PROJECT PERIODIC REPORT

November 15, 2013

Grant Agreement number	FP7-ICT-247914
Project acronym	MOLTO
Project title	Multilingual Online Translation
Funding Scheme	STREP
Date of latest version of Annex I	21 Sep 2011
against which the assessment will be made	
Periodic report	3rd
Period covered	from M25 to M39
Name, title and organisation of the scientific	Aarne Ranta, Professor,
representative of the project's coordinator	University of Gothenburg
Tel	+46317721082
Fax	+46317723663
E-mail	aarne@chalmers.se
Project website address	http://www.molto-project.eu/

Self-declaration: separate document.

1. Publishable summary

The project MOLTO - Multilingual Online Translation, started on March 1, 2010 and ran until 31 May 2013. Its goal was to develop tools for translating texts between multiple languages in real time with high quality. MOLTO's grounding technology is multilingual grammars based on semantic interlinguas and grammar-based translation. It also explores ways to use statistical machine translation without sacrificing quality.

MOLTO uses specific interlinguas that are based on domain semantics and are equipped with reversible generation functions. Thus translation is obtained as a composition of parsing the source language and generating the target language. An implementation of this technology is provided by GF, Grammatical Framework, which in MOLTO is furthermore complemented by the use of ontologies, as in the semantic web, and by methods of statistical machine translation (SMT) for improving robustness and extracting grammars from data. GF has been applied in several small-to-medium size domains, typically targeting several parallel languages. During its lifetime, MOLTO has scaled up this technology in terms of productivity, domain size, and the number of languages.

The size of domains has been increased to involve up to thousands of concepts. and the number of languages to twenty parallel ones. A special focus has been to make the technology accessible to domain experts without GF expertise and minimize the effort needed for building a translator. Ideally, the MOLTO tools will reduce the overall task to just extending a lexicon and writing a set of example sentences.

MOLTO was initially committed to dealing with 15 languages, which included 12 official languages of the European Union - Bulgarian, Danish, Dutch, English, Finnish, French, German, Italian, Polish, Romanian, Spanish, and Swedish - and 3 other languages - Catalan, Norwegian, and Russian. The additional languages also addressed in MOLTO are Chinese, Hebrew, Hindi, Latvian, Persian, and Urdu.

While tools like Systran (Babelfish) and Google Translate are designed for consumers of information, MOLTO's main target is the producers of information. Hence, the quality of the MOLTO translations must be good enough for, say, an e-commerce site to use in translating their web pages automatically without the fear that the message will change. Third-party translation tools, possibly integrated in the browsers, let potential customers discover, in their preferred language, whether, for instance, an e-commerce page written in French offers something of interest. Customers understand that these translations are approximate and will filter out imprecision. If, for instance, the system has translated a price of 100 Euros to 100 Swedish Crowns (which equals 11 Euros), they will not insist to buy the product for that price. But if a company had placed such a translation on its website, then it might be committed to it. There is a well-known trade-off in machine translation: one cannot at the same time reach full coverage and full precision. In this trade-off, Systran and Google have opted for coverage whereas MOLTO opts for precision in domains with a well-understood language.

MOLTO technology is continuously released as open-source software and linguistic modules, accompanied by cloud services, to be used for developing plug and play components to translation platforms and web pages and thereby designed to fit into third-party workflows. The project showcases its results in web-based flagship demos applied in three case studies: mathematical exercises in 15 languages, patent translations and queries in 3 languages, and museum object descriptions and queries in 15 languages. The MOLTO Enlarged EU scenarios add to this an application of MOLTO tools to a collaborative semantic wiki and to an interactive knowledge-based system used in a business enterprise environment.

Software

Foreground specifically created in MOLTO:

http://www.molto-project.eu/view/biblio/type/Software

Major open-source projects that MOLTO built on and contributed to, and which are developing MOLTO's achievements further:

- Grammatical Framework: http://www.grammaticalframework.org/
- Asiya MT evaluation platform: http://asiya.lsi.upc.edu/
- Attempto Controlled English: http://attempto.ifi.uzh.ch/site/

2. Project objectives for the period

The last period of the MOLTO project and of its enlargement MOLTO-EEU has been a very intensive period of work for the Consortium. The major deliverables have been delayed to this period and had to be completed. They included:

- GF in the cloud: a sample collection of web services for GF and MOLTO grammars
- a revision of the GF-OWL interoperability
- a unified grammar for handling natural language queries, customized for the case studies
- a third party translators' platform in which the MOLTO tools were integrated
- the overall evaluation of the MOLTO work packages

To demonstrate the usage of the MOLTO tools and technologies, the partners worked towards joint prototypes for the various case studies listed in the workplan. The coordination work involved agreement on platforms, on formats, and on the overall architecture of each demonstrator.

The final case studies include:

- a proof-of-concept dialog system and reasoner for word problems (WP6)
- patent translation by the robust hybrid approach and multilingual query interface (WP3, WP4, WP5, WP7)

- museum artifacts multilingual query and descriptions (WP4, WP8)
- multilingual semantic wiki AceWiki (WP2, WP11)
- multilingual business modelling by GF (WP2, WP3, WP12).

Followup to the reviewers' report

Recommendation 1

Technical coordination should be strengthened. Continuous and strict monitoring should be applied. Reviewers made several recommendations in the 1st review but most of them have not been implemented or it was unclear what was done with respect to them. As it is shown in the remarks per WP, the adoption of most of these recommendations would support monitoring of the work progress towards the project's objectives.

