Learning IE patterns: a terminology extraction perspective

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

The large-scale applicability of knowledge-based information access systems such as the ones based on Information Extraction techniques strongly depends on the possibility of automatically acquiring the large amount of knowledge required. However, the basic assumption of the IE paradigm, i.e. that the information need is known in advance, limits inherently its applicability since the resulting IE pattern learning algorithms are not generally conceived for the analysis of large corpora if not driven by a specific information need. Since in the terminological studies the corpora and not the information needs already drive the extraction of the knowledge, they offer many insights and mechanisms to automatically model the knowledge content of a coherent text collection. In this paper, we will present a terminological perspective to the acquisition of IE patterns based on a novel algorithm for estimating the domain relevance of the relations among domain concepts. The algorithm and the representation space will be presented. Before starting the discussion, however, we will describe the overall process of building a domain ontology out from a extensional domain model (i.e. the collected domain corpus). Finally, the results of the application of the algorithm over a large domain corpus will be presented and the resulting ontology is discussed.

1. Introduction

The large-scale applicability of knowledge-based information access systems such as the ones based on Information Extraction techniques strongly depends on the possibility of automatically acquiring the large amount of knowledge required. The applicability of these systems over large heterogeneous text collections (e.g. the World Wide Web) may be one of the keys of success of "emerging" information access paradigm such as the Question Answering (QA) and the Automatic Summarisation (AS). In fact, the major strength of the Information Retrieval engines (typically based on the "poor" abstraction of stem) is due more to their wide applicability than to their actual retrieval performances.

A very well assessed approach to Information Access is the paradigm of Information Extraction (MUC-7, 1997; Pazienza, 1997). This latter gave the fertile area where a number of techniques for the automatic acquisition of knowledge have been proposed. However, these learning approaches are focused on the extraction of knowledge needed for the satisfaction of a particular information need (i.e. the one expressed by the template) as the IE paradigm imposes. Therefore, the resulting learning approaches are biased by the fact that they can rely on two important hypothesis limiting their search space. From the one side, the target knowledge domain is generally small and, from the other side, the target information need is very narrow (such as missile launch event in one of the MUC conference). Therefore, the size of the resulting ontology can be kept controlled and the scope of the learning algorithms is a controlled (and small) corpus. In fact, in unsupervised learning techniques as in (Yangarber, 2001; Riloff and Jones, 1999), texts are firstly classified according to their relevance with respect to the particular information need and then particular surface forms somehow related are extracted and retained. The first step narrows the corpus that is given to the second.

However, the basic assumption, i.e. that the information need is known in advance, limits the applicability of the IE paradigm and of the resulting IE pattern learning algorithms. In fact, these latter are not generally conceived for the analysis of large corpora if not driven by a specific information need. If the goal to be achieved is the applicability in large, a different approach has to be undertaken. In such a perspective, the final information needs can not drive the learning phase that should totally rely on the corpus that has to be the source of this information, i.e. it is the final source of information that should suggest the information needs that can be satisfied. This is the typical case a information access system has to face when exposed to an uncontrolled information scenario (e.g. the Web).

Since in the terminological studies the corpus is already the major source of knowledge, they offer many insights and mechanisms to automatically model the knowledge content of a coherent text collection. Here, in fact, the corpus plays the central role of extensional model for the target domain where a domain ontology (i.e. a thesaurus) is extracted from. In this latter, terms and relations among them are generally described. The "operational" notion of *term*, i.e. that the term is the surface representation of a domain concept, allows to define two different levels of analysis: the notion of *admissible surface forms* and the notion of *domain relevance*. The target is generally the extraction of concepts conveyed by nominal phrases and the investigated relations are IS-A and PART-OF. Neverthless this terminological perspective to the extraction of IE patterns can be adopted for widening the applicability. IE patterns may be considered as domain relations among specific concepts, i.e. typical concepts of the domain and named entity classes that hold by definition the special status of domain concepts.

In this paper, we will present a novel algorithm for estimating the domain relevance of the relations among domain concepts. As for the term, the application of a terminolog-

ical approach to the problem of the discovering the domain relations among concept has to establish:

- which are the surface representations of the target relations;
- which is the estimator of the "domain importance" for the discovered relations.

The algorithm and the representation space will be presented in Sec. 4.. Before starting the discussion, however, we will describe the overall process of building a domain ontology out from a extensional domain model (i.e. the collected domain corpus) in Sec. 2. Finally, the results of the application of the algorithm over a large domain corpus will be presented and the resulting ontology discussed (Sec. 5.).

