OPENMT-2:

Traducción automática híbrida y evaluación avanzada

http://ixa.si.ehu.es/openmt2

TIN2009-14675-C03-01 y 02

Kepa Sarasola
Ixa Group. Euskal Herriko Unibertsitatea
Lluís Màrquez
TALP, Universitat Politècnica de Catalunya

Jornadas de Seguimiento de Proyectos MICINN

5 de Septiembre de 2011, A Coruña

Grupo Ixa (UPV-EHU)

http://ixa.si.ehu.es

Grupo Ixa de procesamiento de la lengua

Universidad del País Vasco / Euskal Herriko Unibertsitatea

- Doctores: 7
- No doctores: 4
- Funcionarios ó profesorado permanente : 5
- Otro profesorado: 4
- Becarios FPI: 0
- Otros becarios: 2

NLP Research Group, TALP, UPC

http://nlp.lsi.upc.edu

Universidad Politécnica de Cataluña

- Doctores: 7
- No doctores: 4
- Funcionarios: 3
- Otro profesorado: 3
- Becarios FPI: 1
- Otros becarios: 4

Objetivos (Goals)

Machine Translation (MT) technology to generate:

- robust, high-quality combined MT systems.
- improved evaluation metrics and methodologies.

5 main areas:

- Collection, annotation and exploitation of multilingual corpora.
- Further development of the current single-paradigm translation systems.
- Pre-edition, post-edition and system improving based on collaboration with a web2.0 community.
- Combining and hybridizing MT paradigms.
- Advanced evaluation for MT.

Four different languages: English, Spanish, Catalan and Basque.

Objetivos (Goals)

The main innovative points of the proposal are:

- Design of hybrid systems combining traditional linguistic rules, example-based methods and statistical methods
- Use of advanced morphological, syntactic and semantic processing in MT
- Research taking advantage of a web community of posteditors (2.0 oriented)
- Use of more realistic metrics in MT evaluation, improving its reliability
- Creation of a web-based suite of MT systems and tools for evaluation and error analysis

Objetivos (Goals)

The consortium is composed by two universities:

- University of the Basque Country (UPV/EHU)
- Technical University of Catalonia (UPC)

Several companies and foundations with activities in closely related areas serve as supervision EPOs for the project:

- Foundations: Elhuyar. i2CAT, eu.wikipedia.
- Companies: Eleka, Imaxin, Semantix, Translendium SL

Objetivos alcanzados

- Collection, annotation and exploitation of multilingual corpora **(%100)**.
  - 10 corpora and tools availables in the project website http://ixa.si.ehu.es/openmt2/baliabideak_html
- Further development of the current single-paradigm translation systems. **(%100)**
  - http://ixa2.si.ehu.es/openmt-demo
- Combining and hybridizing MT paradigms. **(%70)**

Objetivos parcialmente alcanzados

- Pre-edition, post-edition and system improving based on collaboration with a web2.0 community. **(%50)**
  - Ongoing Wikiproject (openmt2 - eu.wikipedia) to collect 100.000 words corpus with hand made corrections for MT outputs (2011 December).
  - Experimentation in 2012.
- Advanced evaluation for MT **(%50)**
  - Creation of Asiya opensource evaluation system. http://www.lsi.upc.edu/~nlp/Asiya/
  - Adaptation to enable its use with Basque (2012)
  - The main researcher is working now at Google

Producción Científica

Número de JCR y Número de congresos CORE aceptados.

2010 y 2011

59 publicaciones en total: http://ixa.si.ehu.es/openmt2/argitalpenak_html

24 sobre Traducción Automática + 35 sobre Tecnología de la Lengua

Número de JCR/Scopus:	**4**
(Machine Translation Journal, Natural Language Engineering, Applied soft computing)	
Número de congresos CORE aceptados:	**16**
(ACL, Coling, EAMT, MT-Summit, AMTA, EMNLP, Interspeech)	
Número de artículos en revistas o libros:	8
Número de artículos en congresos con 3 evaluadores y (admisión<%50) :	16
Número de otras publicaciones :	15

Producción Científica

Listado de las 5 publicaciones más relevantes.

