Web-based information access: Multilingual Automatic Authoring

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

Abstract

The needs for managing similar documents in different languages increases with the growing amounts of electronic information available in documents of the same type (e.g. news streams). This paper proposes a viable approach to information access emphasizing the hypertextual paradigm in a multilingual framework. This task of processing/structuring text so that cross-lingual hypertext links are generated will be called Multilingual Authoring (MA). Methods from Natural Language Processing, especially Information Extraction, to both monolingual and Multilingual Authoring will be described and a general architecture for MA will be defined. Effectiveness of the proposed approach will be discussed the description of the NAMIC prototype system[1]

1 Introduction

Access to the growing amounts of multilingual information is made critical by the widespread adoption of distributed and multimedia technologies. The Web offer to large communities of naive users relevant information in different languages. This is sparse in a heterogeneous space difficult to govern by means of existing search technologies. Selecting, filtering and managing multilingual streams of news is even more critical for international information providers.

From one side, traditional Information Retrieval (IR) approaches are too general largely because of the adopted shallow matching techniques. User is too often forced to read a significant number of irrelevant documents before reaching interesting information. On the other side, traditional NLP technologies, like Information Extraction (IE) [10],[15] approaches, are very specific and too biased to restricted sets of information.

Automatic Authoring aims to create hypertextual organizations of (possibly multilingual) documents. This kind of information is an 'added value' to the information implicitly embodied in the text and it is not in contrast with other retrieval paradigms. Multilingual Authoring is the activity of *processing* documents, *detecting* and *extracting* relevant information from them and, accordingly, *organising source texts in a non-linear fashion*. The result is an information access paradigm that is a good example of how IR and IE methods can be improved via their integration. In Automatic Authoring (AA), the hypertextual structure provide navigation guidelines. The final user is responsible for crossing links after evaluating the relevance of the system suggestions.

Of course the challenge of AA is not free from problems. While IE systems like the ones participating in the Message Understanding Conference (MUC, see [14]) are oriented towards specific phenomena (e.g. *joint ventures*) in restricted domains, the scope of Automatic Authoring is wider. The size of the required IE is thus very large in AA. This requires more flexible architectures where more general information is required to lead the matching and extraction phases. This would allow higher coverage and a not-so-small precision. Notice that in the AA framework, large-scale knowledge bases are also required for modeling either linguistic and world knowledge.

Moreover, when faced with multilingual information homogeneous representations should be provided. Semantic information should be extracted from texts and represented in a language independent form, so that the authoring process can transparently apply to any text (whatever its source

[1] NAMIC is a HLT EU-funded project, devoted to the Multilingual Authoring of news agency text. EFE and ANSA, the major news agencies in Spain and Italy respectively, and the Financial Times are all members of the NAMIC consortium.

language is). Making the fact extraction stage in IE, a language neutral process has been already studied (e.g. [1]), but over small domains. A multilingual IE applied to AA requires large scale (lexical and ontological) knowledge bases harmonized throughout the different languages.

In this paper an architecture for Multilingual Automatic Authoring (MA) is presented based on knowledge intensive and large-scale Information Extraction. The general architecture is presented capitalising robust methods of Information Extraction [6] and large-scale multilingual resources (e.g. EuroWordNet [16]). The system is developed within a HLT European project, called NAMIC (News Agencies Multilingual Information Categorisation)[2].

Section 2 will introduce the main notion of multilingual automatic authoring as proposed by this paper. Section 3 will define the principles behind the NAMIC approach to authoring, while describing details of the proposed architecture. Section 4 will motivate the strengths (and benefits) that makes the approach viable on a large scale.

2 Authoring

2.1 Automatic Authoring

The complexity of Multilingual Automatic Authoring (MA) requires a suitable decomposition:

- **Text processing** requires at least the detection of morphosyntactic information characterising the source texts: recognition, normalisation, and assignment of roles is required for the main participants for the different events/facts described
- **Event Matching** is then the activity of selecting the relevant facts of a news article, in terms of their general type (e.g. selling or buying companies, winning a football match), their participants and their related roles (e.g. the company sold or the winning football team)
- **Authoring** is thus the activity of generating links between news articles according to relationships established among facts detected in the previous phase.