The greatest effort undertaken to strengthen the coordination of the partners was to define a number of "flagships" aimed at demonstrating the integration of the MOLTO technologies. These showcase demonstrators have been developed during the final months of the project by tight cooperation of the partners, each flagship adopting and reviewing some tool or technology from a different partner.

Recommendation 2

The recommendation from the 1st review "How grammar rules are extracted (from lexical databases, ontologies, text examples) needs to be specified in detail and a concrete schedule should be included in the updated workplan (D1.1)" has not been included in D1.1.It should be included in D2.3 "Grammar tool manual and best practices", due in M27. This is a crucial deliverable since the best practices with respect to the other work packages should be included here

Recommendation 3

The recommendation from the 1st review "Details on the integration steps (the integration of the vocabulary editor with the translation editor, the integration of the vocabulary editor with TermFactory (TF), and the integration of TF with the Knowledge Representation Infrastructure (KRI) of WP4) need to be provided in the updated workplan (D1.1). Concerning the integration of TF and KRI, it seems that there are overlaps between these tools. The partners must clarify which functions of these tools will be used in the case studies in order to exploit complementarities of the tools and avoid overlaps." has not been addressed properly and is presented as still "less understood" by the WP leader. This is a major issue of concern. The problems of the integration of WP3 tools remain. These should be discussed in an updated D1.1. Follow-up: D4.3A makes comparison between KRI and TF and suggests steps to be taken to integrate TF and KRI. The integration requires a mirror of a KRI site, whose semantic repository is open to edit with TermFactory. The resulting knowledge base in the mirror site will be grammatically enriched, so that the

new information is presented to the user. Moreover, the integration can facilitate lexicon extraction for the GF grammars and the query language of KRI.

Recommendation 4

The translator's tool s that should be developed in WP3 should not be given up. Although the WP's leader's impression is that the MT quality is too low for the tools to ever be used, the developed tools can be useful for those subdomains/language pairs where MT quality is better.

Follow-up: The development of the translator's tool has been continued, but with another platform. Deliverables 3.1 and 3.2 use GlobalSight, a translation management system, and an external editor that supports GF. However, we found that GlobalSight was not maintained, and changed to Pootle, a modern and lightweight translation platform with an active user base. D3.3 describes the integration of the GF translation to Pootle. A demo video is found at MOLTO's youtube channel.

Recommendation 5

The recommendation from the 1st review "Critical issues with respect to the semi-automatic creation of abstract grammars from ontologies, as well as deriving ontologies from grammars, are still to be clarified. Concrete steps to handle these issues need to be specified in detail and a schedule should be included in the updated work plan (D1.1). In addition, as noted with respect to WP3, complementarities between KRI and TF should be exploited avoiding possible overlaps. Terminology should be added and abbreviations explained in Deliverable D4.1 in order to facilitate reading by non-experts in the field" should still be addressed. The issue of the two-way interoperability between ontologies and GF grammars still remains unclear, although as noted in the DoW this represents one of the two most research-intensive parts of MOLTO. This should be solved in the new versions of D4.2 and D4.3 The current version of deliverables D4.2 "Data Models, Alignment Methodology, Tools and Documentation", and D4.3 "Grammar-Ontology Interoperability" are not approved. D4.2 is too general. For instance, a lot is said about LOD and the museum case and not on the alignment methodology. D4.3, on the other hand, does not give a clear picture of the interoperability issues and the degree of automation that can be expected. What is required for porting this to a new application? Concrete steps should be provided making clear what can be automated and what cannot with the provided infrastructure.

Follow-up:

- D4.1 includes a "Glossary" with definitions and explanations of all technical terms. (see the .pdf version of the document, available on the website)
- D4.3A gives details on the further work in the field of grammar-ontology interoperability and summarizes the achieved in WP4, WP7 and WP8 in the field(since the prototypes of these work packages are child projects of the KRI prototype). The highlights are using GF to generate SPARQL,

observing it as yet another language, and also the details of (semi-)automatic verbalization of RDF facts with the help of GF.
- D4.2 was updated
- D4.3 was updated and D4.3A (annex deliverable) was published. D4.3A gives more specific details than D4.3 and explains what was achieved in the WP4 goals after M24 of MOLTO.
- In the end of D4.3A we list the steps of customization of the KRI prototype for other specific domains. A highlight is that the need of GF expert and knowledge engineer cannot be overcome; the two experts should work in collaboration to design the prototype's query language(auto-suggestions are possible) and a few iterations of mutual work might be needed to refine the result.

Recommendation 6

The current description of work in WP6 lacks details on the prototype multilingual dialogue system to be developed. As recommended in the 1st review, an example dialogue and specifications of this prototype should be provided. These can be included in D9.1E.

Example dialog and description are available at D6.3 cover document,

http://www.molto-project.eu/sites/default/files/D6.3.pdf

Sections 1, 3 and 4).

Recommendation 7

WP7 work should focus on the major issues examined in MOLTO, especially in relation to the grammar – ontology interoperability automation. Specific scenarios are needed for the exploitation of MOLTO tools in this case study. It is recommended to include such scenarios in deliverable D9.1E.

Follow-up:

In response, two use case scenarios were described: UC-71 and UC-72.

