Intelligent Building of Language Resources for HLT Applications

Anna Samiotou[§], Lambros Kranias[§], Dimitrios Kokkinakis^{*}

[§]ESTeam AB Sikelianou 8, 146 71 Athens, Greece <u>esteam@otenet.gr</u> ^{*}University of Gothenburg Box 200, 405 30 Göteborg, Sweden dimitrios.kokkinakis@svenska.gu.se

Abstract

It is generally agreed by human language technology (HLT) practitioners (information scientists, computational linguists etc.) that the collection and processing of legacy data as well as the definition and development of the knowledge structure for the purpose of building language resources (LRs) for HLT applications, such as machine translation and knowledge management systems, is a crucial issue. This paper deals with a semi-automatic, data-driven scenario for building language resources for multilingual HLT applications. The scenario is *language* and *domain independent* to a great extend and *conforms to international standards*.

1. Introduction

This paper proposes a semi-automatic, data-driven scenario for building and structuring domain-specific and application-tailored LRs for multilingual HLT applications. The scenario deploys (a) selected available and relevant textual, lexical, terminological or ontological resources together with the targeted legacy data from the domain(s) under investigation and (b) the ESTeam Translator[©] (ET) software¹, a commercial translation system for building domain-specific lexical. terminological and translation memory resources. The built LRs are re-usable and dynamic in the sense that they can be re-targeted, extended or adapted to other applications.

The proposed scenario is used in both HLT commercial applications and R&D projects. Current applications include the building of new LRs for the coming members of the EU and their linking to all the existing EU languages, the building of LRs for the translation needs of the organisation of the Olympic Games 2004 and the Mumis² R&D project for building the domain ontology as well as fully integrating the lexicon entries with the ontological concepts in multiple languages

The paper is organised as follows: Section 2 gives a brief overview of the methodological approach employed in two concrete case studies. Sections 3 and 4 provide a brief outline for the application of the methodology in both an R&D and a commercial application. Finally, Section 5 summarises the paper.

2. Building and Structuring Multilingual LRs

The scenario for building and structuring LRs involves the collection of legacy data, data analysis and processing. By legacy data, we basically mean all the data that an organisation holds on a variety of computers, in multiple locations, and in a variety of different formats. The focal

² MUltiMedia Indexing and Searching environment; see <u>http://www.ctit.utwente.nl/projects/mumis.html</u>

point is that legacy data describes a world that can be structured into domains on which a specific application operates. In case of a knowledge intensive application, legacy data is studied and analysed further in order to identify as many important concepts that describe the domain as possible. Examples of textual legacy data within the Mumis domain (i.e. soccer games) were ticker reports, match reports and comment reports. The scenario includes the integration of domains, ontological concepts and surface lexicon entries in a unique multilingual manner.

2.1 Knowledge or Domain Structure

A top level knowledge or domain structure provides a common semantic platform for the distribution and access of information, the maintenance of resources and the linking of data to domains, facilitating indexing, archiving and retrieval at the top level. The knowledge structure can be used not only for textual but also for multimedia information such as video or audio content.

2.2 Domain Ontology Modelling

A deeper structure is the development of an ontology for each domain. The domain ontology is a hierarchical set of concepts and relationships between them which describe the domain itself internally and links to the top level domain structure. Semantic robustness towards representational changes as well as multilingual structure for lexical access are crucial for the development of the domain ontology. The initial creation of the set of concepts is ruled by the nature of the legacy data and by the available terminology. The concepts describe the relevant situations and actions within the domain. Each concept is linked to a surface term in one or more languages, ensuring cross-lingual indexing and retrieval. Whereas terms might change, and are different in each language, the semantic meaning and interpretation of the terms' abstract concept stays the same. For a comparative study of ontology building methodologies see Fernandez et al. (1999) while Guarino (1998) provide a set of methodologies for ontology-driven conceptual analysis.

¹ <u>http://www.esteam.gr</u>

2.3 Domain Ontology Development

We start by studying and analysing the legacy data, in order to identify as many of potentially important concepts as possible (i.e. ontology caption and trimming process), while brainstorming completes this first identification phase, by applying the so called middle-out approach, which identifies the most relevant terms to the more abstract and most concrete; cf. Uschold & Gruninger (1996) and López (1999). Then, the material developed in the first stage is refined and labels are produced providing information for the different classes. Furthermore, we produced definitions for most of the concepts. Finally, we decide the format of the ontology by transforming the design representation into a more formal language representation using XML. For a detailed and relevant to our scope discussion of criteria for ontology-driven conceptual analysis see Welty & Guarino, 2001.