Aingeru Mayor, Iñaki Alegria, Arantza Díaz de Ilarraza, G. Labaka, M. Lersundi, K. Sarasola 2011.
Matxin, an open-source rule-based machine translation system for Basque.
Machine Translation Journal. volume 25, number 1, July 2011.

Jesús Giménez and Lluís Màrquez. 2010.
Linguistic Measures for Automatic Machine Translation Evaluation.
Machine Translation Journal, volume 24, numbers 3-4, December 2010.

Jesús Giménez and Lluís Màrquez. 2010.
Asiya: An Open Toolkit for Automatic Machine Translation (Meta-)Evaluation.
The Prague Bulletin of Mathematical Linguistics, No. 94, 2010.

Cristina España, Gorka Labaka, Lluís Màrquez, Arantza Díaz De Ilarraza and Kepa Sarasola 2011.
Hybrid Machine Translation Guided by a Rule-Based System.
XIII Machine Translation Summit, Xiamen, China

Pighin, Daniele, and Màrquez Lluís 2011.
Automatic Projection of Semantic Structures: an Application to Pairwise Translation Ranking.
Fifth Workshop on Syntax, Semantics and Structure in Statistical Translation (SSST-5),
ACL, Portland, Oregon, 2011.

Para más información: http://ixa.si.ehu.es/openmt2/argitalpenak_html

Producción Científica

Eventos organizados

Organización conjunta de un workshop internacional

"Using Linguistic Information for Hybrid Machine Translation"

Barcelona, 18-Noviembre- 2011.

http://ixa2.si.ehu.es/lihmt2011/

Colaboración con investigadores punta en Traducción Automática a nivel mundial:

- Comité de programa.

  David Farwell, Josep M. Crego (LIMSI/CNRS, France), Chris Dyer (Carnegie Mellon University, US), **Marcello Federico** (Fondazione Bruno Kessler, Italy), **Mikel Forcada** (University of Alacant, Alicante), **Adrià de Gispert** (University of Cambridge, UK), **Kevin Knight** (Information Sciences Institute, US), **Philipp Koehn** (University of Edinburgh, UK), Patrik Lambert (Universiteé du Maine, France), **José B. Mariño** (Technical University of Catalonia, TALP, Barcelona), **Hermann Ney** (RWTH-Aachen, Germany), Aarne Ranta (Chalmers University of Technology, Gothenburg, Sweden), Marta R. Costa-jussà (Barcelona Media, Barcelona), Felipe Sánchez-Martínez (University of Alacant, Alicante), **Dekai Wu** (Hong Kong University of Science and Technology, China)...

- Ponentes invitados.

  - **Ondřej Bojar** (Charles University, Czech Republic)

  - **Alon Lavie** (Carnegie Mellon University, Pennsylvania)

  - **Lucia Specia** (University of Wolverhampton, UK)

