Multilingual Authoring: the NAMIC approach

R. Basili, M.T. Pazienza
F. Zanzotto
Dept. of Computer Science
University of Rome, Tor Vergata
Via di Tor Vergata,
00133 Roma
Italy
basili@info.uniroma2.it
pazienza@info.uniroma2.it
zanzotto@info.uniroma2.it

R. Catizone, A. Setzer
N. Webb, Y. Wilks
Department of Computer Science
University of Sheffield
Regent Court
211 Portobello Street,
Sheffield S1 4DP, UK
R.Catizone@dcs.shef.ac.uk
A.Setzer@dcs.shef.ac.uk
N.Webb@dcs.shef.ac.uk
Y.Wilks@dcs.shef.ac.uk

L. Padró, G. Rigau
Dept. Llenguatges i Sistemes Informàtics
Universitat Politècnica de Catalunya
Centre de Recerca TALP
Jordi Girona Salgado 1-3,
08034 Barcelona
Spain
padro@lsi.upc.es
g.rigau@lsi.upc.es

Abstract

With increasing amounts of electronic information available, and the increase in the variety of languages used to produce documents of the same type, the problem of how to manage similar documents in different languages arises. This paper proposes an approach to processing/structuring text so that Multilingual Authoring (creating hypertext links) can be effectively carried out. This work, funded by the European Union, is applied to the Multilingual Authoring of news agency text. We have applied methods from Natural Language Processing, especially Information Extraction technology, to both monolingual and Multilingual Authoring.

1 Introduction

Modern Information Technologies are faced with the problem of selecting, filtering and managing growing amounts of multilingual information to which access is usually critical. Traditional Information Retrieval (IR) approaches are too general in their selection of relevant documents where as traditional Information Extraction (IE) (Gaizauskas and Wilks, 1998; Pazienza, 1997) approaches are too specific and inflexible. Automatic Authoring is a good example of how these two methods can be improved and used to create a hypertextual organisation of (multilingual) information. This kind of information is 'added value' to the information embodied in the text and is not in contrast with other retrieval paradigms. Automatic Authoring is the activity of *processing* news items in streams, *detecting* and *extracting* relevant information from them and, accordingly, *organising texts in a non-linear fashion.*

While IE systems like the ones participating in the Message Understanding Conference (MUC, 1998) are oriented towards specific phenomena (e.g. *joint ventures*) in restricted domains, the scope of Automatic Authoring is wider. In Automatic Authoring, the hypertextual structure has to provide navigation guidelines to the final user which can also refuse the system suggestions.

In this paper an architecture for Automatic Multilingual Authoring is presented based on knowledge-intensive and large-scale Information Extraction. The general architecture is presented capitalising robust methods of Information Extraction (Cunningham et al., 1999) and large-scale multilingual resources

(e.g. EuroWordNet). The system is developed within a European project in the Human Language Technologies area, called NAMIC (News Agencies Multilingual Information Categorisation)[1]. It aims to extract relevant facts from the news streams of large European news agencies and newspaper producers[2], to provide hypertextual structures within each (monolingual) stream and then produce cross-lingual links between streams.

2 Authoring

2.1 Automatic Authoring

As Automatic Authoring is the task of automatically deriving a hypertextual structure from a set of available news articles (in three different languages English, Spanish and Italian in our case), the complexity of the overall framework 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 referred to in one (or more) news items as:

- *Intel, the world's largest chipmaker, bought a unit of Danish cable maker NKT that designs high-speed computer chips ...*
- *The giant chip maker Intel said it acquired the closely held ICP Vortex Computersysteme, a German maker of systems ...*
- *Intel ha acquistato Xircom inc. per 748 milioni di dollari.*

The hypothesis underlying Authoring is that all the above news items deal with facts in the same area of interest to a potential class of readers. They should be thus linked and links should suggest to the user that the underlying motivation (used to decide whether or not to follow an available link) is that they all refer to *Intel acquisitions*.

Notice that a link generation process based only upon words would fail in the above case as the common word (that could play the role of anchor in linking) is the proper noun *Intel*. As no other information is available, the resulting set of potential matches can be huge and the connectivity too high.