2.4 Building Multilingual Domain-Specific LRs

All collected textual data for the specific domain is used for building the multilingual domain language resources. All languages whose character set is covered by the Unicode standard can be treated. A *first step* for a rough LR building is done through automatic processes by using the ET software (for more details on the technology see ESTeam AB, 2004), first on a monolingual and then on a multilingual level:

Monolingual level:

- Data is analysed in three processing levels per domain which includes text segmentation into sentences, sub-sentences ((sub-)sentence splitting) and words (tokenisation).
- The segmented text is sorted according to frequency of appearance and frequent collocations are extracted.
- A statistic analysis report is produced for all processing levels (sentence, sub-sentence, word) indicating the internal repetition of the data. Since repetition is very relevant to the length of the (sub-)sentences i.e. shorter sentences tend to be more repetitive, the repetition rate is calculated with relation to their length in characters (without the spaces)
- Import of all-level segmented text in the Lexicon (LeX) and/or Translation Memory (TM) database.

Multilingual level:

- Import of available multilingual domain terminology in the LeX database.
- Alignment of available parallel texts in different languages and in sentence and sub-sentence level (see Cranias, 1995 & Papageorgiou et al, 1994). The Alignment is based both on the lay-out of the texts, on lexicon look-up and verification as well as on statistics on co-occurrences of translation links within the pairs of texts.
- Import of high quality alignment results in the TM and/or LeX database. The quality threshold is user-defined.
- Import of medium quality alignment results in the ALIGN database for browsing and editing

these alignment results (see also the *second step* below). Import of the accepted and edited results in the TM and/or LeX database.

- Automatic multilingual linking of entries through a common language link. This is a unique feature of the ET which generates multilingual indirect links for entries of a language that have no direct links to other languages. This is possible since the ET system is not pair-based but fully multilingual. For example if the following translation links are imported in the database:
 - o (German) Tisch <-> (English) table

• (Spanish) mesa <-> (English) table

then the link:

 (German) Tisch <-> (Spanish) mesa is automatically generated

- Re-use of any available multilingual LRs by importing them in the LeX database and automatically filtering out the candidate entries that do not exist in the legacy data. This filtering is optional, however, it is recommended for the purpose of having a control over the domain-specific language resources (a) by not allowing the import of non-existing words and/or phrases in the database (b) by minimising the import of possible erroneous links (c) by avoiding the import of non-relevant to the domain links
- Run the ET Machine Translation (MT) on the high frequent untranslated sentences, subsentences and collocations using the built LeX. Import of the translation results in the TM database.

A second step consists of manually browsing and editing the high frequent imported entries in order to refine translations as well as to provide semantic links to the terms and fine-tune the LeX/TM on the basis of the domain knowledge. Once semantic links (domain and/or ontological concepts) are assigned on the monolingual level either automatically if the information was already in the legacy data or manually, the translation links automatically inherit the common semantic links of the relevant source and target entries through the ET software. Inheritance also applies to other features, if available, such as part-of-speech and capitalisation (i.e. if a word is correctly spelled with a first upper-case letter like proper names or nouns in German, etc.). Finally, the automatic multilingual linking feature of the ET is also active during the manual editing session.

3. The Mumis Project (R&D)

Mumis is a EU-funded project within the 5th Framework Programme Information Society Technologies (IST) of the European Union (see Declerck et al., 2003). Mumis has developed basic technology to support automatic multimedia indexing and to facilitate search and retrieval from multimedia databases in specific domains.

Since Mumis was a knowledge intensive application, using ontology-based Information Extraction as the driving force, legacy data was studied and analysed in order to identify as many important concepts as possible using the ontology caption and trimming process. There was a further refinement of concepts, namely the domain modelling. The results were structured according to XML Schema of W3C.



Figure 1: Example of Domain Ontology



Figure 2: Example of the XML Schema

A Document Type Definition (DTD) was defined for the lexicon paradigm. The ontological concepts were integrated with the surface lexicon entries in a unique multilingual manner.