For instance, a company acquisition can be described in news items as:

1. *Intel, the world's largest chipmaker, bought a unit of Danish cable maker NKT that designs high-speed computer chips ....*
2. *The giant chip maker Intel said it acquired the closely held ICP Vortex Computersysteme, a German maker of systems for ....*
3. *Intel ha acquistato Xircom inc. per 748 milioni di dollari.*
4. *Le dichiarazioni della Microsoft, infatti, sono state precedute da un certo fermento, dovuto all'interesse verso Linux di grandi ditte quali Corel, Compaq e non ultima Intel (che ha acquistato quote della Red Hat) ...*

The above news items (1-4) deal with facts in the same area of interest of (potentially large classes of) readers. Links should be provided to support fast access via browsing to all these facts and suggest the underlying motivations. The criterion here used to decide whether or not to create (and use) a link is that all refer to *Intel acquisitions*.

Notice that a link generation process based only upon words would use common words (i.e. the proper noun *Intel* as potential anchor in linking) resulting in a huge set of potential matches. Such a connectivity would bring more noise than information in the user navigation phase.

It is important to stress that the relevant information concerning Intel is mainly related to the following kernel information in the examples 1,2 above:

1' *Intel* ***buys*** *a unit of NKT*

2' *Intel* ***acquires*** *ICP Vortex.*

Suitable links seem characterized by the equivalence between senses of *bought* and *acquired*. Mechanisms like query expansion or thesauri of synonyms (e.g. WordNet [13]) are highly affected by word polisemy and noise. The contextual information, i.e. grammatical and semantic role, is critical here. *Intel* as '*agent*' and *NKT* or *ICP Vortex* as the sold companies motivates the relatedness. In fact, news telling facts like *Intel buys silicon* represents irrelevant information for the user class that is a target of the linking process. Such unwanted sense of the verb *buy* should thus be distinguished.

The example semantic descriptions although very shallow provide a core information able to support relatedness judgments among documents (i.e. among the mentioned events). If such basic event descriptions are available links can be traced when enough relatedness can be detected. In this way, the authoring problem is thus a side effect of the overall language-processing task.

The example suggests that the decomposition suggested in the beginning of this section includes mandatory steps. First *text processing* is the responsible of the morphosyntactic recognition, building grammatical structures (i.e. graphs) out from sentences. Notice how co-reference resolution (see the role played by the pronoun *it* in the second example that co-refers the subject *Intel* in the key part of the sentence) is also useful. The capability of interpreting the different grammatical relations, resolving potential coreferences and mapping syntactic structures in event descriptions is under the responsibility of the *event matching* phase. In order to derive interpretation (i.e. *events*) from syntactic representations, references to a target ontology are required. In such an ontology, equivalence among facts (e.g. *buying*

[2] See `http://namic.itaca.it`.

companies) is represented. For instance, the relation among *buy* and *acquire* can be encoded under a more general notion of *financial acquisition*. Ontologies thus *define* the set of relevant facts in a target domain. A *financial acquisition* or *hiring of players* are examples of relevant event types in *corporate industrial* and *sports* news, respectively.

Conceptual differences among facts (detected during event matching) motivate a selective notion of hyperlinking. Links are generated during the latter *automatic authoring* phase. They are ontologically justified by the underlying conceptual representations: link types like *same financial acquisition*, *same person*, or *same company* are defined for links. They keep links separated by class and serve as explanations available to the user in the navigation phase.

2.2 Multilingual Automatic Authoring

Most of the above-mentioned phase for automated linking require language neutral information (i.e. conceptual and not simple lexical constraints). Notice that from a multilingual perspective (i.e. to establish links among news in different languages), the full-text approaches to linking can rely only on language independent phenomena (e.g. proper nouns like *Intel*). Unfortunately these are very limited and not comprehensive in texts.