- UC-71 focuses on grammar-ontology interoperability. User queries, written in CNL, are used to query the patent retrieval system. We defined a query language and a new query grammar in order to a) decrease the number of ambiguities in the queries and b) increase the coverage of the ontology. As a result, we come up with a more reusable grammar (Y AQL), easier to maintain, that facilitates the lavour of building query grammars for the application domain and languages. NL queries are translated into SPARQL using this approach. Additional details are given in D7.3 and D4.3A.
- UC-72 focuses on high-quality machine translation of patent documents, and the ultimate goal is to endow the retrieval system with the information required to enable multilinguality. We used an SMT baseline system to translate a big dataset of patents and feed the retrieval databases. The automatic translation included the semantic annotation, available only

in English documents. This mechanisms allowed to extract multilingual lexicons for the domain ontology, which were used also to build the query grammars. More details are also given in D7.3.

- Finally, the exploitation plans for the technologies developed within this WP, which are further discussed in D10.4, are focused on multilingual text processing and cross-lingual translation of various domain data within search and retrieval techniques.

Recommendation 8

The recommendation from the 1st review "Preparation of a new version of D9.1 is recommended including prototype specifications and scenarios for the three case studies (WP6, WP7, WP8)" should still be addressed. A concrete evaluation methodology is needed focusing on MOLTO's major goals: How will the consortium prove that its objectives were fully/partially met? We expect to see this in D9.1E "Addendum to the MOLTO test criteria, methods and schedule" hoping that the recommendations suggested above as well as in the 1st review, in relation to D9.1, will be included there.

Follow-up: D9.1A "Appendix to MOLTO test criteria, methods and schedule" addresses these issues.

Recommendation 9

The way the project's web site is structured, although it contains the necessary content, affects its readability in some cases. It should contain a structure according to the work packages, including all documentation related to a specific work package.

The content published on the web site can be navigated according to the way the producer has tagged it. If the author has decided to tag a certain item as belonging to a work-package then this content will display when selecting the proper tag: e..g http://www.molto- project.eu/category/dow/potential-impact/dissemination or, for publications,

http://www.molto- project.eu/biblio/keyword/88

will select the WP7-related bibliography. However, to the casual reader of the website, the distinction in work-packages is not very informative and the results are best viewed independently of the contingent organization in the work-plan. Following this principle, we have created a navigation menu that distinguishes the internal, work-plan related items from the public more general publications.

Recommendation 10

The deliverables on the work plan (D1.1) and the dissemination plan (D10.1) should be updated at the beginning of the 3rd year.

We have adopted the methodology to continuously use online publication tools on the internal section of our web pages in order to maintain the work

plan, the dissemination plan and their updates. Partners that are undergoing new activities use the news feed to inform the Consortium. Work package leaders have been given the option to create tasks, allocate and manage them. Some of the work planning has been coordinated by the partners using third party specific tools such as Trello (trello.com) and Symphonical (https://www.symphonical.com).

3. Work progress and achievements during the period

WP2 Grammar Developer's Tools

Summary of progress

GF Eclipse plugin http://www.grammaticalframework.org/eclipse/index.html has grown to Version 1.5.1 by June 2012, when this work package finished. It has been adopted by Be Informed and Ontotext. Camilleri and Ranta gave a GF crash course at Be Informed using Eclipse. There are moreover two publications: one in EAMT (a poster), one at FreeRBMT (full paper).

The Resource Grammar Library has been enhanced by two languages as external contributions: Japanese and Latvian. The MOLTO Phrasebook has been extended to Latvian. Work on Chinese and Maltese is going on.

With the release of D2.3, Grammar Tools and Best Practices, the work in this WP finished on 30 June. But there was further work planned, as dissemination (carried out e.g. in GF Summer School, August 2013) and publication as a journal article.

A preview version of libpgf, a C-based reimplementation of the GF runtime, is available since July 2012. When finished, it should make GF technology accessible to applications that cannot make use of the current Haskell- and Java-based runtimes either due to resource constraints or interoperability concerns. In particular, libpgf should be easier to access from non-JVM-based programming languages. Bindings for Python are available since September.

Note: the C-based GF runtime was released before the end of MOLTO and also made available as a web service in the "GF Robust Translator", http://cloud.grammaticalframework.org/translator/

Highlights:

- GF Eclipse plugin http://www.grammaticalframework.org/eclipse/index.html
- D2.3 http://www.molto-project.eu/biblio/deliverable/grammar-tools-and-best-pr...

Deviations from Annex I

- The delivery date of D2.3 was postponed from M24 to M27 to be able to profit from the initial experiences with the new partners' scenarios.

Use of resources: in accordance with Annex I.

WP3 Translator's Tools

Summary of progress

We have developed translator's tools further: lexicon extraction as an essential core technology, integration of GF translation to the Pootle platform as a concrete example. Preparing for the final review, UHEL has had the flagship of lexicon extraction and prepared a presentation. UGOT has contributed to the lexicon work with Shafqat Virk's PhD research about resources for Indo-Aryan languages. We have also collaborated with Ontotext in D4.3A to include contrast and comparison of TermFactory and KRI, as suggested by the reviewers.

Highlights

- Deliverable 3.3 which describes the GF integration to Pootle
- Lexicon conversion to GF format added to TermFactory
- ontoR (ontology based term alignment) tool providing lexicon output to TermFactory format

Deviations from Annex I

- A part of the work on the web-based translation tool originally scheduled for UHEL was carried out by UGOT.

Use of resources: a shift of workload from UHEL to UGOT of 3 person months.

WP4 Knowledge engineering

In the final period of MOLTO, we have finalized the model for SPARQL generation and RDF facts verbalization. The D4.3A annex deliverable was published as a follow-up to the reviewers' recommendations and to summarize the progress of work in the field of grammar- ontology interoperability in the descendants of the KRI prototype.