A corpus of collected textual data in three languages, English, Dutch and German was used to build a domain ontology and multilingual lexicon for the football domain.

Figure 3: Example of an XML LeX entry

Based on the common domain model, off-line IE components, one per language, were developed to extract the key events and participants from football reports and to produce XML output. The result of IE components was used by a cross-document co-reference mechanism to merge the information from the different IE components. Audio and video material was also used and processed by speech recognition and image keyframes extraction techniques respectively. The on-line part of Mumis consists of a user interface allowing the user to query the multimedia database. The interface makes use of the multilingual domain lexica interleaved with the domain ontology to guide the user while typing a query.

4. The Olympic Games 2004 Application (Commercial)

In the translation department of $ATHOC^3$, the body that organises the 2004 Olympic Games in Athens, there was the need for building Greek, English and French language resources for an automatic and computer assisted translation scenario with on-line language resources support tools for look-up and automatic updating.

The analysis of the legacy data led to the following domain structure:

- Administration
- Media
- Sports

There was no need, on behalf of the client, for deep internal structure for each domain in the specific application.

Legacy data was first processed monolingualy. The repetition analysis indicated approximately a 40% repetition rate for the sentences. The sentences that were not repeated even once were further segmented into subsentence level and their repetition rate was approximately 6%. So the total repetition of the legacy data was 46%. These were imported in the TM as source units structured into the above-mentioned domains. A source unit could be linked to more than one domains.

<tu segtype="sentence"> <prop type="x-DOMN">23~sports~*/5~administration~*</prop></tu>
for the "DD".
<tuv lang="FR"></tuv>
<seg>Catégories</seg>

Figure 4: Example of a TMX monolingual TM entry

Pre-translated parallel documents, provided by ATHOC, were aligned per domain to populate the TM. In this way, the source units were linked to one or more translations. Further links where automatically generated through the multilingual linking feature of the ET.

³ ATHens Organising Committee for the Olympic Games



Figure 5: Example of a TMX bilingual TM entry

Existing multilingual terminology was also imported both in the LeX, for manual reference and further enhancement by the translators in this application, as well as in the TM, for use by the translation system.



Figure 6: Example of a TMX bilingual LeX entry

5. Conclusions

Language resource building and structuring varies according to the needs of the specific application. In this paper we presented a data-driven scenario for building and structuring LRs for multilingual HLT applications. We presented two successful current applications, an R&D and a commercial, where this scenario was used in different ways. This work will be demonstrated at the conference.

References

- Chen H. (2001). Knowledge Management Systems: A Text Mining Perspective. University of Arizona. Tucson, Arizona.
- Declerck T., Cunningham H., Saggion H., Kuper J., Reidsma D. and Wittenburg P. (2003). MUMIS advanced information extraction for multimedia indexing and searching. *Proceedings of the 4th European Workshop on Image Analysis for Multimedia Interactive Services*. London, UK.
- Fernandez L.M., Gomez-Perez A., and Pazos S. J. (1999) Building a Chemical Ontology using Methontology and the Ontology Design Environment. *IEEE Intelligent Systems*. 14:1. pp. 37--46.
- Guarino N. (1998). Formal Ontology and Information Systems. *Proceedings of FOIS'98, Trento, Italy.* Amsterdam, IOS Press, pp. 3-15.

- López F.M., (1999). Overview of Methodologies for Building Ontologies. Proceedings of the IJCAI-99 workshop on Ontologies and Problem-Solving Methods. Stockholm, Sweden.
- Uschold M. and Gruninger M. (1996) Ontologies: Principles, Methods and Applications, *Knowledge Engineering Review*. Volume 11(2).
- Welty C. and Guarino N. (2001). Supporting Ontological Analysis of Taxonomic Relationships. *Data & Knowledge Engineering*. Volume 39(1).
- Cranias L. (1995). A New Optimal Algorithm for the Solution of a Generalised Assignment Problem -Application in Automatic Text Alignment. *Proceedings* of the International Conference on Systems, Man & Cybernetics, Vancouver, Canada
- Papageorgiou, H., L. Cranias, Piperidis S. (1994). Automatic Alignment in Parallel Corpora. Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics, ACL-94, Las Cruces, New Mexico, USA
- ESTeam AB. (2004). ESTeam Translator© White Paper. URL: www.esteam.gr