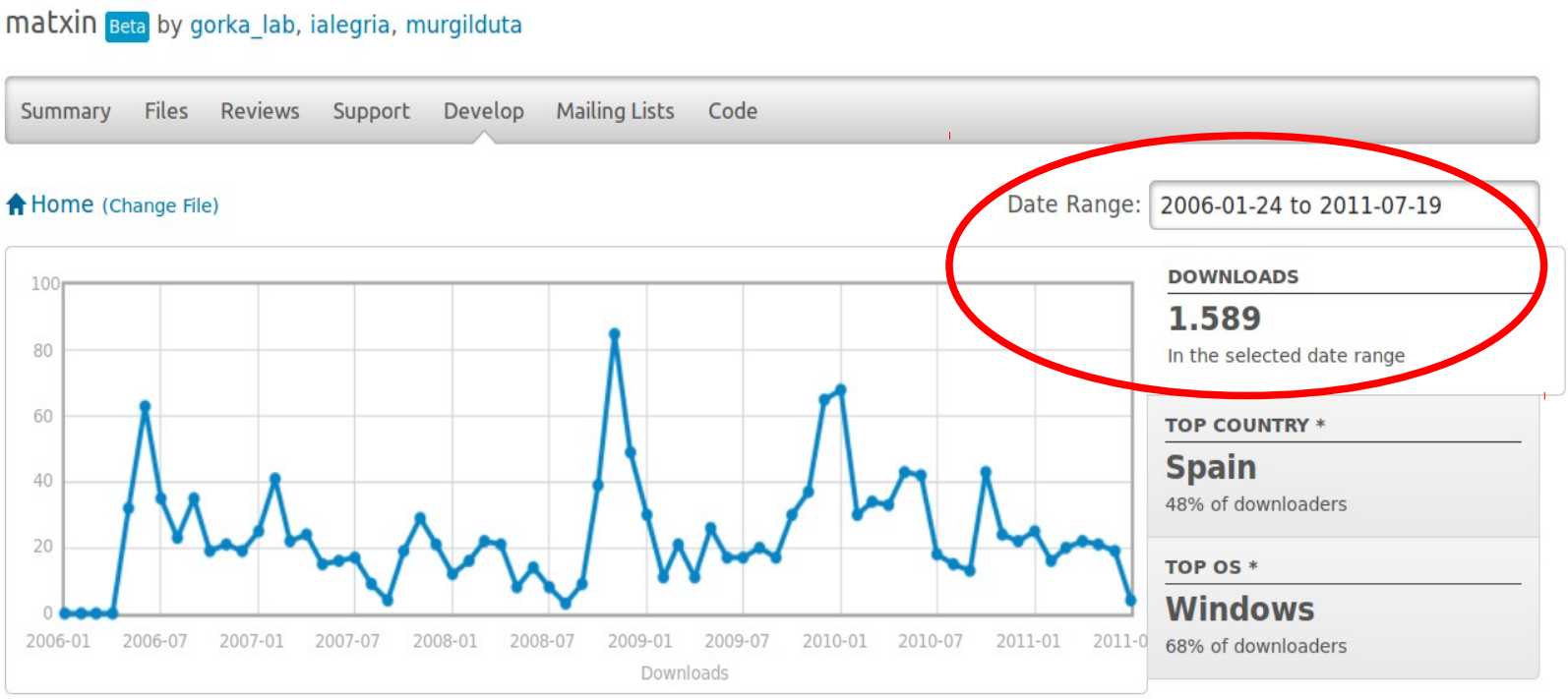
Transferencia

Uso de resultados del proyecto en la industria.

Matxin. Software registrado (SS-0409-2008)

1589 descargas del motor de traducción Matxin en Sourceforge.

http://sourceforge.net/projects/matxin/files/stats/timeline?dates=2006-01-24+to+2011-07-19