Highlights

- We deliver D4.3A annex deliverable to D4.3, that addresses reviewers' remarks and recommendations from M24 and also describes the subsequent work we have done after M24
- Development of the Y AQL core module, that was propagated afterwards in WP7, WP8, and WP12. It allowed exploring SPARQL as yet another natural language and hence the translation from NL to SPARQL can be fascilitated directly by GF
- Research on verbalization of RDF facts
- A conceptual model for automatic verbalization of RDF facts through GF means, and its implementation for WP4 prototype(single facts) and WP8 prototype(painting descriptions)
- Addition of answer generation grammars for the molto-kri prototype - (http://molto.ontotext.com/); Swedish and Finnish grammars were added to demonstrate the multilinguality power of GF.

Deviations from Annex I: none.

Use of resources: in accordance with Annex I.

WP5 Statistical and robust translation

Summary of progress

The main goal of this WP has been to develop a hybrid system between GF and SMT specialised in patent translation and has implied the construction of new resources on the domain and conceiving techniques to integrate both technologies. The WP has also tasks devoted to widen GF and have been focused on building general purpose lexicons. Besides, a more robust GF has been achieved by the use of the robust statistical parser that will allow to translate free text or, at least, the parts covered by the grammars without being affected by unknown elements.

Regarding the development of different types of general lexicons it has been used GF's core idea of common abstract syntax and multiple concrete syntaxes to produce multilingual morphological lexicons. The abstract syntax is based on data from the Princeton WordNet and the Oxford Advanced Learner's dictionary. The concrete syntaxes are produced using data from already existing lexical resources (i.e. Bilingual dictionaries and Universal WordNet), and GF's morphological smart paradigms. Because words can have multiple senses, and it is often very hard to find one-to-one word mappings between languages, two different types of multi- lingual lexicons have been developed: Uni-Sense and Multi-Sense. In a uni-sense lexicon each source word is restricted to represent one particular sense of the word, and hence it becomes easier to map it to one particular word in the target language. These type of lexicons are useful for building domain specific NLP applications. A multi-sense lexicon, on the other hand, is a more comprehensive lexicon and contains multiple senses of words and their translations to other languages. This type of lexicons can be used for open-domain tasks such as arbitrary text translation. These lexicons cover a number of language including English, German, Finnish, Bulgarian, Hindi, Urdu and their size ranges from 10 to 50 thousand lemmas.

In WP5 we also experimented with open-domain robust translation based solely on GF. This is a huge step since the traditional application domain of GF is in controlled languages where the domain is small and well defined, while in the task of translating running text the source language is not clearly defined anymore. As a simple numerical measure for the leap, we can say that the typical GF applications deal with grammars containing hundreds of lemmas while in this experiment our grammars contain more than 50,000 lemmas. We developed an entirely new runtime system for GF in C which has the advantage to be more portable and more efficient. The efficiency was the first requirements that we had to satisfy since otherwise interpreting these huge grammars would be intractable. Furthermore, we turned the original non-probabilistic algorithms for parsing and reasoning into probabilistic ones. The introduction of probabilistic models is crucial for the disambiguation of the grammars which are by necessity highly ambiguous. The third major contribution to the project is that we also made the GF parser robust, i.e. when faced with sentences which are not parseable, it returns a sequence of recognized chunks rather than an error.

We evaluated our implementation with state-of-the- art statistical parsers for related grammatical formalisms, and we found that for sentences longer than 25 tokens, our implementation is at least two orders of magnitude faster. We also tried to use our new architecture in machine translation but here the results are not conclusive yet. We found that the two main limitations are in the quality of the translation dictionaries which we built and the still limited coverage of the grammars. Furthermore, we need to better address the word sense disambiguation and the proper translation for multiword expressions.

The translation of patents using this robust parsing is still in an embryonic state, but we have developed a complete translation system that combines GF and SMT to overcome the input controlled language assumption. This hybrid system implies the construction of in-domain dictionaries and grammars that make use of probabilistic components, and the integration with an SMT engine that is able to complement GF translations. Regarding these resources for patent translation, we emphasise the generation of static lexicons obtained from SMT translation tables, and the on-line generation of lexicons with unseen vocabulary but available in the monolingual dictionaries. For German, also a dictionary of compounds has been built. A grammar for dealing with patents in English, French and German has been built on top of the resource grammar with several additions devoted to deal with chunks instead of sentences. Particular constructions appearing in patents are also covered by this new in-domain grammar. As a demand of the selected domain, we have also developed a detector and tokeniser of chemical compounds. A full translation system uses this tokeniser and prepares the patent to be translated. This involves chunking and parsing the source sentences which are first translated by GF and afterwards sent to an SMT decoder which is fed with this information. An SMT engine trained on the domain is also used by the top decoder. The final hybrid system is available for download and has several options that take into account which method to build the lexicon has to be used and which kind of integration is to be applied.

