-linking types, i.e. $lt(et_1, et_2)$, and the related rules may still be a cumbersome problem since prototypical relations among facts may not be so evident. The problem is strictly related to the size of the set EC of the possible event classes the underlying information extraction system is able to deal with. A large EC set results in a rich language for expressing the link types and very specific rules may be foreseen. But, as the size of EC grows, the hyperlinking rule definition activity becomes more complex as the number of possible linking types to be considered grows squarely. Trivially, the

number of types $lt(et_1, et_2)$ with $et_1, et_2 \in EC$, whose relevance has to be judged, is $|EC|^2/2$. Aiming to reduce this space, we here propose a method to suggest (among all the possible ones) the more promising link types among event classes that may be easily translated into hyper-linking rules.

3 Shallow Discourse Analyis for Automatically Deriving Hyperlinking Rules

Establishing relations among facts and ideas is a relevant cognitive process. It supports not only the activity of tracing hyper-links among different texts but also the writing process. We argue that if a type of relation among events should be relevant in a particular domain it has been expressed in some of the texts belonging to the analysed collection. Therefore, an analysis of stable relations among facts within the texts may suggest the relvant types of hyper-links and, consequently, the hyperlinking rules. In this perspective, the relations expressed in the corpus are retained as relevant link types.

The model we here propose analyses the stable relations among event classes expressed in the domain texts. We will analyse the content of the texts in light of the event classes EC trying to find these correlations (i.e. the co-occurrences) among elements in EC that are more promising in the corpus C. The experimental framework will build on robust methods for processing natural language (i.e. a robust syntactic parser (Basili & Zanzotto 02)) and a shallow knowledge-based method for a semantic analysis in line with approaches such as (Humphreys *et al.* 98), and on well assessed statistical measures (i.e. mutual information).

In the model we propose, the syntactic analyser extracting the grammatical structures of a given text cooperates with the semantic analyser to map different surface representations to the implied event class. The corpus C, reduced to a set of forms as in (1), will be used to define the statistical model able to capture the stable relations among fact types by means of mutual information scores.

3.1 Robust Syntactic Parsing

We need a document representation as the (1) to capture relations between fact classes. Then, salient facts expressed in the documents need to be made explicit. The process for the recognition of the events starts from a grammatical analysis of the texts to leave out grammatical differences over the surface forms and to make evident syntactic relations between the textual fragments.

We will rely on a robust approach (Basili & Zanzotto 02) that produces partial and possibly ambiguous syntactic analysis with the available grammatical knowledge. We used the following module cascade: a *tokenizer*, matching words from character streams; a *yellow page look-up module* that matches named entities existing in catalogues; a *morphologic analyser* that attaches (possibly ambiguous) syntactic categories and morphological interpretations to each word; a *named entity matcher* that recognizes complex named entities according to special purpose grammars; a rule-based *part-of-speech tagger*; a *POS disambiguation module* that resolves potential conflicts between the results of the POS tagger and the morphologic analyser; a *syntactic parser* based on modularisation and lexicalisation: it builds a chunk-based representation of the input text, including major grammatical dependencies between chunk heads. The formalism adopted to collect syntactic information is called extended dependency graph (XDG). Merging the notions of dependencies and constituents it results in a good representation scheme for describing partial and competing structures (as discussed in (Basili & Zanzotto 02)). An $xdg = (C, L)$ is a graph whose nodes C are constituents and edges L are dependencies between constituents. Generally, due to the capabilities of the used syn-

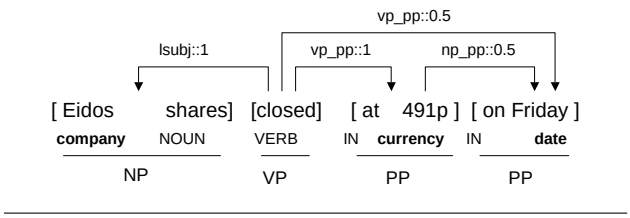

Figure 3: Example XDG over a sentence

tactic processors, chunks (Abney 96) are the constituents represented in the graph. We will provide details on the formalism through the example in Fig. 3. Chunks (i.e. the non-recursive kernels of noun phrases, prepositional phrases, and verbal phrases respectively NPK, PPK, and VPK) are shown in square brackets. Sub-components of

