Cross-Lingual Information Retrieval and Semantic Interoperability for Cultural Heritage Repositories

Monti Johanna
University of Sassari
jmonti@uniss.it

Mario Monteleone
University of Salerno
mmonteleone@unisa.it

Maria Pia di Buono
University of Salerno
dibuono@unisa.it

Federica Marano
University of Salerno
fmarano@unisa.it

Abstract

This paper describes a computational linguistics-based approach for providing interoperability between multi-lingual systems in order to overcome crucial issues like cross-language and cross-collection retrieval. Our proposal is a system which improves capabilities of language-technology-based information extraction. In the last few years various theories have been developed and applied for making multicultural and multilingual resources easy to access. Important initiatives, like the development of the European Library and Europeana, aim to increase the availability of digital content from various types of providers and institutions. Therefore the accessibility to these resources requires the development of environments enabling to manage multilingual complexity. In this respect, we present a methodological framework which allows mapping both the data and the metadata among the language-specific ontologies. The feasibility of cross-language information extraction and semantic search will be tested by implementing an early prototype system.

1 Introduction

The growing need by users to access information on the web in languages different from their own is fostering the research in the field of Cross-language Information Retrieval (CLIR) applications.

Typically in state-of-the-art CLIR applications, information is searched by means of a query expressed in the user's mother tongue. This query is automatically translated in the desired foreign language and the results are translated back in the user's mother tongue.

This process is based on two different translation stages: query translation and document translation. The query translation concerns the translation in the desired foreign language of the query expressed in the user's mother tongue, whereas the document translation is the back translation in the user's language of the relevant documents found by means of the translated query.

CLIR success obviously depends on the quality of translation and therefore inaccurate translations may cause serious problems in retrieving the relevant information in a foreign language.

A very frequent source of mistranslations in specific domain texts is represented by multi-word units (MWU). MWUs designate a wide range of lexical constructions, composed of two or more words with an opaque meaning, i.e. the meaning of a unit is not always the result of the sum of the meanings of the single words that are part of the unit. MWUs are not always easy to identify since co-occurrence among the lexemes forming the units may vary a great deal. A particular type of MWUs are term compounds, i.e. various types of compounds, but mainly noun compounds, which belong to a language for special purposes (LSP). In all languages there is a close relationship between terminology and multi-words and, in particular, word compounds. In fact, word compounds account in some cases for 90% of the terms belonging to an LSP.

Contrary to generic simple words, terminological word compounds are mono-referential, i.e. they are unambiguous and refer only to one specific concept in one special language, even if they may occur in more than one domain. Their meaning, similar to all compound words, cannot be directly inferred by a non-expert from the different elements of the compounds because it depends on the specific area and the concept it refers to.

Processing and translating these different types of compound words is not an easy task since their morpho-syntactic and semantic be-

Proceedings of Recent Advances in Natural Language Processing, pages 483–490,
Hissar, Bulgaria, 7-13 September 2013.

havior is quite complex and varied according to the different types and their translations are practically unpredictable.

The main contribution of this paper is the experimentation of a bilingual ontology-based CLIR system designed to overcome the current limitations of the state-of the-art CLIR systems and in particular to take into account a proper processing and translation of MWUs. This experiment has been set up for the Italian/English language pair and it can be easily extended to other language pairs.

The remaining of this paper is organized as follows. The next section briefly explains the related work in the area of CLIR. Section 3 describes the methodology and the tools used in the experiment. Then, section 4 is devoted to the system overview, and in particular it presents the data modeling and the system architecture extension. Finally, experiments and conclusions and future work are reported in sections 5 and 6, respectively.

2 Related work

There are several approaches to CLIR: they are either based on bilingual or multilingual Machine Readable Dictionaries (MRD), Machine Translation (MT), parallel corpora and finally ontologies. For a description of the different approaches see Hull and Greffenstette (1996), Oard and Dorr (1996), Pirkola (1999) and more recently Oard (2009).

Both MRD-based and MT-based CLIR are very popular but they present several shortcomings especially in relation to domain-specific contexts because of the lack of consideration for MWUs, a very frequent and productive linguistic phenomenon in LSPs.