In order to get the suitable links the equivalence between the senses of *bought* and *acquired* in the first two news items must be known. Although such a relation can be drawn by mechanisms like query expansion or thesauri of synonyms (e.g. WordNet (Miller, 1990)), word polysemy and noise may result in an inherent proliferation of irrelevant matches. Contextual information is critical here. Notice that the senses of 'buy' and 'acquire' are constrained by the role played by *Intel* as '*agent*' and *NKT* or *ICP Vortex* being the sold companies. In fact, *Intel buys silicon* represents an unwanted sense of the verb and should be distinguished.

The relevant information concerning Intel should be thus limited to:

- *Intel* ***buys*** *a unit of NKT*
- *Intel* ***acquires*** *ICP Vortex.*

These descriptions provide the core information able to establish equivalence among the underlying events. Whenever base event descriptions are available the linking process can be carried out via simpler equivalence inferences. The Authoring problem is thus a side effect of the overall language-processing task.

[1] See http://namic.itaca.it.

[2] EFE and ANSA, the major news agencies in Spain and Italy respectively, and the Financial Times are all members of the NAMIC consortium.

According to the suggested decomposition all the above steps are mandatory. First *text processing* is responsible for morpho-syntactic recognition. Morphological units and syntactic relations are produced for each sentence at this stage. However, syntactic relations (e.g. among subjects and verbs) are not sufficient for proper event characterisation. In the example(s), the subject of the verb *acquire* is a pronoun only anaphorically referring to *Intel*. Co-reference resolution is usually applied to this kind of mismatch at the surface level. This capability is under the responsibility of the *event matching* phase. Moreover, in order to keep track of *events* over syntactic representations, references to a target ontology are required. In such an ontology, equivalence among facts (e.g. *buying companies*) is represented. For instance, the relation among *buy* and *acquire* can be encoded under a more general notion of *financial acquisition*. Ontologies also *define* the set of relevant facts of the target domain. A *financial acquisition* is a perfect example of what is needed in *corporate industrial* news but is less important, for example, in *sports* news, where *hiring of players* seems a more relevant event class.

Conceptual differences among facts (detected during event matching) motivate a selective notion of hyperlinking. These links can be thus generated during the *automatic authoring* phase. They are ontologically justified as their conceptual representation is already available at this stage. Types as *same acquisition fact*, *same person*, or *company* can be used to distinguish links and make explanations available to the user.

2.2 Multilingual Automatic Authoring

¿From a multilingual perspective, the problem is to establish links among news in different languages. Full-text approaches can rely only on language independent phenomena (e.g. proper nouns like *Intel*) that are very limited in texts. Most of the abovementioned inferences require language neutral information (i.e. conceptual and not lexical constraints). The inherent overgeneration related to word polysemy affects the results of translation-based approaches. Again principled representations made available by IE processes (i.e. templates) provide a viable solution. The different event realisations (in the different languages) can be handled during the overall event matching. A lexical interface to the ontology is able to factor the language specific information. As syntactic differences are handled during text processing, the result is a common domain model for IE plus independent lexical interfaces. The unified representation of the set of facts activates multilingual linking at a conceptual level, thus making the Authoring a language independent process. Some challenges of such a framework are:

- 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)
- the size of the lexical interfaces to the ontology available for the different languages
- the amount of task dependent knowledge. For example the definition of the set of events useful for the target application is underspecified.

In the following, we propose a complex architecture where the above problems are approached according to well-assessed techniques presented elsewhere. Robust Information Extraction is adopted (Humphreys et al., 1998) as an overall method for *text processing* and *event matching*. Target events are semiautomatically derived from domain texts and represented in the IE engine ontology. Finally, multilinguality is realised by assuming a large-scale multilingual lexical hierarchy as a reference ontology for nominal concepts. The resulting architecture for Multilingual Automatic Authoring is presented in Section 3.4.

3 The NAMIC system

3.1 Large scale IE for Automatic Authoring

Information Extraction is a very good approach to Automatic Authoring for a number of reasons. The key components of an IE

system are events and objects - the kind of components that trigger hyperlinks in an Authoring system. Coreference is a significant part of Information Extraction and indeed a necessary component in Authoring. Named Entities - people, places, and organisations, etc. - play an important part in Authoring and again are firmly addressed in Information Extraction systems.