2. Building an ontology for a large-scale IE system

A large-scale IE system for a news agency should be able to scan news streams. The activity of building the needed knowledge base is therefore a huge task. However, in our opinion, this may be undertaken using some insight given by the terminology extraction practice. News streams are, in fact, coupled with a news classification scheme that can be more or less complex (cf. IPTC standards (IPTC,)). This rough or fine-grained classification over the news items allows the definition of coherent knowledge areas over which terminology extraction techniques can be helpful. Each collection of news items belonging to a class is in fact the extensional model for the underlying domain according to the classifiers.

The process of the knowledge modelling is sketched in the following. Given the corpus as model for the knowledge domain (or class) under investigation, the activities that have to be carried out for building the domain ontology are the following:

1. the definition of the named entity classes
2. a first analysis of the corpus for the acquisition of the most important concepts and relations among the concepts
3. the analysis of the extracted domain knowledge for the definition of the top "event" classes
4. the extraction of all the important concepts and relations among the concepts and their clustering under the defined event classes

For the activities 2 to 4, terminology extraction practice may be very useful with the notions of *admissible surface forms* and of *domain relevance*. The latter is a key notion that helps in showing to the ontology builder only the most relevant IE patterns (a combination of the domain concepts and domain relations). These patterns sorted according the domain relevance estimated by the importance function can drive the definition of the top event classes. The event classes elsewhere referred as "template types" will represent the knowledge the final IE system is able to make explicit over the particular domain. Finally, since IE patterns are ranked according to their importance, in the activity of clustering this guarantees that the most important events (and generally the most frequent) may be captured by the resulting IE system.

The attention on the clustering activity is somehow one of the major difference between the construction of a domain ontology for an IE system and the one of a terminological knowledge base (TKB) (or thesaurus). This is mainly because of the nature of the typical target knowledge domains. Terminology extraction is mainly conceived for giving a systematic representation of scientific or technological knowledge domains where certain terms are stable and a relatively small number of surface forms are used to convey a domain concept. On the other hand, in the news streams (the areas in which IE system has to find the information) domain concepts and, more often, domain relations are generally conveyed by more than one surface form. It is the equivalence between different event prototypes, i.e. prototypes that specifies the possible instances of the "Who? Where? What? When? Why?" events, that may make the difference.

3. Domain relations among concepts as event prototypes

Event prototypes (or IE patterns) used by IE systems to perform the activity of extracting information are very similar to what a domain relation among domain concepts may look like. Given for instance the financial domain, the prototype necessary to extract a *"sell event"* from the following news items:

Example 1 *Financial news excerpts*

(a) Eon, the German utility formed by the merger of Veba and Viag, is poised to sell its electronics arm to an Anglo-American consortium for about $2.3bn.

(b) It is understood to be near a deal to sell the Longview smelter for $150m to McCook Metals.

may have the following form:

Example 2 *Sell event prototype*

```
sell(  (agent:companyNE),
       (patient:object),
       (to:companyNE),
       (for:currencyNE))
```

i.e. a company typically sells something to a company for a certain amount of money (`currencyNE`). Here, the two named entity categories, `companyNE` and `currencyNE`, are typical concepts of the financial domain and the showed event prototype is a typical domain relation among these concepts.

Due to the difference on the perspective and on the application domain, some adjustments of the techniques developed in terminology extraction are mandatory in the IE pattern extraction problem. As suggested in the example, in IE, a major role is played by named entities. They are not important as surface forms but as generalised forms (i.e. their category). This is a major difference with the general terminology extraction where named entities are important as instances. For instance, *Newton's law* and *Zipf's law* convey very different meaning and are relevant as such and not in a generalised form `personNE's law`. The adoption of TE techniques on the IE tasks requires that named entity categories are considered as typical concepts of the domain. Admissible surface forms also consider the possibility of selecting forms with named entities (e.g. `companyNE_share` where `companyNE` is a named entity category that may be used for detecting *IBM shares* in target text).

Furthermore, in the IE perspective, the definition and the extraction of the domain relations plays a major role. Such a problem is generally neglected in the TE studies because major efforts are spent in the definition of algorithm for extracting and using catalogues for the general relations among terms such as IS-A or PART-OF (Morin, 1999; CON, 1998). The resulting methods are not suitable for the extraction of domain relations.

In order to adopt an TE perspective to the IE pattern learning these two issues have to be faced. In the following section we will present our approach to the extraction of domain relations over large collection of texts.