Various techniques have been proposed to reduce the errors due to the presence of MWU introduced during query translation. Among these techniques, phrasal translation, co-occurrence analysis, and query expansion are the most popular.

Concerning phrasal translation, techniques are often used to identify multi-word concepts in the query and translate them as phrases. Hull and Grefenstette (1996) showed that the performance achieved by manually translating phrases in queries is significantly better than that of a word-by-word translation using a dictionary. Davis and Ogden (1997) used a phrase dictionary extracted from parallel sentences in French and English to improve the performance of CLIR. Ballesteros and Croft (1996) performed phrase translation using information on phrase and word usage contained in the Collins machine readable dictionary. More recently, Gao et al. (2001) propose that noun phrases are recognized and translated as a whole by using statistical models and phrase translation patterns and that the best word translations are selected based on the cohesion of the translation words. Finally, Saralegi and López de Lacalle (2010) use a simple matching and translation technique based on a bilingual MWU list to detect and translate them.

Co-occurrence statistics is used to identify the best translation(s) among all translation candidates using text collections in the target language as a language model, assuming that correct translations occur more frequently than wrong ones (Maeda et al., 2000; Ballesteros and Croft, 1998; Gao et al., 2001, Sadat et al., 2001).

As for query expansion techniques, Ballesteros and Croft (1996 and 1997) assume that additional terms that are related to the primary concepts in the query are likely to be relevant and that phrases in query expansion via local context analysis and local feedback can be used to reduce the error associated with automatic dictionary translation.

Concerning MT-based CLIR, MWU identification and translation problems are far from being solved. Recently, increasing attention has been paid to MWU processing in MT since it has been acknowledged that MT cannot be effective without proper handling of MWUs of all kinds. MWU processing and translation in Statistical Machine Translation (SMT) started being addressed only very recently and different solutions have been proposed so far, but basically they are considered either as a problem of automatically learning and integrating translations or as a problem of word alignment.

Current approaches to MWU processing move towards the integration of phrase-based models with linguistic knowledge and scholars are starting to use linguistic resources, either handcrafted dictionaries and grammars or data-driven ones, in order to identify and process MWUs as single units. Monti (2013) provides a thorough overview of the problem.

Ontologies are also used in CLIR and are considered by several scholars a promising research area to improve the effectiveness of Information Extraction (IE) techniques particularly for technical-domain queries. Volk et al. (2003) use ontologies as interlingua in CLIR for the medical

domain and show that the semantic annotation outperforms machine translation of the queries, but the best results are achieved by combining a similarity thesaurus with the semantic codes. Yapomo et al. (2012) perform ontology-based query expansion of the most relevant terms exploiting the synonymy relation in WordNet.

3 Methodology

Our linguistic methodology is based on the Lexicon-Grammar (LG) theoretical and practical analytical framework, formulated by the French linguist Maurice Gross (Gross, 1968; 1975; 1989).

LG presupposes that linguistic formal descriptions should be based on the observation of the lexicon and the combinatory behaviors of its elements, encompassing in this way both syntax and lexicon. Linguistic Resources (LRs) developed according to the LG framework are used in NLP applications and are helpful to achieve effective Information Retrieval (IR) Systems (Marano F., 2012).

In the field of MT-based CLIR, the LG methodology tries to overcome the shortcomings of statistical approaches as in *Google Translate* or *Bing* by Microsoft concerning MWU processing in queries, where the lack of context represent a serious obstacle to disambiguation. LG linguistic framework is grounded in the analysis of the so-called "simple sentence", achieved by considering rules of co-occurrence and selection restriction, i.e. distributional and transformational rules (active/passive, positive/interrogative, etc.) based on predicate syntactic-semantic properties in the wake of the Operator-Argument Grammar (Harris, 1982).

Thanks to the above-mentioned research studies, LG range of analysis concerns the concept of MWU as "meaning unit", "lexical unit" and "word group", for which LG identifies four different combinatorial behaviors (see De Bueriis et al., 2008).