Highlights

- The novelties since the last report correspond on the one hand to the improvement of the previous hybrid MT systems, its portage to German, and the development of new hybrid systems. On the other hand, we highlight the generation of lexical resources from WordNet, Apertium dictionaries, and SMT translation tables and the development of a statistical robust parser which results two order of magnitude faster than comparable state-of-the-art probabilistic parsers. The last points allow to extend the coverage of GF and are useful for a general translation or a translation in any domain. The first one, on the contrary, starts from the translation on a concrete domain and tries to extend the coverage outside the coverage of the grammar
- The main goals were achieved:
  - GF grammar for the patents domain
  - an SMT system for patents
  - an SMT-driven combination GF-SMT translators

– a prototype of robust GF-driven translator, together with a web demo

Deviations from Annex I

- The hybrid system that depends on using GF tree fragment pairs is in a less mature state than we expected. The dependence on the performance of the robust parser showed to be crucial. Efforts were devoted into this direction on the level of basic algorithms. As the integration with the patent system did not give competitive results, this techniques was not included in the final patent system delivered.

Use of resources: in accordance with Annex I.

WP6 Case study: mathematics

Summary of progress

In the first part of the project we developed a GF Mathematical Grammar Library (mgl) based on several OpenMath content dictionaries. This encompasses the OpenMath layer of the mgl. For the next part we developed, on top of it, the module Commands that allows the use of human language at commanding a Computer Algebra System (cas) into computing the objects described in the OpenMath layer, and getting the answers delivered in natural language too.

For the final part we undertook creating a prototype for assisting students into modeling and solving word problems: The statements of these problems relates to notions of ordinary life and the goal of proposing these to the students is for they to learn how to describe mathematically the relevant relations in the statement into equations (modeling) and then, how to solve these to get the numeric solutions interpreted in terms of the original statement (solving).

The kind of reasoning needed in this the description logic used by WP4 (OWL reasoners) was found wanting in its arithmetical capabilities. We needed a dialog system more than a query/answer system. This moved into creating a new reasoner based on Prolog, along the lines of WP11, able to cope with basic arithmetic settings. That means, being able to automatically decide whether a problem statement is free from contradictions and whether it contains enough information to deduce the solution. On the other hand, since we want the system to guide the student into the proper equations, we need to account for the state of the modeling process, storing new facts discovered by the student and automatically providing next-step hints to him/her. All this took much time that originally planned and forced us to concentrate in the novel challenge (modeling) and keep aside the solving part.

Highlights

- We developed a tool that runs on a Scala shell for constructing simple word problems, sentence by sentence, using one of the four languages supported: Catalan, English, Spanish and Swedish. It checks that the sentences written so far are consistent and complete to make a problem and saves it as Prolog code with comments in GF.
- We developed an assitant that runs on a text terminal and engages the student in a dialog in one of the aforementioned languages. This dialog

starts with the statement of the problem and the proceeds by providing hints on how to do next or answering questions about the information that has been discovered. The process ends when the student provides an equation that captures the relevant information to solve the problem. Then, the system delivers the solution in natural language.

Deviations from Annex I

- We could not use the grammars of WP4 as stated since the reasonong and language are different. In the Query Technologies worpackage, questions are about objects in classes having properties, while in our case the questions are about cardinalities of sets of objects. On the other hand, we departed form the query/answer form and went into the more general structure of a dialog. All this required new grammars.
- Time constraints, as mentioned above, forced us to leave the integration of the solving step into the prototype. Apart from this, a vital component that mediated the communication between the GF side and the cas side (Sage simple server) was deemed obsolete by the Sage community, so it was no advisable to pursue it further until a clear standard for communicating with Sage arises. At the moment of writing this document such a candidate seems to dominate (sagecell) but still is not distributed among the standard packages of Sage (and fails to install in some platforms for the last version of Sage (5.9)).

Use of resources: in accordance with Annex I.

WP7 Case study: patents

Summary of progress

The aim of "WP7:Patents Case Study" was to create a prototype for automatic translation and multilingual retrieval of patents. The online prototype is publicly available at: http://molto- patents.ontotext.com/.

This patents case study has set up the grounds where to put together several technologies in order to come up with a useful platform for multilingual patent retrieval system. The main challenges addressed in the prototype are a) to translate semantically enriched patent documents, including the original markup, b) to design the mechanisms to enable the multilingual indexing and retrieval of the patents, c) to define and develop a query language and the query grammar to enable a user-friendly interaction with the system, and d) to set up an on-line application for retrieval of patent document that serves as a testbed of our work.

The patents prototype combines semantic annotations, retrieval techniques and two different approaches for machine translation. The integration of different translation methodologies into the system has been crucial to increase its capabilities and make possible extended features and functionalities, with respect to preliminary version of the system.

For the massive translation of text, a statistical system has been trained and adapted to translate the text and transfer the semantic annotations into the target languages. One of the challenges in this task was to come up with

a mechanism to translate the semantics of the source texts to the target files. As a result, the patent documents are semantically enriched and translated using the statistical system. Then, the multilingual documents are used to feed the databases and indexes of the retrieval system. What remains as a future challenge is the use of these annotations to still increase either the accuracy of the annotations or the quality of the translations.

On the other hand, a rule-based system is built in order to translate from (controlled) natural language to the semantic query language (SPARQL), in the interface. The GF has been proved an efficient way of generating the SPARQL queries, as if it was "Yet Another Query Language". In other words, it allows to translate a natural language query from the user's language to SPARQL, which makes the system accessible to a broader community rather than just skilled users. This automation facilitates also the interoperability between the query grammar and the ontologies and speeds up the development and maintenance of the querying subsystem.

Finally, the patent prototype is not comparable with the interfaces exposed by the European Patent Office, namely because they were conceived for different purposes. Nonetheless, the MOLTO patents prototype demonstrates that a patents retrieval system that addresses multilingualism by means of automatic translation techniques is commercially viable.

Highlights -The preliminary version of the prototype, described in Deliverable 7.1 had only original patent documents in the databases and the system was only available in English and French.