Again principled representations made available by IE processes (i.e. templates) provide a viable solution. If event descriptions (i.e. templates instantiated from news in different languages) are made available over a uniform semantic formalism, this unified representation can multilingually activate linking. At a conceptual level no difference should exist between English, Spanish or Italian instances. *Intel* ***ha acquistato*** *Xircom inc.* can be derived as a kernel information of a **financial acquisition event** (see the example 3 in the previous section). If a unified representation of roles and concepts is here used, the event type, the fact that *Intel* is the 'Agent' are also available and links can be decided as much as in the monolingual case. This makes the authoring a language independent process.

The described framework poses some challenges. First, the *size* of the *ontological resources* required in terms of taxonomic (i.e. `IS_A` relations) and conceptual information (i.e. classes of events and implied participant-event relations) can be very large. Moreover, *availability of language-specific lexical interfaces* to the ontology, for the different involved languages is not trivial. Differences in the linguistic realisations of events should be modeled in such lexicons that are by no means small. Finally, the required task-dependent knowledge, that defines the set of useful events for the user community, is hard to be designed and encoded.

In the following section, a complex architecture is proposed to tackle the above problems.

3 Multilingual Authoring in the NAMIC system.

3.1 The NAMIC Architecture

The complexity of the AA framework proposed in the previous section requires a modular architecture where robust Information Extraction for *text processing* and *multilingual event matching*, and linking are integrated. In NAMIC, the aim is to extract relevant facts from the news streams of large European news agencies, to provide hypertextual structures within each (monolingual) stream and then produce cross-lingual links between streams. The NAMIC system is a distributed object oriented system where services (e.g. text processing or Multilingual Authoring) are provided by independent components and asynchronous communication is allowed. All the servers are (Java) objects within a CORBA architecture integrating libraries written in different languages (e.g. C, C++, Prolog, and Perl). The communication interfaces among the components are specified via XML DTDs, reflecting current standards (e.g. NewsXML and IPTC subject codes [8]) within the news business process.

Independent news streams for the different languages (English, Spanish, and Italian) are input to specific processors (LPs), responsible for text processing and event matching of independent text units in each stream. LPs produce an *objective representation* (see Fig. 1) for each source texts, including the detected morphosyntactic information. A modular and lexicalised shallow morpho-syntactic parser [4] is responsible of providing name entity matching and extracting dependency graphs from source sentences. Topical categorisation ([2]) of each news is also carried out at this stage according to news standards (IPTC classes). The description of the relevant events in a canonical (language neutral) form is called *objective representation* (OR). An OR includes the set of relevant information described in the news different from any subjective use of the same text made by any generic users. The later authoring activity is based on this canonical representation. In particular a monolingual linking process is carried out within any stream by the three monolingual *Authoring Engines* (English AE, Spanish AE, and Italian AE). A second phase is foreseen to take into account links across streams, i.e. multilingual hyper-linking: a *Multilingual Authoring Engine* (M-AE) is here foreseen. The main difference between the monolingual ande cross-lingual authoring is the more granular set of constraints that can be used in the former task and the selection among the languages possible in the second one.

Figure 1 represents the overall flow of information. The Language Processors are composed by a morphosyntactic (Eng, Ita and Spa MS) and an event-matching component