Our LRs consist of (i) electronic dictionaries morphologically and semantically tagged, (ii) local grammars in the form of Finite State Transducers/Automata (FST/FSA) and (iii) tables in which the syntactic-semantic properties of lexical entries are described (see 5.1, 5.2).

4 System overview

In CLIR systems "the complexity of the grammatical structures and the quality of parsing are the main cause of the errors" (Vossen P. et alii, 2012). Indeed, the most frequent error is the assignment of wrong Part Of Speech (POS) to lexical meaning units. In this sense, as for IR and IE, we will see that our research framework allows to achieve major improvements both in recall and precision.

We propose an architecture, which when applied to a given language, maps data and metadata exploiting the morpho-syntactic and semantic information stored inside both electronic dictionaries and Finite State Automata/Finite State Transducers (FSA/FSTs) (presented in 5.2). Furthermore, this architecture can also map linguistic tags (i.e. POS) and structures (i.e. sentences, MWU) to domain concepts.

The first step performed by our system is a linguistic pre-processing phase in which natural language texts are analysed, tokenized and indixed and textual meaning units are assigned relevant morpho-grammatical and terminological information. During this first phase we also extract information from free-form user queries, and match this information with already available ontological domain conceptualizations.

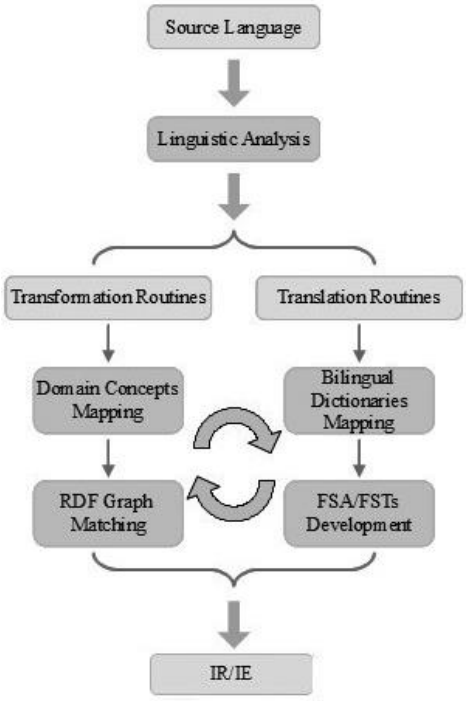

Figure 1: System Workflow

As described in Fig. 1, prior to the execution of a query against a knowledge base, it is necessary to apply the translation and transformation routines. The system is based on two workflows which are carried out simultaneously but independently.

The benefits of keeping separate these two workflows are (i) the development of an architecture with a central multilingual formalization of the lexicon, in which there is no specific target language, but each language can be at the same time target and source language, (ii) the development of extraction ontologies and SPARQL/SERQL adaptation systems which could represent a standard not only for our multilingual electronic dictionaries, but also for any lexical and/or language data-base for which translation is required.

With this dual-structure system, it is easier to successfully achieve the CLIR process since the separation of the RDF matching from the translation process allows to preserve semantic interoperability and translation quality.

5 Experiments

To test the feasibility of our architecture, we are carrying out a transfer experiment from Italian to English, using all ontological constraints defined for the Italian model.

We have chosen the Archaeological domain to test the applicability of our approach. This choice allows us to demonstrate that the modularity of our architecture may be applied to a domain which is variable by type and properties and is semantically interlinked.

In the next sections we will present the linguistic resources which have been developed for our experiment, together with the required semantic annotation and the translation system.

5.1 Electronic dictionaries

An electronic dictionary is a lexical database homogeneously structured, in which the morphologic and grammatical characteristics of lexical entries (gender, number and inflection) are formalized by means of distinctive and non-ambiguous alphanumeric tags (Vietri et al. 2004). The electronic dictionaries, used in this experiment and built according to the LG descriptive method, belong to the DELA system and are (i) the simple word dictionaries, which include semantically autonomous lexical units formed by character sequences delimited by blanks, such as *home*, and (ii) the compound word dictionaries, which include lexical units composed of two or more simple words with a non-compositional meaning, such as *rocking chair*. Terminological entries (the most common source of mistranslations in CLIR) are mainly lemmatized in compound word electronic dictionaries.