4. Learning domain relations from large textual collections

The approach to the extraction of domain relations should be completely corpus driven since information needs are not stated in advance. Therefore, given the corpus C, all the relations have to be analysed in order to detect the more important ones. Since the corpus should suggest the typical domain relations in the first phase of the construction of the domain model (cf. Sec. 2.), the target relations should then not to be too far from the admissible surface form as happens for the concept spotting in TE. As for the concept detection, we should then define the admissible surface forms and a function for estimating the domain importance of the given form. However, a minimal abstraction is needed to take into account the relatively free order of the participants when they appear in the actual text as in the above example (Ex. 1). In the following section (Sec. 4.1.), the admissible surface forms and their equivalence are stated and the size of the problem is estimated. On the other hand, an efficient algorithm for the estimation of the importance function based on the frequency of the relations in the target corpus is presented in Sec. 4.2.

4.1. Admissible surface forms: the size of the problem

A relation $r = (rv, (ra_1, ra_2, ..., ra_n))$ (as the one of the Ex. 2) may be represented in a number of different surface forms. Due to the fact that the corpus should suggest the important relations, we will only consider the realisation of r in verbal phrases. The corpus C s then seen as a collection of verb contexts $c = (v, (a_1, a_2, ...a_n))$ where v is the governing verb and each argument a_i is a couple (g_i, c_i) representing its grammatical role g_i (e.g. subject, object, pp(for), pp(to), etc.) and the concept c_i semantically governing it. A context $c \in C$ is a positive example of the target relation $r \in R$ if $rv = v$ and r partially cover c, i.e. the arguments of r should then appear in any order in the context c.

Given the domain corpus C represented as a collection of verb contexts, the objective is to evaluate the relevance of each possible relation $(r, (ra_1, ra_2, ..., ra_n))$. The first problem is to estimate how many different relations have to be analysed. This may be obtained after partitioning the corpus C according to the verb governing the contexts. For each verb v, a subset of the corpus is then defined as:

$$C(v) = \{(a_1, ..., a_n) | (v, (a_1, ..., a_n)) \in C\} \quad (3)$$

Notice that the notion of context that we use is open to two different 'views'. A lexicalized notion of context is obtained by relying on the full definition. A context $c = (v, ((g_1, c_1), (g_2, c_2), ..., (g_n, c_n)))$ expresses the governing verb v with the lexical (c_i) and its syntactic role (g_i) for each argument found within a given corpus fragment. c_i is usually a partially generalized surface form. c_i denote thus partially generalized surface forms like `companyNE` (for fragments like *IBM*, *Financial Times*, *Apple Ltd.*) or `companyNE_shares` for structures like *IBM's shares*. If we neglect this rich $lexical$ information, and make use a generic concept (e.g. `object`) for the arguments, the remaining information is purely syntactic, making explicit only the grammatical role in the context:

$$c = (v, ((g_1, object), (g_2, object), ..., (g_n, object)))$$

As a result the following two sets of arguments in contexts of $C(v)$ remain defined:

$$A_\Lambda(v) = \{a | \exists (a_1, ..., a_n) \in C(v) \wedge \exists i . a_i = a\} \quad (4)$$

$$A_\Sigma(v) = \{ \begin{array}{l} (s, object) | \exists i . g_i = s \wedge \\ \exists ((g_1, c_1), ..., (g_n, c_n)) \in C(v)\} \end{array} \quad (5)$$

Given the above sets, $A_\Lambda(v)$ and $A_\Sigma(v)$, the set $R(v)$ of the possible relations for a given v is the following:

$$R(v) = \bigcup_{i=1...MC(v)} R_i(v) \quad (6)$$

where $R_i(v)$ are the collection of individual combinations of exactly i arguments in the set $A(v) = A_\Lambda(v) \cup A_\Sigma(v)$ that are syntactically meaningful. The distinction between lexicalised and syntactic arguments is useful to take into account the fact that some relations may have a recurrent syntactic argument whose filler concept is not recurrent.

If $R(v)$ is the set of all the relations for the investigated verb v, the domain importance of each $r(v) \in R(v)$ should be assessed. Therefore, at least the evaluation of the frequency of the relation $r(v)$ over the corpus $C(v)$ has to be used.