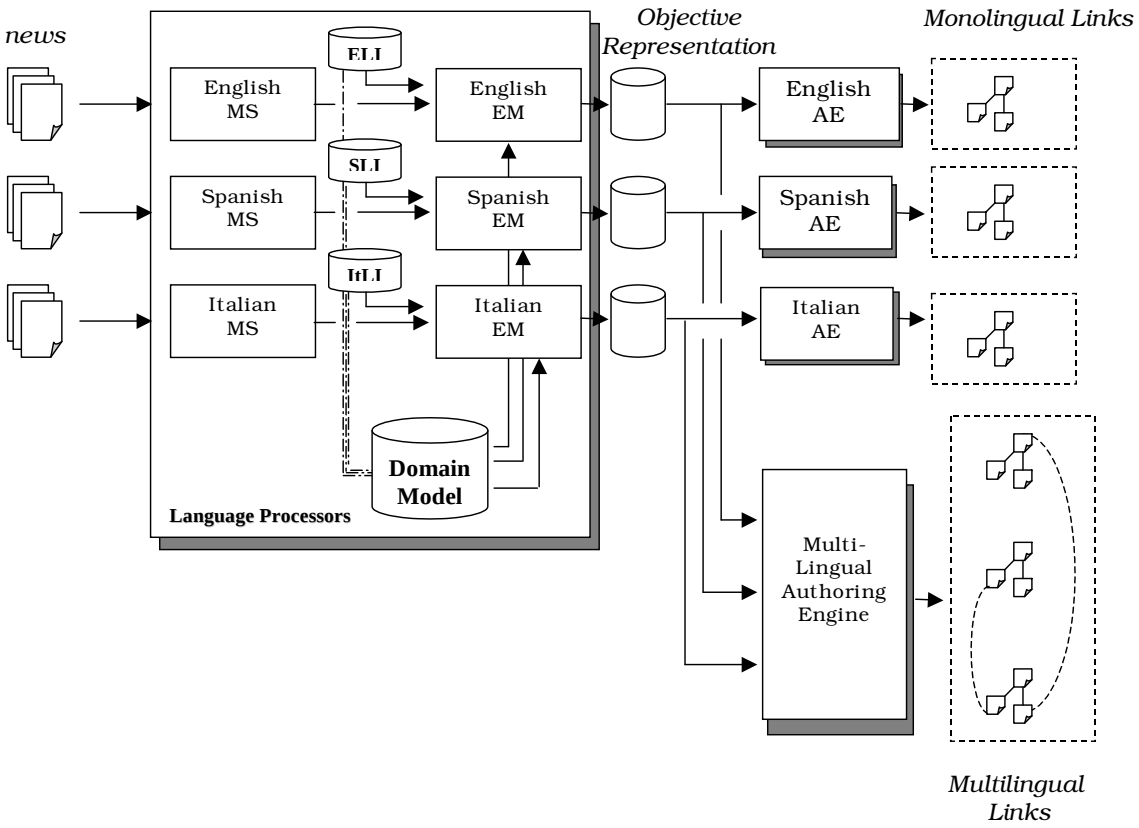

Figure 1. Namic Architecture

(EM). Notice that the lexical interfaces (ELI, SLI and ItLI) allow reference to a common (i.e. unified Domain model) that includes all the ontological information required during event matching.

The next section will add some details on the information extraction process while Section 3.2.1 will describe the proposed multilingual ontological representation that allow event descriptions to be shared among languages.

3.2 Multilingual Information Extraction.

The key components of an IE system are *events* and *objects*, i.e. facts and participants that justify hyperlinks in automatic authoring (AA). Coreference is also a necessary component in Authoring as Named Entities and their referent expressions bring important information in AA. In IE, the notion of *world model* has been used as an ontological representation of events and objects for one (or more) domain(s). The world model describes event and object types, with attributes. Event types characterise a set of events and are usually expressed in a text via verbs. Object Types on the other hand, are best thought of as characterising a set of domain entities and usually represented in a text by nouns (both proper and common). In Information Extraction, instances of each type are inserted/added to the world model and those instances that refer to the same thing are linked in coreference resolution.

In NAMIC, the world model is created using the XI cross-classification hierarchy [9], proposed for LaSIE [11]. The resulting XI-based hierarchy is referred to as an ontology. It associates nodes in the ontology with attributes and supports inheritance. Processing a text works by populating the initially bare world model with the various instances and relations mentioned in the text and converting it into a discourse model specific to the particular text. The entire process of *event matching* in NAMIC is thus designed as a LaSIE-like discourse processing task.

The Discourse Processor module maps the semantic representation produced by the morphosyntactic component (MS) into a representation of instances, their ontological classes and their attributes. An effective coreference algorithm is then applied to attempt to resolve, or in fact merge, the newly added with the current instances. Merging involves the removal of the least specific instances (i.e. the

highest in the ontology) and the merging of all known attributes. This results in a single instance (multiply realised in the texts) with several attributes and active relations in the texts. Events are then matched against the known relations involving verbs (i.e. realisations of event types) and some object types as participants.