The following example represents an excerpt from the Italian-English dictionary of Archaeological Artifacts[1]

anfora di terracotta, N + NPN + FLX=C41 +DOM=RA1 + EN=earthenware amphora,
N+AN+FLX=EC3
cerchi concentrici, N + NA + FLX=C601 + DOM=RA1 + EN=concentric ridges,
N+AN+FLX=EC4
cottura ad alte temperature, N + NPAN + FLX =C611 + DOM=RA1 + EN=high fired,
N+AN+FLX=EC4
fregio dorico, N + NA + FLX = C523 + DOM=RA1 + EN=doric frieze,
N+AN+FLX=EC3
fusto a spirale, N + NPN + FLX = C7 + DOM=RA1 + EN=spiral stem,
N+AN+FLX=EC3

For instance, the compound word *fregio dorico* («Doric frieze») is marked with the domain tag «DOM=RA1», which stands for «Archaeological Artifacts – Building – Architectural Elements – Structural Elements».

For each entry, a formal and morphological description is also given with (i) the internal structure of each compound, as in *fregio dorico* where the tag «NA» indicates that the given compound is formed by a Noun, followed by an Adjective and (ii) the inflectional class, for which the tag «+FLX=C523» indicates the gender and the number of the compound *fregio dorico*, together with its plural form. The inflectional class refers to a local grammar and indicates that *fregio dorico* is masculine singular, does not have any feminine correspondent form, and its plural form is *fregi dorici*.

Together with electronic dictionaries, local grammars are used in NLP routines to parse texts. Local grammar design is based on syntactic descriptions, which encompasses transformational rules and distributional behaviours (Harris, 1957). We develop local grammars in the form of FSA/FST (Silberztein, 1993; 2002).

[1] In order to develop the Italian-English dictionary of Archaeological Artifacts, we relied on the Thesauri and Guidelines of the Italian Central Institute for the Catalogue and Documentation (ICCD) available at http://www.iccd.beniculturali.it/index.php?it/240/vocabolari For each dictionary we developed a taxonomy, therefore all entries have a terminological and domain label usable for ontologies population.

5.2 Semantic annotation

As for ontologies, the formal definition we rely upon is the one given by the International Council of Museums - Conseil Interational des Musees (ICOM – CIDOC) Conceptual Reference Model (CRM), which states that "a formal ontology (is) intended to facilitate the integration, mediation and interchange of heterogeneous cultural heritage information" (Crofts N., Doerr M., Gill T., Stead S., Stiff M. 2008). The CIDOC CRM ontology is composed of two different hierarchies, one composed of 90 classes (which includes subclasses and superclasses) and another one of 148 unique properties (and subproperties). The object-oriented semantic model and its terminology are compatible with the Resource Description Framework (RDF)[2]. This ontology is constantly developed and updated. At the same time**,** our methodology shows that a given linguistic knowledge can be reused independently from the domain to which it pertains.

LRs are used for analyzing corpora to retrieve recursive phrase structures, in which combinatorial behaviours and co-occurrence between words identify properties, also denoting a relationship. Furthermore, electronic dictionaries also include all inflected verb forms allowing to process queries expressed also with passive and more generally non-declarative sentences.

Consequently we use FSA variables for identifying ontological classes and properties for subject, object and predicate within RDF graphs.

This matching of linguistic data to RDF triples and their translation into SPARQL/SERQL path expressions allows the use of specific meaning units to process natural language queries.

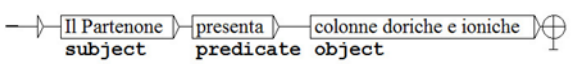

Figure 2: Simple FSA/FST with RDF Graph

Figure 2 is a sample of an automaton showing an associated RDF graph for the following sentence:

Il Partenone (subject) presenta (predicate) colonne doriche e ioniche (object)

According to our approach, electronic dictionaries entries (simple words and MWUs) are the subject and the object of the RDF triple.