the chunks are typed with grammatical labels as IN (i.e. preposition), NOUN, and VERB or with named entity labels as *company*, *date*, and *currency*. Relations between constituents are shown using oriented arcs. It is worth noticing that a plausibility score is attached to each relation. Dependency types and plausibility scores are represented in Fig. 3 as TYPE::PLAUS where TYPE is the relation type and PLAUS is the plausibility score. This latter ranges in the interval (0,1] and determines the ambiguity of the syntactic dependency. The value 1 means an unambiguous link. For instance the vp_pp dependency between the VPK *[closed]* and the PPK *[at 491p]* is depicted as unambiguous (the plausibility score is 1) whilst the VP_PP one between the VPK *[closed]* and the PPK *[on Friday]* is considered ambiguous (0.5). An alternative interpretation considers an NP_PP dependency between the PPK *[at 491p]* and the PPK *[on Friday]*. No choice is done between competing interpretations.

The XDG formalism is well suited for the model we propose since it allows to easily isolate the text fragments that are interesting for our ligth discourse analysis. As we want to consider relations between the events expressed in verbal phrases, the analysis will focus on verb contexts, i.e. the verbs with their arguments. These can be straightforwardly extracted from the XDG grammatical representation selecting the subgraphs with all the constituents that can be directly reached from a verb. These contexts will be represented as pairs of elements: the verb lemma and the list of the actual arguments. For instance, the context c of the verb *close* in the sentence in Fig. 3 extracted as a subgraph of the related XDG will be:

$$c = (close, \quad \{lsubj(share\_of(company(Eidos)), \; at(currency(491p)), \; in(date(Friday))\}) \tag{5}$$

The semantic carrier with its possible structure and named entity type is retained as representing each chunk. Similarly, the two representations for the salient events (the contexts c_1 and c_2) in Fig. 1 will be expressed by the following:

$$\begin{array}{ll} c_1 = (buy, & \{lsubj(company(Intel), \; lobj(unit\_of(company(NKT))\}) \\ c_2 = (acquire, & \{lsubj(company(Intel), \; lobj(company(ICP))\}) \end{array} \tag{6}$$

The syntactic processor SP can be seen as a function mapping documents in raw texts t to syntactically analysed documents sd, i.e. $SP(t) = sd$. The syntactic representation sd of a document t is a set of verb contexts c, i.e. $sd = \{c_1, ..., c_n\}$.

3.2 Knowledge-based Semantic Analysis

Syntactically processed documents are still not completely generalised to allow the analysis of the co-occurrences for two given event types. The differences between the surface forms still exists, e.g. no explicit equivalence is stated between the verb context c_1 and c_2. In line with (Humphreys *et al.* 98), we will produce this equivalence in the semantic analysis by matching the current verb contexts with event prototypes contained in a domain knowledge base KB. This latter stores the different surface representations associated to each event type postulated in the domain (i.e. for each element in the set EC).

For the semantic analysis we propose a function S that maps documents in the syntactic space to their semantic representation, i.e. $d = S(sd)$ where sd is a document syntactically represented while d is the document seen as a sequence of salient facts. The key element of the semantic analysis is the event knowledge-base KB that contains elements, *event prototypes*, as follows:

$$(t, v, [a_1, ..., a_n]) \tag{7}$$

describing that a possible form of an event typed $t \in EC$ is a verb context governed by the verb v with a_1, ..., a_n as arguments. An instance of event prototype of the type *buy_event* may be:

$$(buy\_event, acquire, \quad \{lsubj(company(X)), \; lobj(company(Y))\}) \tag{8}$$

that is a *buy_event* may be represented in a context of the verb *acquire* if the logical subject X (referred to as $lsubj$) is a *company* and the logical object Y (referred as $lobj$) is *company*. The knowledge-base KB is then used to define the function S as in the following:

$$S(sd) = \{s(c_1), ..., s(c_n)\} \tag{9}$$

where:

$$s(c) = \begin{cases} (t, args) & if \quad \exists (t, v, args) \\ & \quad \in best\_match(c, KB) \\ none & otherwise \end{cases} \tag{10}$$

That is, a verbal context $c = (v, sargs)$ is transformed into an event $(t, args)$ of the type t if an event prototype $(t, v, args)$ is foreseen in the knowledge base KB whose arguments $args$ are matched against the arguments of the actual instance $sagrs$. If more than one prototype satisfies the constraints, the one matching the largest number of arguments is selected, i.e. $best\_match(c, KB)$. The default value *none* is assigned when the verb context does not match any event prototype.