- A complete version of the prototype, described in Deliverable 7.2, included resources also for German, and patent documents translated using the Statistical Machine Translation (SMT) system trained on the domain, and described in Deliverable 5.2.
- The news introduced with respect to previous versions of the prototype are: 1. A new process for statistical-based translation of patents that allows to transfer the semantic annotations and the original mark-up in the source documents to the target language.
    1. The development of the patent translator API in order to integrate the translation system into remote applications, such as online patent translation in the GF cloud.
    2. The updates on the retrieval architecture in order to improve the response time, such as snippeting. + A new querying approach for SPARQL generation based on the grammar – ontology interoperability automation, driven by the Grammatical Framework.
    3. A new query grammar for the biomedical patents domain, which has been improved in terms of coverage and compliance to the patent domain ontology that is behind the information retrieval system.
    4. The new functionalities integrated in the user interface in order to improve the usability of the application, such as the integration of the free-text search as a back-off mechanism for the query language, based on free text search.
    5. Some updates on the on-line user interface that address usability

aspects and further functionalities.

Deviations from Annex I

- This workpackage started with six months of delay because the WP leader, Matrixware, left the MOLTO Consortium during Month 3. After the rescheduling, the tasks related to this workpackage were kept up to date according to the calendar. The final version of the prototype was agreed to be delayed till M36 due multiple dependencies with other workpackages. The new calendar allowed to incorporate the latest developments (grammar and ontologies interoperability in WP4 and hybrid translation from WP5), in the final demoed applications.
- The main objectives of the work package werefulfilled:
  - i) create a commercially viable prototype of a system for MT and retrieval of patents in the bio- medical and pharmaceutical domains,
  - ii) allowing translation of patent abstracts and claims in at least 3 languages
  - iii) exposing several cross-language retrieval paradigms on top of them.

Use of resources: in accordance with Annex I.

WP8 Case study: cultural heritage

Summary of progress

The multilingual Semantic Web system covers semantic data from the Gothenburg City Museum database and DBpedia. The grammar enables automatic coherent descriptions of paintings and answering to queries over them in 15 languages for baseline functionality and in 5 languages with an extended semantic coverage. The system contains an automatic process for translating museum names from Wikipedia. The process can be easily extended to translate names of painters, places, etc. The system provides a public SPARQL endpoint against which the user can explore the knowledge base with manually written natural language queries.

Highlights

- We created the Museum Reason-able View where several ontologies were linked, including: the CIDOC-CRM, the Painting ontology and the Museum Artifacts Ontology (MAO).
- We build an ontology-based system for communication of museum content on the Semantic Web and made it accessible in 15 languages. The multilingual system automatically generates coherent Wikipedia-like articles. It has been made available online for cross-language retrieval and representation using Semantic Web technology.
- We were able to reuse the query technology that has been developed in WP4 and adapt it successfully to our needs.
- We extended the semantic coverage of the grammar to five languages and demonstrated the benefits of exploiting a modular approach in the context of multilingual Semantic Web.

Deviations from Annex I

- The time-line of the work has been shifted from M30 to M39

Use of resources: in accordance with Annex I.

WP9 User requirements and evaluation

Summary of progress

Due to the progress of other work packages, the actual evaluation work was started at Spring 2013. Some of the evaluations were made within work packages, for instance the patent cases (WP7) were evaluated with automatic evaluation metrics, and the semantic multilingual wiki (WP11) was evaluated internally for usability. WP9's contribution to the project is translation quality evaluation with native or near-native speakers.

In the evaluations, human evaluators were presented with translations by MOLTO tools and references by other MT systems (Google, Bing, Systran), and they chose the most adequate, either for post-editing or to accept as such. From these results we calculated error rates, and in addition, the percentages to what extent the evaluators preferred MOLTO translations over other systems. The results vary between languages and use cases, but in general, both automatic evaluation metrics and the percentage of the evaluators' preferred translations suggest that MOLTO method fares better in the chosen domains.

During the evaluations, some errors were detected and the grammars in question were sent to be corrected. The time and effort needed to fix the languages that get the poorest results is another factor which is favorable to MOLTO tools: a systematic fix in the grammars corrects all instances of an erroneous construction.

Some methodological issues about the qualitative evaluation were raised during the project, especially concerning the evaluation of Phrasebook. MOLTO's goal has been publishable quality automatically, but the evaluation results have been less than perfect—however, this doesn't mean that the results are incorrect, but simply that there are many ways to say the same thing, and an evaluation method that compares an edit distance to a reference doesn't capture the whole picture. This discrepancy between the human perceptions of quality and post-editing operations is discussed in the project deliverable, and has been a topic of two conference papers by Maarit Koponen, one between M31-M39 period in AMTA 2012 Workshop on Post- editing Technology and Practice, and one presentation at the XI Symposium on Translation and Interpreting: Technology and Translation in Turku, Finland.

Highlights

- Deliverable D9.2 published
- Development of methodology of evaluating limited domain publishing quality systems

Deviations from Annex I. none

Use of resources: in accordance with Annex I.

WP10 Dissemination and exploitation

Summary of progress

The major work has been to produce the final deliverables for this work-package, a report on dissemination and exploitation and the final version of the MOLTO web services. In order to produce these, we have tweaked the website and added a number of ways to generate and view the publication activity of the Consortium. Part of the work has also included the delivery of an archival version of the software prototypes as bibliographical items, with describing metadata, on the project's publication list and on a devoted page: http://www.molto-project.eu/view/biblio/type/Software.