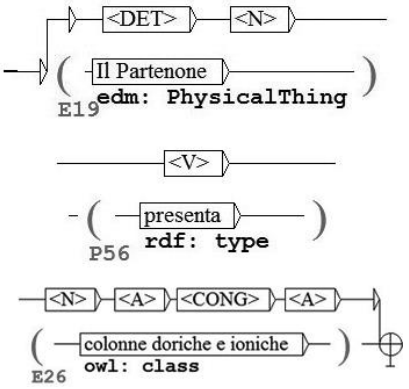

Figure 3: Sample of the use of the FSA variables for identifying classes for subject, predicate and object

In Figure 3 we develop an FSA with a variable which applies to the sentence the following classes and properties: (i) E19 indicates "Physical Object" class, (ii) P56 stands for "Bears Feature" property, (iii) E26 indicates "Physical Feature" class. So, the FSA variables transform our sentence into:

Il Partenone (E19) *bears feature* colonne doriche e ioniche (E26).

The role pairs *Physical Object/name* and *Physical Feature/type* are trigged by the RDF predicate *presenta*.

Besides in Fig. 3 we also indicate specific POS for the first noun phrase *Il Partenone* (DETerminer + Noun), the verb *presenta* (V) and the second noun phrase *colonne doriche e ioniche* (Noun+Adjective+Conjunction+Adjective).

By applying the automaton in Fig. 3 (built using the high variability of lexical class and not of the original form) we can recognize all instances included in E19 and E26 classes, the property of which is P56.

5.3 Query Translation

In our model, the Translation Routines are applied independently of the mapping process of the pivot language. This allows us to preserve the semantic representation in both languages.

Indeed, identifying semantics through FSA guarantees the detection of all data and metadata expressed in any different language.

Figure 4 shows a FST in which a translation process from Italian to English is performed on

[2] Information about the Resource Description Framework (RDF) can be found at http://www.w3.org/RDF/

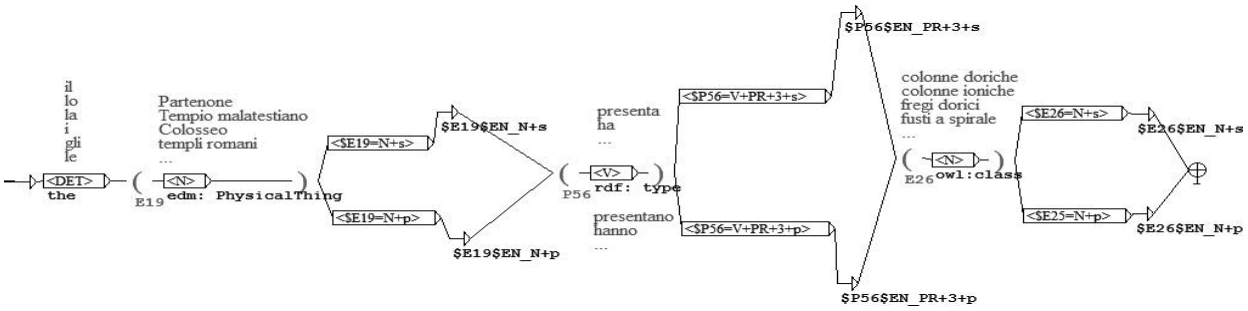

Figure 4: Translation FST with variables for identifying classes for subject, predicate and object

the basis of a dictionary look-up, a morpho-syntactic and semantic analysis. This translation FST, in fact, recognizes and annotates the different linguistic elements of declarative sentences such as "Il Partenone presenta fregi dorici", "I templi romani hanno fusti a spirale", etc, with their morpho-syntactic and semantic information and performs automatic translations on the basis of a well-crafted LG bilingual dictionary.

For instance, if a grammar variable, say $E26, holds the value "fusti a spirale", the output $E26$EN will produce the correct translation "spiral stems", on the basis of the value associated to the +EN feature in the bilingual entry "*fusto a spirale, N + NPN + FLX = C7 + OM=RA1 + EN=spiral stem*, N+AN+FLX=EC3" and the morpho-syntactic analysis performed by the graph in Figure 4, which identifies and produces the plural form of the compound noun "fusto a spirale".