The role of a world model as a method for event matching and coreferencing The world model is an ontological representation of events and objects for a particular domain or set of domains. The world model is made up of a set of event and object types, with attributes. The event types characterise a set of events in a particular domain and are usually represented in a text by verbs. Object Types on the other hand, are best thought of as characterising a set of people, places or things and are usually represented in a text by nouns (both proper and common). When used as part of an Information Extraction system, the instances of each type are inserted/added to the world model. Once the instances have been added, a procedure is carried out to link those instances that refer to the same thing - achieving coreference resolution.

In NAMIC, the world model is created using the XI cross-classification hierarchy (Gaizauskas and Humphreys, 1996). The definition of a XI cross-classification hierarchy is referred to as an ontology, and this together with an association of attributes with nodes in the ontology forms the world model. Processing a text acts to populate this initially bare world model with the various instances and relations mentioned in the text, converting it into a discourse model specific to the particular text.

The attributes associated with nodes in the ontology are simple attribute:value pairs where the value may either be fixed, as in the attribute `animate:yes` which is associated with the person node, or where the value may be dependent on various conditions, the evaluation of which makes reference to other information in the model.

3.1.1 The Description of LaSIE

LaSIE is a Large-scale Information Extraction system, developed for MUC (Message Understanding Conference) competitions, comprised of a variety of modules, see (Humphreys et al., 1998; MUC, 1998). Although we are not using the complete LaSIE system in NAMIC, we are using 2 of the key modules - the Named Entity Matcher and the Discourse Processor. Below is a description of each of these modules.

Named Entity Matcher The Named Entity Matcher finds named entities through a secondary phase of parsing which uses a named entity grammar and a set of gazetteer lists. It takes as input parsed text from the first phase of parsing and the named entity grammar which contains rules for finding a predefined set of named entities and a set of gazetteer lists containing proper nouns. The Name Entity Matcher returns the text with the Named Entities marked. The Named Entities in NAMIC are PERSONS, ORGANISATIONS, LOCATIONS, and DATES. The Named Entity grammar contains rules for coreferring abbreviations as well as different ways of expressing the same named entity such as Dr. Smith, John Smith and Mr. Smith occurring in the same article.

Discourse Processor The Discourse Processor module translates the semantic representation produced by the parser into a representation of instances, their ontological classes and their attributes, in the XI knowledge representation language (see Gaizauskas(1996)). XI allows a straightforward definition of cross-classification hierarchies, the association of arbitrary attributes with classes or instances, and a simple mechanism to inherit attributes from classes or instances higher in the hierarchy.

The semantic representation produced by the parser for a single sentence is processed by adding its instances, together with their attributes, to the discourse model which has been constructed so far for the text.

Following the addition of the instances mentioned in the current sentence, together with any presuppositions that they inherit,

the coreference algorithm is applied to attempt to resolve, or in fact merge, each of the newly added instances with instances currently in the discourse model.

The merging of instances involves the removal of the least specific instance (i.e. the highest in the ontology) and the addition of all its attributes to the other instance. This results in a single instance with more than one realisation attribute, which corresponds to a single entity mentioned more than once in the text, i.e. a coreference.

3.2 Ontological Modeling

As we have seen in section 3.1, some critical issues of the NAMIC project rely on the performance of the lexical and conceptual components of all linguistic processors. As NAMIC faces large-scale coverage of news in several languages we decided to adopt EuroWordNet (Vossen, 1998) as a common semantic formalism to support:

- lexical semantic inferences (e.g. generalisation, disambiguation)
- broad coverage (e.g. lexical and semantical) and
- a common interlingual platform for linking events from different documents.

The NAMIC ontology consists of 40 predefined object classes and 46 attribute types related to Name Entity objects and nearly 1000 objects relating to EuroWordNet base concepts.

3.2.1 EuroWordNet as a Multilingual Lexical Knowledge Base

Since the world model aims to describe the language used in a given domain via events and objects, the accuracy and breadth of the model will impact how well the information extraction works.