3.3 Highlighting Stable Relations between Fact Types using Mutual Information

The proposed standardised document content model and the processing techniques introduced in the previous sections offer the possibility of studying the stable relations between event types relevant in a knowledge domain. The domain corpus C seen as the implicit domain model is in fact the sample space where to study the correlations between fact types using statistical measures.

As the final step of our model, we propose here to study this correlation using mutual information on a statistical model of the corpus. Pairs of event classes found with a positive level of mutual information will be retained as useful to build hyper-linking rules. We assume that the link type is defined by the pair of correlated events.

In order to evaluate mutual information (MI), two sample spaces S_1 and S_2 have been defined on the analysed corpus C. S_1 is the space of all the events occurred in the corpus C, i.e.:

$$S_1 = \{e | d \in C, e \in d\} \tag{11}$$

while S_2 represents all the events that jointly appeared in one of the documents, i.e.:

$$S_2 = \{(e_1, e_2) | d \in C, e_1, e_2 \in d, e_1 \neq e_2\} \tag{12}$$

The probability that the single event e is of type t can be estimated using the frequency of the events of type t in the space S_1:

$$p(t(e) = et) = \frac{freq(e \in S_1 | t(e) = et)}{|S_1|} \tag{13}$$

Similarly, the probabilities of (et_1, et_2) pairs can be estimated on the space S_2 with the following measure:

$$p(t(e_1) = et_1, t(e_2) = et_2) = \tag{14}$$

$$= \frac{freq((e_1, e_2) \in S_2 | t(e_1) = et_1, t(e_2) = et_2)}{|S_2|} \tag{15}$$

Then, mutual information $mi(et_1, et_2)$ of two event types et_1 and et_2 belonging to the event classes foreseen in the model of the knowledge domain is:

$$mi(et_1, et_2) = \log \frac{p(et_1, et_2)}{p(et_1)p(et_2)} \tag{16}$$

The mutual information in the corpus C of the pairs (et_1, et_2) makes evident the relations considered relevant by the writers of the analysed documents. The higher is the value of mutual information the wider is the perception of a relevant link between the two event classes. Event type pairs with a positive value of mutual information can be retained as relevant link type $lt(et_1, et_2)$ and easly translated to hyper-linking rule. Furthermore, in a supervised environment, the use of mutual information as index offers the possibility to rank the choices according to their relevance.

4 Analysis of the Results in a Financial Domain

Our model has been applied to a financial domain and the experiment has been carried out on a large document collection with these parameters: (1) the set of event classes EC and (2) the semantic knowledge base KB. We will hereafter refer to $EC + KB$ as the model for the domain. The event classes foreseen for this domain are shown in Tab. 1 structured into a hierarchy. These 20 classes are the same as proposed in (Basili *et al.* 02) to describe the relevant events that may occur in a financial news stream. This class hierarchy covers event types involving the activities of the companies in the market (e.g. the class 5) as well as the events that may happen in the stock exchange (i.e. the types 6, 6-1, and 6-2). On the other hand, the knowledge base KB including about 500 event prototypes has been extracted from the corpus with the semi-automatic extraction method based on a terminological perspective (Basili *et al.* 02). The terminological model enables the extraction of event prototypes from the corpus and their ranking according to the relevance score. The ranked prototypes are then shown to one or more annotators that provide a classification in one of the foreseen classes in EC. It is worth noticing that also the modular