5.4 Translation Quality Evaluation (TQE)

Often using smart technologies for MT involves the lowering of Translation Quality (TQ). In LG methodology, instead, we take advantage of well-formed LRs to maintain a high level of TQ. The Translation Quality Evaluation (TQE) methodology adopted to solve this problem is based on a hybrid approach, that encompasses human and automatic evaluation.

The process is composed of two cycles. The first cycle can be outlined as follows (i) a query expressed in a Source Language (SL) is the input of the CLIR application, (ii) the MT system produces sample queries (i.e. sample texts) in the Target Language (TL), (iii) the resulting translated queries are examined by humans (Linguists, Translators, Terminologists/Domain Experts) to evaluate their quality. The human judgements are based on common criteria of TQ – i.e. adequacy and fluency – and are expressed using a Likert scale with scores 1-5 (for instance using following judgements: 1. Strongly disagree, 2. Disagree, 3. Neither agree nor disagree, 4. Agree, 5. Strongly agree), (iv) only texts which obtained scores 4-5 become "validated" and "supervised" texts which represent the gold standard, (v) this gold standard is the training set for the Automatic Evaluation process, that can be carried out using METEOR[3] and GTM[4], that are the most suitable methods according to our opinion, as well as other ones[5].

During the second cycle, human evaluation is skipped and the SL queries directly become the input for automatic evaluation.

It is necessary to periodically repeat the first cycle in order to enrich the training set and to increase the quality cycle.

6 Conclusions

The proposed architecture ensures not only the coverage of a large knowledge portion but preserves deep semantic relations among different languages.

Future work aims at implementing further Linguistic Resources to achieve translation accuracy in CLIR applications and semantic search.

Note

Johanna Monti is author of sections 1, 2 and 5.3, Mario Monteleone is author of sections 5.1 and 6, Maria Pia di Buono is author of sections 4, 5 and 5.2 and Federica Marano is author of section 3 and 5.4.

[3] http://www.cs.cmu.edu/~alavie/METEOR

[4] http://nlp.cs.nyu.edu/GTM

[5] BLEU and NIST (based only on precision measure), F-Measure (based also on recall).

References

Ballesteros L. and Croft B. 1996. *Dictionary Methods for Cross-Lingual Information Retrieva*l. Proc. of the 7th DEXA Conference on Database and Expert Systems Applications, Zurich, Switzerland, September 1996: 791-801.

Ballesteros L. and Croft B. 1997. *Phrasal translation and query expansion techniques for crosslanguage information retrieval*. In Proceedings of the 20th annual international ACM SIGIR conference on Research and development in information retrieval.

Ballesteros L. and Croft B. 1998. *Resolving Ambiguity for Cross-language Retrieval*. SIGIR'98, Melbourne, Australia, August 1998: 64-71.

Bloomfield L. 1933. *Language*. Henry Holt, New York.

Crofts N., Doerr M., Gill T., Stead S., Stiff M. (eds.). 2008. *Definition of the CIDOC Conceptual Reference Model, Version 5.0*.

Davis M. W., and Ogden W. C. 1997. *Free resources and advanced alignment for cross-language text retrieval.* In: The Sixth Text Retrieval Conference (TREC-6). NIST, Gaithersbury, MD.

De Bueris G., Elia, A. (eds.). 2008. *Lessici elettronici e descrizioni lessicali, sintattiche, morfologiche ed ortografiche*. Plectica, Salerno.

Gao J., Nie J., Xun E., Zhang J., Zhou M., Huang C. 2001. *Improving Query Translation for Cross-Language Information Retrieval using Statistical Models.* Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval. ACM.

Gross M. 1968. *Grammaire transformationnelle du français. – I – Syntaxe du verbe*, Larousse, Paris.

Gross M. 1975. *Méthodes en syntaxe, régime des constructions complétives*, Hermann, Paris.