Given the defined sets, the size of the $R(v)$ set is, in the worst case, the following:

$$|R(v)| = \sum_{i=1...MC(v)} \left(\begin{array}{c} |A(v)|+i-1 \\ i \end{array} \right)^1 \quad (7)$$

where $MC(v)$ is the maximum context size for the verb v in $C(v)$. It is worth noticing that $|R(v)|$ values lie in a very large range, due to the size of $A(v)$. In the next section we concentrate on a measure of relevance (for the target domain) that allows to systematically reduce the size of the space where pattern selection is applied for each verb v.

4.2. Estimating the importance: Counting efficiently instances of event prototypes

Given the corpus C, the space of the possible relations is huge. This inherent complexity is the result of tackling the argument order freedom that is neglected in (Yangarber, 2001). In order to tackle with the problem, an informed exploration strategy may be settled. This strategy can not take advantage on the biasing given by the awareness of the final information need that is typical of the IE pattern extraction algorithm. However, some observations may be useful for the purpose:

- the target of the analysis is to emphasize the more important relations arising from the domain corpus
- the frequency of a specific relation strictly depends on the frequency of a more general relation

A very simple but effective domain relevance estimator is represented by the frequency of the relation in the corpus. In this perspecitive, the more important relations are the more frequent. Therefore, the above considerations may reduce the complexity of the search algorithm if only promising relation are explored, i.e. patterns whose generalisations are over a frequency threshold.

The idea is then to drive the analysis using the pattern generalisation that may be obtained projecting the patterns on their "syntactic" counterpart. The projection $\widehat{\Sigma}(r)$ of the relation r over the syntactic space Σ is defined as follows:

$$\widehat{\Sigma}(r) = (\widehat{\Sigma}(ra_1), ..., \widehat{\Sigma}(ra_m))$$

where $\widehat{\Sigma}(ra_i) = ra_i$ if ra_i is a "syntactic" argument ($ra_i \in A_\Sigma(v)$) or $\widehat{\Sigma}(ra_i) = (s_i, object)$ if $ra_i = (g_i, c_i)$ is a lexicalised argument ($ra_i \in A_\Lambda(v)$). The resulting search space $R_\Sigma(v) = \{\widehat{\Sigma}(r) | r \in R(v)\}$ is greatly smaller than $R_\Sigma(v)$ since $|A_\Lambda(v)| >> |A_\Sigma(v)| = \#preposition + 2$. This search space can be used for the extraction of the more promising generalised relations. This subset $\overline{R_\Sigma}$ can be used for narrowing the search space of the following step. In fact, when the acceptance threshold is settled, the resultant admissible relations are confined in the following set:

$$\overline{R}(v) = \{r | \widehat{\Sigma}(r) \in \overline{R_\Sigma}(v)\} \quad (8)$$

[1]Notice that, in syntactically meaningful contexts, arguments may appear with multiplicity higher than 1, so that the factorial expression is a useful approximation.

The overall domain importance estimation procedure may take also advantage from the fact that the order of the relation arguments may be fixed after the analysis of the promising syntactic patterns. The final counting activity can be thus performed with a simple sorting algorithm with the $O(nlog(n))$ complexity. In this case n is directly related to the number of context samples in the corpus $C(v)$. The procedure is sketched in the following:

procedure SelectAndRankRelations($R(v)$,$C(v)$)
begin
Select $\overline{R_\Sigma}(v) = \{r \in R_\Sigma(v) | hits(r, C(v)) \geq K\}$;
Set $L = \emptyset$;
for each $r \in \overline{R_\Sigma}(v)$
$L := L \cup prj(C(v), r)$;
$RankedR(v)$:= CountEquals(L);
return $RankedR(v)$;
end

where $hits(r, C(v))$ is the number of instances of the relation r in $C(v)$ e $prj(C(v), r)$ is the projection of the contexts in $C(v)$ on the syntactic relation r. The procedure $CountEquals(L)$ using a standard sorting algorithm counts the repetition of each element in L. Finally, $RankedR(v)$ is the set of couples (f, r) where f the frequency of the relation $r \in \overline{R}(v)$ on the corpus.

5. A case study: IE patterns for the financial domain

The above methodology has been applied for the definition of an ontology for a financial domain. The ontology construction steps have been followed. Firstly, an homogeneous collection of texts has been prepared as the model for the target domain, namely a collection of 13,000 news stories of the *Financial Time* over a period of time ranging from 2000 to 2001. The corpus will be hereafter called $FinTimeNews$. The analysis of the corpus has been carried out with the Chaos robust parser (Basili et al., 2000).