EuroWordNet (Vossen, 1998) is a multilingual lexical knowledge base (LKB) 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 developed at Princeton (Miller, 1990) containing 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. WordNet 1.6 is also connected to the ILI as another English WordNet (Daude et al., 2000). Via this index, the languages are interconnected so that it is possible to go from the words in one language to words in any other language having similar meaning. The index also gives access to a shared top-ontology and a subset of 1024 Base Concepts (BC). The Base Concepts provide a common semantic framework for all the languages, while language specific properties are maintained in the individual wordnets. The LKB can be used, among others, for monolingual and cross-lingual information retrieval, which has been demonstrated in other projects (Gonzalo et al., 1998).

3.3 Multilingual Event description

The traditional limitations of a knowledge-based information extraction system such as LaSIE have been the need to hand-code information for the world model - specifically relating to the event structure of the domain.

For the NAMIC project, we have decided to semi-automate the process of adding new 'event descriptions' to 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 a Galois lattice (as described in (Basili et al., 2000b)). These frames can then be uploaded into the event hierarchy of the discourse interpreter world model.

The world model can have a structure which is essentially language independent in all but the lowest level - at which stage lexicalisations relating to each representative lan-

guage are required. Associated with these lexicalisations are language dependent scenario rules which control the behaviour of instances of these events with a Discourse Model. These rules are expected to differ across languages in the way they control coreference for languages which are constrained to lesser or greater degree.

The lattice generates patterns which refer to synsets in the WordNet hierarchy. For our purposes, we will use patterns referring to Base Concepts in the EuroWordNet hierarchy - which allows us to exploit the Inter-Lingual-Index as described in the previous section.

These Base Concepts serve as a level of multilingual abstraction for the conceptual constraints of our events, and allow us to extend the number of semantic classes from seven (the MUC Named Entity classifications) to 1024 - the number of base concepts in EWN.

3.4 The NAMIC Architecture

The complexity of the overall NAMIC system required the adoption of a distributed computing paradigm in the design. The system is a distributed object oriented system where services (like text processing or Multilingual Authoring) are provided by independent components and asynchronous communication is allowed. Independent news streams for the different languages (English, Spanish, and Italian) are assumed. Language specific processors (LPs) are thus responsible for text processing and event matching in independent text units in each stream. LPs compile an *objective representation* (see Fig. 1) for each source texts, including the detected morphosyntactic information, categorisation in news standards (IPTC classes) and description of the relevant events. Any later Authoring activity is based on this canonical representation of the news. In particular a monolingual 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. Figure 1 represents the overall flow of information. The Language Processors are composed of a morphosyntactic (Eng, Ita and Spa MS) and an event-matching component (EM). The lexical interfaces (ELI, SLI and ItLI) to the unified Domain model are also used during event matching.

The linguistic processors are in charge of producing the *objective representation* of incoming news. This task is performed during MS analysis by two main subprocessors:

- a modular and lexicalised shallow morpho-syntactic parser (Basili et al., 2000c), providing name entity matching and extracting dependency graphs from source sentences. Ambiguity is controlled by part-of-speech tagging and domain verb-subcategorisation frames that guide the dependency recognition phase.
- a statistical linear text classifier based upon some of the derived linguistic features (Basili et al., 2000a) (lemmas, POS tags and proper nouns)

The results are then input to the event matcher that by means of the discourse interpreter (Humphreys et al., 1998) derive the objective representation. As discussed in section 3.1, coreferencing is a side effect of the discourse interpretation (Humphreys et al., 1998). It is based on the multilingual domain model where relevant events are described and nominal concepts represented.

The overall architecture is highly modular and open to load balancing activity as well as to adaptation and porting. The communication interfaces among the MS and EM components as well as among the AEs and the M-AE processors are specified via XML DTDs. This allows for user-friendly uploading of a back-end database with the detected material as well as the easy design and management of the front-end databases (available for temporary tasks, like event matching after MS). All the servers are objects in a distributed architecture within a CORBA environment. The current version includes the linguistic processors (MS and EM) for all the three languages. The English and Italian linguistic processors are fully object oriented modules based on