We have checked the Open Access policy of the partners and requested the publication on OAI-PMH compatible repositories. The listing of such archives is documented in Deliverable D10.4.

The presence and dissemination of MOLTO via social sites has been constant throughout the lifetime of MOLTO and in the last period we have started to plan how to sustain the MOLTO Community after the project's end. We have been testing various platform, most recently a Google+ Community, where we also streamed the talks from the final Open Day and archived them on YouTube.

The final demonstrations are reachable from the website and they are accompanied by videos in order to supply documentation also in the far future, when the technologies will be obsolete and not available any longer.

Proper documentation and archiving of all these resources is underway. The resources produced by the project are very different in nature and present a challenge in terms of sustainability and future accessibility. They include software (often depending on third-party libraries), technical reports and publications in digital and/or printed form, and multimedia material. We intend to store all of these on an archival media however it is not clear how persistent they will remain.

Highlights

- Deliverable D10.3 MOLTO web service, final version documents the software that allows to deploy web services for any GF application grammar and describes some of the sample grammars made available online at http://www.molto-project.eu/cloud/gf- application-grammars
- D10.4 MOLTO Dissemination and Exploitation Report contains an analysis of the possible venues of impact of MOLTO from an exploitation point of view. The industrial partners came together to discuss how to adopt the tools and technologies developed during the project, how to ensure a sustainable future for them that would benefit all partners in the Consortium.
- Several new demonstration web sites have been linked from the project's web page: patent translation, query for museum artifacts, query for patents, multilingual semantic wiki, cloud services, translation services and the term factory.

- Some deliverables are published in the MOBI format. The production of a e-Book reader compatible version has been discussed and the final result checked for readability within the Consortium.

Deviations from Annex I: none
Use of resources: in accordance with Annex I.

WP11 Multilingual semantic wiki

Summary of progress

We continued working on our two main projects: (1) developing ACE-in-GF (multilingual grammar of ACE) and (2) developing AceWiki-GF (multilingual CNL-based semantic wiki).

ACE-in-GF was extended to almost all the languages supported in the GF resource grammar library (˜20 languages), although only the languages reported in D11.1 are fully implemented and tested.

The main work on AceWiki-GF was completed, and reported in D11.2 and a ESWC 2013 conference paper. Smaller extensions and improvements continue.

In the last 5 months of the project we focused on the evaluation of both ACE-in-GF and AceWiki-GF. The design and results of both of these evaluations are reported in D11.3.

Events

- Norbert E. Fuchs and Kaarel Kaljurand were invited to participate and speak at the 3rd Symposium on Advances in KRDB Technologies (SAKT-2012), Bolzano, 19-20 Nov 2012
- Kaarel Kaljurand gave a 2-day GF course at BeInformed (Dec 2012)
- Kaarel Kaljurand and Tobias Kuhn presented a paper at ESWC 2013, Montpellier 26-30 May 2013

Links to online resources

- ACE-in-GF project website: https://github.com/Attempto/ACE-in-GF
- AceWiki-GF demo wikis: http://attempto.ifi.uzh.ch/acewiki-gf/
- AceWiki-GF source code: https://github.com/AceWiki/AceWiki

Highlights

- Multilingual GF-based ACE grammar (ACE-in-GF)
- AceWiki-GF
- Evaluations of ACE-in-GF and of AceWiki-GF (translation accuracy, user acceptance, both with good results)
- RACE extensions (arithmetic, improved wh-queries)
- paper on ACE-in-GF and AceWiki-GF accepted at ESWC 2013 (25% acceptance rate)

Deviations from Annex I: none
Use of resources: in accordance with Annex I.

WP12 Interactive knowledge-based systems

Summary of progress

This period we continued the work on the further adoption of GF in the Business Process Platform of Be Informed. On of our goals to leverage the adoption of GF was to create a framework in which models could automatically be verbalized. Domain experts usually do not have a background in modeling and thus checking whether a rule or law is modeled correctly usually proves to be a difficulty for them. Be Informed wants to take away these barriers by creating verbalizations of their models. These verbalizations however should not be only a textual representation the models, but it wants the possibility to create verbalizations of the same models for a set of the distinguished tasks.

In order to do this Be Informed created the 3D framework together with the University of Bielefeld. An article on this work will be published in "Jeroen van Grondelle, Christina Unger: A 3-Dimensional Paradigm for Conceptually-scoped Language Technology" in Towards the Multilingual Semantic Web, Paul Buitelaar and Philip Cimiano, eds., Springer, Heidelberg, Germany, 2013. This orthogonal modularization supports specification of the conceptualization and lexical information per dimension, i.e. specifying domains independent from tasks and vice versa. The dimensions can then be freely combined by choosing the particular domains, tasks and languages supported for a specific application.

While the task grammars are written once by hand, each of the domain grammars is created automatically from a Be Informed or OWL ontology and plugged right into the grammars already created for the framework. In order to create these domain grammar automatically Be Informed created three verbalizers, each one with its own heuristics to create verbalizations.

In an evaluation the likelihood of the verbalizations created with grammars from these verbalizers were compared to the verbalizations created by the velocity templates, the verbalizer which is currently implemented in the Be Informed product suite. The results show that the GF based verbalizers are better than these velocity verbalizers.

Work on the adoption and evaluation has been finished and reported in D12.2 (http://www .molto-project.eu/biblio/deliverable/d122-user-studies-bis-exp...).