Gross M. 1989. *La construction de dictionnaires électroniques*. Annales des Télécommunications, vol. 44, n° 1-2: 4-19, CENT, Issy-les-Moulineaux/Lannion.

Harris Z.S. 1957. *Co-occurrence and transformation in linguistic structure*. Language 33,: 293-340.

Harris Z.S. 1964. Transformations in Linguistic Structure. *Proceedings of the American Philosophical Society* 108:5:418-122.

Harris Z.S. 1982. *A Grammar of English on Mathematical Principles*. John Wiley and Sons, New York, USA.

Hull D. A. and Grefenstette G. 1996. *Querying across languages: a dictionary-based approach to multilingual information retrieval,* Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval: 49-57.

Knoth P., Collins T., Sklavounouy E., Zdrahal Z.. 2010. *Facilitating cross-language retrieval and machine translation by multilingual domain ontologies.*

Maeda, A., Sadat, F., et al. 2000. *Query Term Disambiguation for Web Cross-Language Information Retrieval using a Search Engine*. in Proc. of the Fifth Int'l Workshop on Info. Retrieval with Asian Languages, Hong Kong, China: 173-179.

Marano F. 2012. *Exploring Formal Models of Linguistic Data Structuring. Enhanced Solutions for Knowledge Management Systems Based on NLP Applications.* PhD Dissertation, University of Salerno, Italy.

Monti, J. 2013. *Multi-word unit processing in Machine Translation: developing and using language resources for multi-word unit processing in Machine Translation*. PhD dissertation. University of Salerno, Italy.

Oard D. W. 2009. *Multilingual Information Access.* in Encyclopedia of Library and Information Sciences, 3rd Ed., edited by Marcia J. Bates, Editor, and Mary Niles Maack, Associate Editor, Taylor & Francis.

Pirkola A. 1998. *The Effects of Query Structire and Dictionary Setups* in Dictionary-Based Cross-language Information Retrieval. In Croft, W.., et al., 21st Annual ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2008), Melbourne, Australia, August 24-28:55-63.

Sadat F., Maeda A., et al. 2002. *A Combined Statistical Query Term Disambiguation in Cross-language Information Retrieval*. Proc. of the 13th Int'l Workshop on Database and Expert Systems Applications, Aix-en-Provence, France, September 2002: 251-255.

Saralegi X. and de Lacalle M. L. 2010. *Dictionary and Monolingual Corpus-based Query Translation for Basque-English CLIR.*

Silberztein M. 1993. *Dictionnaires électroniques et analyse automatique de textes*, Masson, Paris.

Silberztein M. 2002. *NooJ Manual*. Available for download at: www.nooj4nlp.net.

Szpektor I., Dagan I., Lavie A., Shacham D., Wintner S. 2007. *Cross Lingual and Semantic Retrieval for Cultural Heritage Appreciation*. Proceedings of the ACL Workshop on Language Technology for Cultural Heritage Data, Prague, Czech Republic.

Vietri S., Elia A. and D'Agostino E. 2004. *Lexicon-grammar, Electronic Dictionaries and Local Grammars in Italian,* in Laporte, E., Leclère, C.,

Piot, M., Silberztein M. (eds.), Syntaxe, Lexique et Lexique-Grammaire. Volume dédié à Maurice Gross, Lingvisticae Investigationes Supplementa 24, John Benjamins, Amsterdam/Philadelphia.

Volk M., Vintar S., and Buitelaar P. 2003. *Ontologies in cross-language information retrieval.* Proceedings of WOW2003 (Workshop Ontologie-basiertes Wissensmanagement), Luzern, Switzerland.

Vossen P., Soroa A., Zapirain B. and Rigau G. 2012. *Cross-lingual event-mining using wordnet as a shared knowledge interface.* Proceedings of the 6th Global Wordnet Conference, C. Fellbaum, P. Vossen (Eds.), Publ. Tribun EU, Brno, Matsue, Japan, January 9-13:382-390.

Yapomo M., Corpas G., and Mitkov R. 2012. *CLIR- and ontology-based approach for bilingual extraction of comparable documents*. The 5th Workshop on Building and Using Comparable Corpora.