In the tables 1 and 2, excerpts of the lists related to the complex concepts and the relations governed by the verb *to make* are respectively shown. The lists are sorted according to their frequency in the *FinTimeNews* corpus (f in the tables). A manual assessed domain relevance is then reported (DR in the tables). The rate of the complex concepts retained as useful exceeds the 60% in the presented top 50 positions. It is worth noticing that many of the complex concepts that have not been judged important for the domain are in fact relevant time indicator. These are not useful for understanding the nature of the domain knowledge but they are precious in the perspective of a IE system for the characterisation of the time stamp of the event. Some of these expression such as `first_half` are in any case typical of the financial jargon, in particular they are used in the declaration of the companies' economic performance.

In the case of the relations governed by the verb *make*, the number of domain relevant relations in the top 50 is around 28%. The other presented relations are generally phraseological use of the same verb.

The sorted lists allows the definition of the top level hierarchy of the possible events in the financial domain.

f	Surface form	DR
2924	last_year	
1739	chief_executive	√
1138	last_week	
1086	next_year	
956	percentNE_stake	√
946	entityNE_share	√
834	last_month	
737	oil_price	
687	joint_venture	√
641	first_half	
631	pre-tax_profit	√
618	interest_rate	√
583	entityNE_yesterday	
575	entityNE_company	√
551	stake_in_entityNE	√
499	prime_minister	√
453	first_time	
438	entityNE_market	√
431	entityNE_index	√
429	earnings_per_share	√
413	share_in_entityNE	√
412	mobile_phone	
396	profit_of_currencyNE	√
374	next_month	
361	second_quarter	
358	entityNE_official	
348	second_half	
341	few_year	
341	same_time	
337	entityNE_government	√
332	next_week	
318	last_night	
316	percentNE_rise	√
316	end_of_the_year	
309	end_of_dateNE	
299	entityNE's_share	√
291	economic_growth	√
285	recent_year	
281	loss_of_currencyNE	√
281	central_bank	√
275	entityNE_deal	√
269	percentNE_increase	√
267	percentNE_stake_in_entityNE	√
248	public_offering	√
240	executive_of_entityNE	√
237	net_profit	√
234	past_year	
234	entityNE_economy	√
230	acquisition_of_entityNE	√
229	entityNE_shareholder	√

Table 1: Complex concepts in $FinTimesNews$

f	Surface form	DR
150	*(make,[(dirobj,sense)])*	
132	*(make,[(dirobj,money)])*	√
121	*(make,[(dirobj,profit)])*	√
118	*(make,[(dirobj,decision)])*	
108	*(make,[(for,entityNE)])*	
106	*(make,[(dirobj,sense),(subj,null)])*	
102	*(make,[(in,locationNE)])*	
100	*(make,[(to,entityNE)])*	
100	*(make,[(dirobj,null),(for,entityNE)])*	
95	*(make,[(subj,company)])*	√
87	*(make,[(dirobj,acquisition)])*	√
83	*(make,[(for,null),(subj,entityNE)])*	
81	*(make,[(dirobj,null),(to,entityNE)])*	
80	*(make,[(dirobj,null),(in,locationNE)])*	
79	*(make,[(dirobj,progress)])*	√
76	*(make,[(in,entityNE)])*	
75	*(make,[(dirobj,null),(subj,company)])*	√
71	*(make,[(subj,locationNE)])*	
71	*(make,[(dirobj,use)])*	
71	*(make,[(dirobj,difference)])*	
66	*(make,[(dirobj,use),(of,null)])*	
65	*(make,[(subj,entityNE),(to,null)])*	
60	*(make,[(dirobj,offer)])*	√
57	*(make,[(subj,null),(to,entityNE)])*	
57	*(make,[(dirobj,null),(in,entityNE)])*	
55	*(make,[(dirobj,profit),(subj,null)])*	√
55	*(make,[(dirobj,null),(subj,locationNE)])*	
54	*(make,[(dirobj,effort)])*	
53	*(make,[(in,locationNE),(subj,null)])*	
53	*(make,[(dirobj,currencyNE)])*	√
51	*(make,[(dirobj,mistake)])*	
50	*(make,[(dirobj,null),(subj,entityNE),(to,null)])*	
49	*(make,[(dirobj,debut)])*	√
48	*(make,[(for,entityNE),(subj,null)])*	
48	*(make,[(dirobj,money),(subj,null)])*	√
48	*(make,[(dirobj,bid)])*	√
47	*(make,[(dirobj,locationNE)])*	
46	*(make,[(on,null),(subj,entityNE)])*	
45	*(make,[(dirobj,null),(for,entityNE),(subj,null)])*	
45	*(make,[(dirobj,entityNE),(dirobj2,null),(subj,null)])*	
45	*(make,[(dirobj,difference),(subj,null)])*	
44	*(make,[(dirobj,sense),(subj,it)])*	
42	*(make,[(dirobj,progress),(subj,null)])*	
42	*(make,[(dirobj,decision),(subj,null)])*	
41	*(make,[(dirobj,investment)])*	√
40	*(make,[(dirobj,payment)])*	√
39	*(make,[(dirobj,case)])*	
38	*(make,[(dirobj2,currencyNE)])*	
37	*(make,[(dirobj,contribution)])*	
35	*(make,[(with,entityNE)])*	
35	*(make,[(dirobj,loss)])*	√