Events

- September 17th 2012, Utrecht, Research Workshop on the use of business models by multiple audiences. Participants : Be Informed Customers from the public sector (SVB, Bureau Forum Standaardisatie, Dutch Prosecution Office, Immigration Department) and members from Molto and Monnet.
- December 11-12 2012, Course GF at Be Informed Apeldoon; course given by Kaarel Kaljurand.
- December 13 and 14, 2012, PortDial members from Bielefeld met with Aarne Ranta (MOLTO), Jeroen van Grondelle, Frank Smit and Jouri Fledderman from Be Informed, and John McCrae (lemon) in order to discuss the mapping from ontology-lexica to grammars, as well as the modular combination of induced domain grammars with dialog task grammars.

- February, 6-8, 2013; April 4-5 2013; Subsequent Lemon/GF workshops at Bielefeld with Christina Unge, John McCrae from PORTDIAL and MONNET and Jouri Fledderman/Jeroen van Grondelle from Be Informed.

Deviations from Annex I: none
Use of resources: in accordance with Annex I.

Table 1: Deliverables

The associated documents are available online:

- http://www.molto-project.eu/workplan/deliverables, via partner login, which is granted to the project officer and the evaluators.
- http://www.molto-project.eu/view/biblio/deliverables, for anyone, without login, but restricted to the public deliverables

D nr	title	date	dissem.	nature
D10.1	Dissemination plan, with monitoring and assessment	1 June 2010	Cons.	Report
D10.2	MOLTO web service, first version	1 June 2010	Public	Prototype
D9.1	MOLTO test criteria, methods and schedule	1 Sep 2010	Public	Report
D1.2	Periodic management report 1	1 Oct 2010	Cons.	Report
D4.1	Knowledge Representation Infrastructure	1 Nov 2010	Public	RegPubl
D2.1	GF Grammar Compiler API	1 Mar 2011	Public	Prototype
D1.3	Periodic management report 2	1 April 2011	Cons.	Report
D4.2	Data Models, Alignment Methodology, Tools and Documentation	1 May 2011	Public	RegPubl
D2.2	Grammar IDE	1 Sep 2011	Public	Prototype
D3.1	MOLTO translation tools API	1 Sep 2011	Public	Prototype
D5.1	Description of the final collection of corpora	1 Sep 2011	Public	RegPubl
D6.1	Simple drill grammar library	1 Sep 2011	Public	Prototype
D8.1	Ontology and corpus study of the cultural heritage domain	1 Sep 2011	Public	Other
D4.3	Grammar-Ontology Interoperability	30 Sep 2011	Public	Prototype
D1.4	Periodic management report 3	1 Oct 2011	Cons.	Report
DX.2	Annual public report	15 Nov 2011	Public	Report
D7.1	Patent MT and Retrieval Prototype Beta	1 Dec 2011	Public	Prototype
D6.2	Prototype of commanding CAS	1 Feb 2012	Public	Prototype
D3.2	MOLTO translation tools prototype	1 Mar 2012	Public	Prototype
D5.2	Description and evaluation of the combination prototypes	1 Mar 2012	Public	RegPubl
D8.2	Multilingual grammar for museum object descriptions	1 Mar 2012	Public	Prototype
D1.5	Periodic management report 4	1 April 2012	Cons.	Report
D11.1	ACE Grammar Library	31 May 2012	Public	Prototype
D2.3	Grammar tool manual and best practices	30 June 2012	Public	RegPubl
D7.2	Patent MT and Retrieval Prototype	1 Sep 2012	Public	Prototype
D12.1	Requirements for BI's explanation engine	1 Sep 2012	Public	Report
D1.6	Periodic management report 5	1 Oct 2012	Cons.	Report
D6.3	Assistant for solving word problems	1 Dec 2012	Public	Prototype
D11.2	Multilingual semantic wiki	31 Dec 2012	Public	Prototype
D8.3	Translation and retrieval system for museum object descriptions	1 Mar 2013	Public	Prototype
D3.3	MOLTO translation tools / workflow manual	31 Mar 2013	Public	RegPubl
D5.3	WP5 final report: statistical and robust MT	31 Mar 2013	Public	RegPubl
D7.3	Patent Case Study Final Report	31 Mar 2013	Public	RegPubl
D1.7	Final management report	31 May 2013	Cons.	Report
D10.3	MOLTO web service, final version	31 May 2013	Public	Prototype
D10.4	MOLTO Dissemination and Exploitation Report	31 May 2013	Public	Report
D9.2	MOLTO evaluation and assessment report	31 May 2013	Public	Report
D11.3	User studies for the multilingual semantics wiki	31 May 2013	Public	Report
D12.2	User studies for BI's explanation engine	31 May 2013	Public	Report
D4.3A	Grammar ontology interoperability - Final Work and Overview	31 May 2013	Public	RegPubl

Table 2: Milestones

MS nr	title	date
MS1	15 Languages in RGL	1 September 2010
MS2	Knowledge Representation Infrastructure	1 September 2010
MS3	Web-based translation tool	1 March 2011
MS4	Grammar-ontology interoperability	1 October 2011
MS5	First prototypes of the cascade-based combination models	1 October 2011
MS6	Grammar tool complete	1 March 2012
MS7	First prototypes of hybrid combination models	1 March 2012
MS8	Translation tool complete	1 September 2012
MS9	Case studies complete	1 December 2012
MS11	Prototype of semantic wiki with ACE Grammars is functional	1 July 2012
MS12	BI's explanation engine is functional	1 July 2012