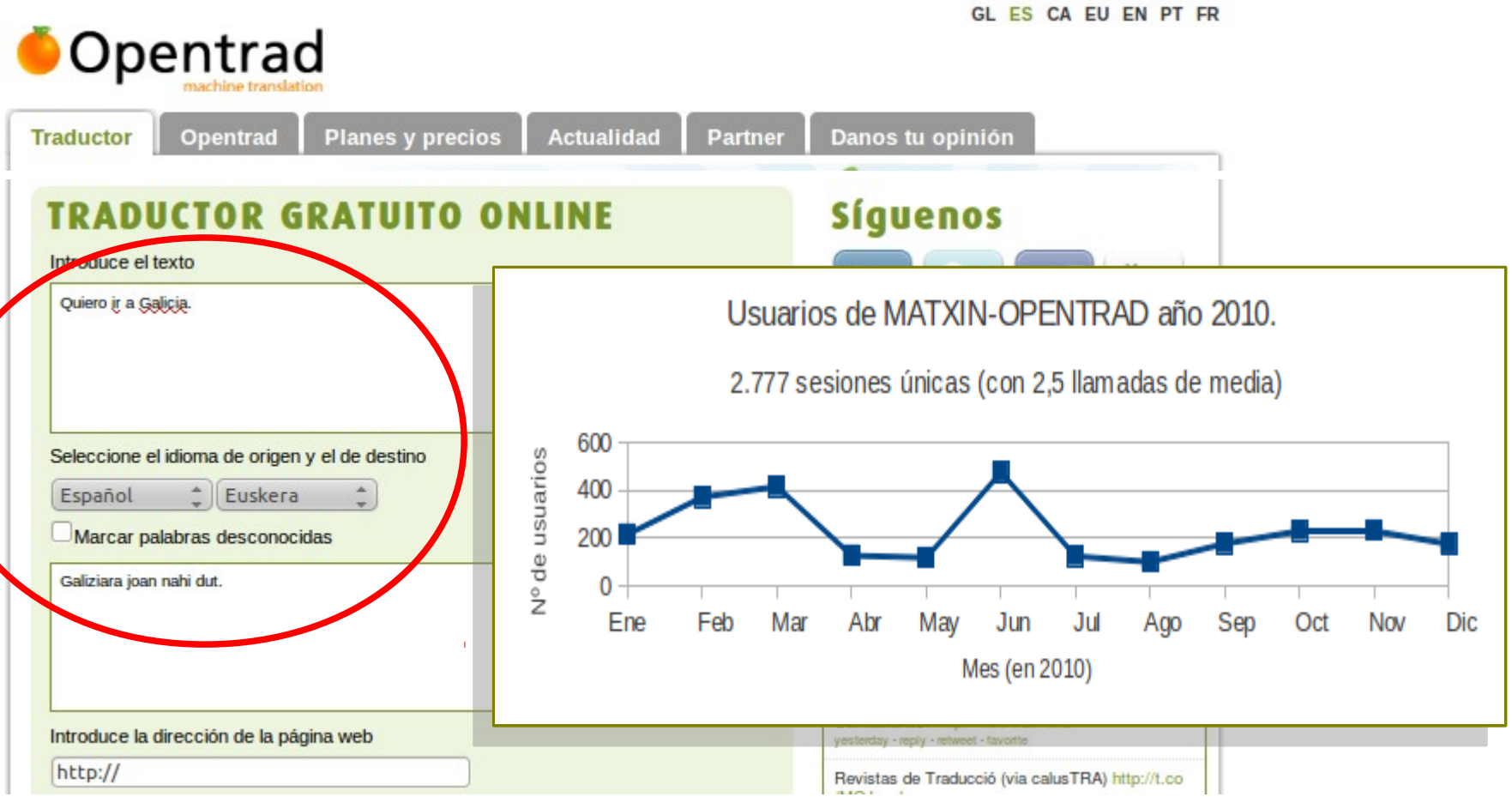


Transferencia

Uso de resultados del proyecto en la industria.

Matxin. Software registrado (SS-0409-2008)

Traducción en web castellano-euskara (Matxin-Opentrad, http://www.opentrad.com/es



Transferencia

Uso de resultados del proyecto en la industria.

Matxin. Software registrado (SS-0409-2008)

Uso del traductor Matxin dentro de un Wikiproyecto web2.0 para incrementar la *Wikipedia en euskara* con 100 artículos nuevos en el 2011 (100.000 palabras)

Transferencia

Uso de resultados del proyecto en la industria.
Asiya. *Sistema de evaluación* de código abierto.

http://www.lsi.upc.edu/~nlp/Asiya/

The Asiya Open Toolkit for Automatic MT (Meta-)Evaluation

Asiya has been designed to assist both system and metric developers by offering a rich repository of metrics and meta-metrics. Asiya has been developed at TALP Research Center NLP group , in Universitat Politècnica de Catalunya, as an evolution, extension, refactoring, and finally a replacement for its predecessor, IQMT.

Download

Latest development version may be downloaded through subversion.