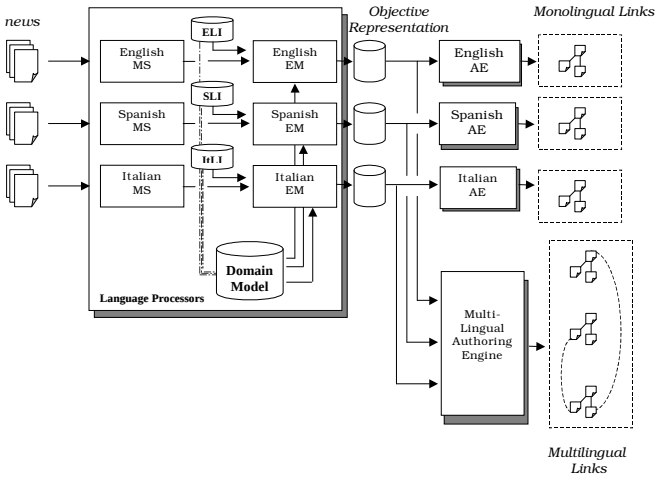

Figure 1: Namic Architecture

Java. They integrate libraries written in C, C++, Prolog, and Perl for specific functionalities (e.g. parsing) running under a Windows NT platform. The Spanish linguistic processor shares the discourse interpreter and the text classifier with the other modules, while the morpho syntactic component is currently a Unix server based on Perl. The use of a distributed architecture under CORBA allowed a flexible solution to its integration into the overall architecture. The servers can be instantiated in multiple copies throughout the network if the amount of required computation exceeds the capability of a current configuration. As the workload of a news stream is not easily predictable, distribution and dynamic load balancing is the only realistic approach.

4 Discussion and Future Work

The above sections have provided the outline of a general NLP-based approach to automatic authoring. The emphasis given to traditional capabilities of Information Extraction depends on the relevance of news content in the target Web service scenarios as well as on their inherent multilinguality. The better is the generalisation provided by the IE component, the higher is the independence from the text source language. As a result, IE is here seen as a natural approach to cross-lingual hypertextual 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. (Chakrabarti et al., 1998; Kleinberg, 1999)). The suggested NAMIC architecture exploits linguistic capabilities for deriving entirely original (*ex novo*) resources, over dynamic, previously unreleased, streams of information.

The result is a 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. It results in a very complex (but realistic) NLP architecture. Its organisation (lexical interfaces with respect to the multilingual ontology) makes it very well suited for customisation and porting to large domains. Although the current version is a prototype, it realises the complete set of core functionalities, including the main IE steps and the distributed Java/CORBA layer.

It is worth noticing that a set of extensions are made viable within the proposed architecture. A first line is the extension of the available multilingual lexical knowledge. The Dis-

course Model can be used to better reflect ontological relationships within a particular domain. These relationships could be examined to confirm known word sense usage as well as to postulate/propose novel word sense usage. Using the mechanism for the addition of events (as categorised by verbs) to the world model, users can specify new events which can be added to the IE system, to achieve User Driven IE, and deliver a form of *adaptive* information extraction.

The instantiated domain models can be thus used as a basis for ontological resource expansion as a form of adaptive process. For example, the stored instantiations of discourse models within a specific domain can be compared: it may be thus possible to recognise new sets of events or objects which are not currently utilised within the system.

The evaluation strategy that is made possible within the NAMIC consortium will make use of the current users (i.e. news agencies) expertise. The agreed evaluation methods will provide evidence about the viability of the proposed large-scale IE-based approach to authoring, as a valuable paradigm for information access.

Acknowledgements

This research is funded by the European Union, grant number IST-1999-12392. We would also like to thank all of the partners in the NAMIC consortium.

References

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

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

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

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

C. Cunningham, R. Gaizauskas, K. Humphreys, and Y. Wilks. 1999. 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.

J. Daude, L. Padro, and G. Rigau. 2000. Mapping wordnets using structural information. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics ACL'00*, Hong Kong, China.

R. Gaizauskas and K. Humphreys. 1996. 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.

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

J. Gonzalo, F. Verdejo, I. Chugur, and J. Cigarran. 1998. Indexing with wordnet synsets can improve text retrieval. In *Proceedings of the COLING/ACL'98 Workshop on Usage of WordNet for NLP*, Montreal, Canada.

K. Humphreys, R. Gaizauskas, S. Azzam, C. Huyck, B. Mitchell, H. Cunningham, and Y. Wilks. 1998. 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. Available at http://www.saic.com.

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

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

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

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

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