3.2.1 The Ontological Information

Notice how the large sets of object and event types used in the NAMIC discourse processing component should be shared among different languages. EuroWordNet [16] has been used as a common semantic formalism. The NAMIC ontology thus consists of 40 predefined object types extended with nearly 1,000 objects that correspond to EuroWordNet Base Concepts [16].

EuroWordNet [16] is a multilingual lexical knowledge (KB) base with wordnets for several European languages (Dutch, Italian, Spanish, German, French, Czech and Estonian). The wordnets are structured in the same way as the American wordnet for English WordNet developed at Princeton [13] in terms of *synsets* (sets of synonymous words) with basic semantic relations between them.

Each wordnet represents a unique language-internal system of lexicalisations. In addition, the wordnets are linked to an Inter-Lingual-Index (ILI), based on the Princeton WordNet 1.5. Via this index, sysnsets in one language are mapped to the 1.5. version. Languages are thus interconnected so that it is possible to go from words in one language to words in any other language via their meaning and the mapping through the ILI. In the index a subset of 1,024 Base Concepts (BC) is represented.

In NAMIC Base Concepts are included in the world model as object types, thus providing a common semantic framework for the common nouns of all the involved languages. Eurowordnet supports (1) lexical semantic inferences (e.g. generalisation, disambiguation among meanings), (2) broad multilingual (lexical and semantical) coverage and (3) a common interlingual platform for event representation.

Once the event type relevant for a domain can be expressed via relations among ontological concepts (i.e. object types) a shared formal expression can be built from the facts matched in documents. Multilingual Automatic linking (modules AE and M-AE in Figure 3.1) is thus a side-effect of the overall IE process.

4 Why multilingual authoring is viable in NAMIC?

The traditional limitations of a knowledge-based information extraction system such as LaSIE has been the need to hand-code information for the world model - specifically relating to the event structure of the domain. This is also valid for NAMIC. At this purpose a semi-automatic booting process has been applied to develop the event type component of the world model. To us, event descriptions can be categorised as a set of regularly occurring verbs within our domain, complete with their subcategorisation information.

These verbs can be extracted with simple statistical techniques and are, for the moment subjected to hand pruning. Once a list of verbs has been extracted, subcategorisation patterns can be generated automatically using machine learning techniques (i.e. Galois lattices as described in [3]).

The lattice derived by the technique proposed in [3] represents patterns whose semantic constraints are expressed via synsets in the WordNet 1.5 base concepts. As an example, `(buy, [Agent:Company,Object:Company])` expresses the knowledge required for matching sentences like "*Intel buys Vortex*". `Company` is a base concept in Wordnet shared among the three languages and reachable via the Inter-Lingual-Index. It is thus included in the world model object hierarchy as described in the previous section. These Base Concepts play the role of multilingual abstractions for the event constraints.

Verb frames (or possibly meaningful clusters of them) can then be uploaded into the event hierarchy. The current set of event types (8 main types in a financial domain ranging from "*Company Acquisitions*" and "*Company Assets*" to "*Regulation*") can be thus connected with lexicalisations in three languages. First, the verb patterns derived for English are mapped to specific event types: a verb pattern like `(buy, [Agent:Company,Object:Company])` as a "*Company acquisition*" event type). Then, translations into Italian and Spanish rules (e.g. `(acquistare, [Agent:Company,Object:Company])`) inherit the same topological position in the ontology. Accordingly, the world model have a structure which is essentially language independent in all but the lowest level - at which stage lexicalisations relating to each representative language are required. Associated with the lexicalisations are the language dependent (verbal) rules which control the behavior of instances of these events in the discourse processing.

The integrated adoption of Eurowordnet and automatic acquisition/translation of verb rules is thus the key of a successful and quick development of the large scale IE component required in automatic authoring.