1		RELATIONSHIPS AMONGS COMPANIES
	1-1	Acquisition/Selling
	1-2	Cooperation/Splitting
2		INDUSTRIAL ACTIVITIES
	2-1	Funding/Capital
	2-2	Company Assets (Financial Performances, Balances, Analysis Sheet)
	2-3	Market Strategies and plans
	2-4	Staff Movement (e.g. Management Succession)
	2-5	External Communications
3		GOVERNMENT ACTIVITIES
	3-1	Tax Reduction / Increase
	3-2	Anti-Trust Control
4		JOB MARKET - MASS EMPLOYMENT/UNEMPLOYMENT
5		COMPANY POSITIONING
	5-1	Position vs Competitors
	5-2	Market Sector
	5-3	Market Strategies and plans
6		STOCK MARKET
	6-1	Share Trends
	6-2	Currency Trends

Table 1: The event class hierarchy of the financial domain

syntactic parser we are using depends on a number of parameters, i.e. the rules and the lexicons it uses. In particular, due to the nature of the information we want to gather, the main parameters are the part-of-speech tagger and the verb argument detector.

Nevertheless, even with all these inherent limitations we carried out the experiments on a large corpus, i.e. a collection of about 12,300 news items of the Finacial Times published in the period Oct.-Dec. 2000. This collection has been assumed to be the implicit domain model from where to extract the domain "linking" knowledge.

After the syntactic/semantic analysis it is made evident that an average of 0.89 events for each document have been matched and this can drastically reduce the expectation of finding correlations between event types. Focussing on the 5,723 out of the 12,308 documents that had at least one detected event, the average number of events per document is more than doubled (i.e. 1.91 events/docs). This value suggests that when the document is covered by the $EC + KB$ model the correlation between fact types in EC is a significant phenomenon and, consequently, can be studied.

The mutual information between elements in EC is depicted in Tab. 3 where hyper-linking rules between the types t_1 and t_2 are referred as $lt(t_1, t_2)$. As expected only a small part (17 out of 210 possible distinct couples in $EC \times EC$) are selected and therefore presented as useful hyper-linking rules. In such a way the 90% of the possible hyper-link types (and related rules) are removed. This demonstrates that our method is a viable solution for discovering relevant link types.

An evaluation of the pairs with respect to human judgements is difficult because two event types randomly picked from EC seem to be correlated. We can only discuss the obtained list with respect to the document collection. We may observe that some event types tend to occur with events of the same type as, for instance, 2-4, 6-1, 1-1, and 2-2. This interesting fact confirms that rules connecting events of the same type can be relevant for a possible user of the linking information. The most relevant suggested relations are those connecting different event classes. Apart from the first pair in the list between a class 6-1 with its superclass 6, these relations suggest that "cause-effect" relation types are active in the corpus according to the domain model given by EC and, therefore, the related linking rules may be foreseen. The secondly ranked pair, for instance, states that a relevant relation may be drawn between the *government activities* (class 3) and the trends in a *market sector* (class 5-2). The third

(a) 20000710L129.141

....
Infogrames is understood to have offered an all-share deal, valuing Eidos shares at between 600p and 700p. **6-1(*Eidos shares closed at 491p on Friday*)**, **2-2(*valuing the loss-making company at {GBP}512m*)**, while shares in Infogrames closed at {XEU}26.20.
....

(b) 20000712L130.107

....
In the first six months of this year, Liberty reported a 4.6 per cent rise in net asset value per share to 728p. **2-2(*Pre-tax profits rose from {GBP}63.1m to {GBP}103.5m*)**, and earnings per share expanded from 12.28p to 22.96p.
....
The group, which already has a 75 per cent stake in Capital Shopping, said it was prepared to invest another {GBP}370m in the future. Capital Shopping said its investment property income for the half year rose from {GBP}56.2m to {GBP}66.8m. **6-1(*Capital shares rose 8p to 433{1/2}p*)**, while **6-1(*Liberty was 2{1/2}p up at 531{1/2}p*)**.
....