Table 2: Relations governed by the verb *to make* in $FinTimesNews$

These have been defined as follows:

1. Relationships among companies
   (a) Acquisition/Selling
   (b) Cooperation/Splitting
2. Industrial Activities
   (a) Funding/Capital
   (b) Company Assets (Financial Performances, Balance Sheet Analysis)
   (c) Staff Movement (e.g Management Succession)
   (d) External Communications
3. Company Positioning
   (a) Position vs. the competitors
   (b) Market Sector
   (c) Market Strategies
4. Governamental Activities
   (a) Tax Reduction/Increase
   (b) Anti-trust Control
5. Job Market - Mass Employment/Unemployment
6. Stock Market
   (a) Share Trends
   (b) Currencies Trends

Once the definition of the top level events has been completed, the discovered event prototypes have been manually clustered according to their class. To give the flavour of the information contained in the produced knowledge base, in the following an excerpt of the event prototypes of the *Company Assets* class are presented:

Company Assets Event Prototypes
(cut,[(subj,entityNE),(dirobj,cost)]))
(rise,[(subj,profit),(to,currencyNE)])
(rise,[(from,currencyNE),(subj,profit),(to,currencyNE)])
(issue,[(subj,entityNE),(dirobj,profit_warning)]))
(suffer,[(subj,entityNE),(dirobj,loss)])
(report,[(subj,entityNE),(dirobj,loss_of_currencyNE)])
(announce,[(subj,entityNE),(dirobj,loss_of_currencyNE)])

The analysis of $1,100$ patterns give rise to 229 patterns retained as useful for the definition of the event prototypes in one of the give class.

6. Conclusions and future work

In this paper we presented a terminological perspective to the extraction of IE patterns. This corpus driven method is more suitable for a wide application of IE-based systems with respect to learning methods driven by the specific information need. The presented method helps in performing the activities required for building a domain ontology since the concepts and the relations are presented according to their relevance for the target domain.

Many issues are still open and are objective of further research. First of all, a more complete evaluation of the method should be performed with respect to the task of

event recognition. The acquired ontology should be evaluated in order to understand if the level of detail of the event prototypes is deep enough for the experts to classify the event prototypes in the correct class. Therefore, we intend to study the possibility of automatically cluster the event prototypes once the domain top level hierarchy has been defined. We will try here to adopt a booting algorithm and we will study the size of the necessary booting data. Finally, domain relations (i.e. IE patterns) not headed by verbs may be an interesting area of research.

7. References

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

1998. In Didier Bourigault, Christian Jacquemin, and Marie-Claude L'Homme, editors, *Proceedings of the First Workshop on Computational Terminology COMPUTERM'98, held jointly with COLING-ACL'98*, Montreal, Quebec, Canada.

IPTC. Iptc standards. In *www.iptc.org*.

Emmanuel Morin. 1999. *Extraction de liens sémantiques entre termes à partir de corpus de textes techniques*. Ph.D. thesis, Univesité de Nantes, Faculté des Sciences et de Techniques.

MUC-7. 1997. Proceedings of the seventh message understanding conference(muc-7). In *Columbia, MD*. Morgan Kaufmann.

Maria Teresa Pazienza. 1997. *Information Extraction. A Multidisciplinary Approach to an Emerging Information Technology*. Number 1299 in LNAI. Springer-Verlag, Heidelberg, Germany.

Ellen Riloff and R. Jones. 1999. Learning dictionaries for information extraction by multi-level bootstrapping. In *Proceedings of the Sixteenth National Conference on Artificial Intelligence (AAAI-99)*.

Roman Yangarber. 2001. *Scenario Customization for Information Extraction*. Ph.D. thesis, Courant Institute of Mathematical Sciences, New York University.