5 Conclusions

In this paper a general NLP-based approach to automatic authoring has been presented. The emphasis has been given to traditional capabilities of Information Extraction in Web service scenarios with their inherent multilinguality. IE is here seen as a powerful solution for cross-lingual hypertex-

tual authoring. Other works in this area make extensive use of traditional IR techniques (e.g. full text search) or rely on already traced (i.e. manually coded) hyperlinks (e.g. [5, 7, 12]). The suggested NAMIC architecture exploits linguistic capabilities for deriving entirely original (*ex novo*) resources, over dynamic, previously unreleased, streams of information.

NAMIC is a novel large-scale multilingual NLP application capitalising existing methods and resources within an advanced software engineering process. The use of a distributed Java/CORBA architecture makes the system very attractive for its scalability and adaptivity. In fact, although its complexity, the overall organisation (lexical interfaces with respect to the multilingual ontology) makes it very well suited for customisation and porting to large domains.

6 Acknowledgements

This research is funded by the European Union, grant number IST-1999-12392. We would also like to thank the NAMIC partners in the University of Sheffield and in the Universitat Politècnica de Catalunya that made the design and development of the prototype possible.

References

[1] S. Azzam, K. Humphreys, R. Gaizauskas, H. Cunningham, and Y. Wilks. Using a language independent domain model for multilingual information extraction. In *Proceedings of the IJCAI-97 Workshop on Multilinguality in the Software Industry: the AI Contribution (MULSAIC-97)*, 1999.

[2] R. Basili, A. Moschitti, and M. Pazienza. Language sensitive text classification. In *In proceeding of 6th RIAO Conference (RIAO 2000), Content-Based Multimedia Information Access, Coll ge de France*, Paris, France, 2000.

[3] R. Basili, M. Pazienza, and M. Vindigni. Corpus-driven learning of event recognition rules. In *Proc. of Machine Learning for Information Extraction workshop, held jointly with the ECAI2000*, Berlin, Germany, 2000.

[4] R. Basili, M. Pazienza, and F. Zanzotto. Customizable modular lexicalized parsing. In *Proc. of the 6th International Workshop on Parsing Technology, IWPT2000*, Trento, Italy, 2000.

[5] S. Chakrabarti, B. Dom, D. Gibson, J. Kleinberg, P. Raghavan, and S. Rajagopalan. Automatic resource compilation by analysing hyperlink structure and associated text. In *Proceedings of the 7th International World Wide Web Conference*, Brisbane, Australia, 1998.

[6] C. Cunningham, R. Gaizauskas, K. Humphreys, and Y. Wilks. Experience with a language engineering architecture: 3 years of gate. In *Proceedings of the AISB'99 Workshop on Reference Architectures and Data Standards for NLP*, Edinburgh, UK, 1999.

[7] J. Dean and M. R. Henzinger. Finding related pages in the world wide web. *WWW8 / Computer Networks*, 31(11-16):1467–1479, 1999.

[8] J. L. Field. It standards in the news, the history of international it standards for news. *Web site:* `http://www.iptc.org/site/history.html`, 2001.

[9] R. Gaizauskas and K. Humphreys. Xi: A simple prolog-based language for cross-classification and inheritance. In *Proceedings of the 6th International Conference on Artificial Intelligence: Methodologies, Systems, Applications (AIMSA96)*, pages 86–95, 1996.

[10] R. Gaizauskas and Y. Wilks. Information Extraction: Beyond Document Retrieval. *Journal of Documentation*, 54(1):70–105, 1998.

[11] 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. Available at http://www.saic.com.

[12] J. M. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46(5):604–632, 1999.

[13] G. Miller. Five papers on wordnet. *International Journal of Lexicography*, 4(3), 1990.

[14] *Proceedings of the Seventh Message Understanding Conference (MUC-7)*. Morgan Kaufman, 1998. Available at http://www.saic.com.

[15] M. Pazienza, editor. *Information Extraction. A Multidisciplinary Approach to an Emerging Information Technology*. Number 1299 in LNAI. Springer-Verlag, Heidelberg, Germany, 1997.

[16] P. Vossen. *EuroWordNet: A Multilingual Database with Lexical Semantic Networks*. Kluwer Academic Publishers, Dordrecht, 1998.