Table 2: Sample marked up documents from the Economic Corpus

Hyper-link prototype	*MI*
lt(6-1,6)	1.0574
lt(3,5-2)	0.9879
lt(2-2,6-1)	0.9513
lt(2-4,2-4)	0.8176
lt(2-2,6)	0.7785
lt(6-1,6-1)	0.5884
lt(5-2,5-2)	0.5499
lt(6,6)	0.3810
lt(1-2,1-2)	0.3462
lt(1-2,4)	0.3090
lt(2-2,2-2)	0.2793
lt(1-1,1-2)	0.1724
lt(2-4,5-1)	0.1516
lt(2-5,6-1)	0.0808
lt(2-1,2-2)	0.0363
lt(1-1,1-1)	0.0254
lt(6,5-2)	0.0210

Table 3: Hyperlinking Rules derived from the Economic Corpus

suggested rule states a relation between the *Company Assets* (2-2) and the *Share Trends* (6-1) similarly stated by the relation (2-2,6). This is reflected in the sample document (a) in Tab. 2. The *Share Trends* (6-1) is seen as the cause of the increase in the value of the company (i.e. the *Company Assets* (2-2)). On the contrary, in the document in Tab 2.(b) the opposite relation is seen, i.e. a change in the company assets *Pre-tax profits rose from* {*GBP*}*63.1m to* {*GBP*}*103.5m*) affects the company's share trend in the stock market.

5 Conclusions

Building on the basic idea that relevant link types are already expressed in the collections of domain documents, we presented a novel method for automatically deriving hyper-linking types $lt(et_1, et_2)$ and, consequently, hyper-linking rules. The proposed method based on natural language processing techniques seems to be a viable solution to address the definition of such types and rules: in fact, when documents are covered by the domain knowledge model, the stated relations between event types may be recovered. As we have seen the method we propose reduces the linking rules that have to be considered.

References

(Abney 96) Steven Abney. Part-of-speech tagging and partial parsing. In G.Bloothooft K.Church, S.Young, editor, *Corpus-based methods in language and speech*. Kluwer academic publishers, Dordrecht, 1996.

(Allan 96) James Allan. Automatic hypertext link typing. In *Proceedings of the the seventh ACM conference on Hypertext*, pages 42–52. ACM Press, 1996.

(Basili & Zanzotto 02) Roberto Basili and Fabio Massimo Zanzotto. Parsing engineering and empirical robustness. *Natural Language Engineering*, 8/2-3, 2002.

(Basili *et al.* 01) Roberto Basili, Roberta Catizone, Luis Padro, Maria Teresa Pazienza, German Rigau, Andrea Setzer, Nick Webb, Yorick Wilks, and Fabio Massimo Zanzotto. Multilingual authoring: the namic approach. In *Proceedings of the workshop on Human Language Technology and Knowledge Management, held jointly with ACL'2001 Conference*, 2001.

(Basili *et al.* 02) Roberto Basili, Maria Teresa Pazienza, and Fabio Massimo Zanzotto. Learning IE patterns: a terminology extraction perspective. In *Proceedings of Workshop of Event Modelling for Multilingual Document Linking at LREC 2002*, 2002.

(Ellis *et al.* 94) David Ellis, Jonathan Furner Hines, and Peter Willett. The creation of hypertext linkages in fulltext documents: Parts i and ii. Technical Report RDD/G/142, British Library Research and Development Department, 1994.

(Glushko 89) R. J. Glushko. Design issues for multidocument hypertexts. In *Proceedings of the second annual ACM conference on Hypertext*, pages 51–60. ACM Press, 1989.

(Green 97) S. Green. *Automatically generating hypertext by computing semantic similarity*. Unpublished PhD thesis, Department of Computer Science, University of Toronto, 1997.

(Humphreys *et al.* 98) K. Humphreys, R. Gaizauskas, S. Azzam, C. Huyck, B. Mitchell, H. Cunningham, and Y. Wilks. University of sheffield: Description of the LASIE-II system as used for MUC-7. In *Proceedings of the Seventh Message Understanding Conferences (MUC-7)*. Morgan Kaufman, 1998.

(Morris & Hirst 91) Jane Morris and Graeme Hirst. Lexical cohesion, the thesaurus, and the structure of text. *Computational linguistics*, 17(1):21–48, 1991.

(Van Dyke Parunak 91) H. Van Dyke Parunak. Ordering the information graph. In E. Berk and J. Devlin, editors, *Hypertext/Hypermedia Handbook*, chapter 20, pages 299–325. McGraw-Hill (Intertext), 1991.