- svn co http://svn-rdlab.lsi.upc.edu/subversion/asiya/public asiya

Everyone is allowed to check out (User: reader, Password: reader).

Asiya is also open to public contribution. If you feel like helping us in the development, please, e-mail us at jgimenez so we grant you with the necessary permissions.

The Asiya framework is released under the GNU Lesser General Public License (LGPL) of the Free Software Foundation. This means that it may be linked to and used by commercial software packages. But the license also enforces that any changes or improvements made to the library must be redistributed under LGPL terms.

Resultados: Formación

Tesis de doctorado leídas (2010-2011)

Gorka Labaka (Marzo 2010, Doctorado europeo)
EUSMT: Incorporating Linguistic Information into SMT for a Morphologically Rich Language. Its use in SMT-RBMT-EBMT hybridation.
Kepa Sarasola y Arantza Diaz de ilarraza.

Beñat Zapirain (Febrero 2011, Doctorado europeo).

Semantic Role Labeling: Role Inventories and Selectional Preferences / Rol Semantikoen Etiketatze Automatikoa: Rol Multzoak eta Hautapen Murriztapenak
Eneko Agirre Bengoa y Lluís Màrquez Villodre

Bertol Arrieta (Enero 2010)
Azaleko sintaxiaren tratamendua ikasketa automatikoko tekniken bidez: euskarako kateen eta perpausen identifikazioa eta bere erabilera koma-zuzentzaile batean.
Iñaki Alegria y Arantza Diaz de Ilarraza

Resultados: Formación

Tesis de máster o PFC

- **Itzulpen-sistema desberdinen irteera bistaratzeko interfazea.**
  Aritz Sala Mayor
  2010-04-15
- **Postedizio-interfazearen diseinua eta inplementazioa itzulpen automatikoko sistemen ebaluaziorako.**
  Mireia Lecea Urrutia.
  2010-09-17. Matrícula de honor
- La FPI del proyecto Eva Martínez ha completado los cursos de Máster satisfactoriamente durante el curso académico 2010-2011. Ahora está empezando la tesis de máster con finalización prevista para Febrero de 2012: "**Robust POS tagging for Machine Translation**"

Colaboraciones internacionales

Grupo Ixa (EHU) http://ixa.si.ehu.es/Ixa/Nazioarteko%20harremanak

Colaboraciones internacionales

Grupo TALP (UPC)

Personas y universidades con las que se colabora

- **Grupo de Investigación del DISI de la Universidad de Trento (Alessandro Moschitti)** y la **Fondazione Bruno Kessler de Trento (profesor Marcello Federico)**. Lluís Màrquez realizó una estancia corta en Agosto-Septiembre de 2010. Como consecuencia de esta relación el grupo UPC ha incorporado en octubre de 2010 un investigador post-doctoral proveniente de estos grupos, Daniele Pighin.
- Contactos con los investigadores **Lucia Specia (University of Wolverhampton) y Mikel Forcada (Universidad de Alicante)** han propiciado las visitas en la primavera de 2011 de un estudiante de doctorado (Mariano Felice, 5 meses) y un investigador post-doctoral (Felipe Sánchez, 3 meses) que presumiblemente se unirán al grupo en la línea de investigación sobre medidas de evaluación y su aprendizaje.
- Los grupos de investigación de los **proyectos europeos FAUST y MOLTO.** Sobre todo **University of Cambridge (Bill Byrne), Charles University, Prague (Jan Hajic) y University of Gothenburg (Aarne Ranta).**

Colaboraciones internacionales

Proyectos internacionales financiados

Proyectos europeos con temática de MT

- **FAUST**: Feedback Analysis for User Adaptive Statistical Translation (FP7-ICT-2009-4-247762, 2010-2012). http://www.faust-fp7.eu/
- **MOLTO**: Multilingual On-Line Translation (FP7-ICT-2009-4-247914, EC, 2010-2012). http://www.molto-project.eu/

Otros proyectos europeos

- **PATHS** - Personalised Access To cultural Heritage Spaces (STREP, ICT-2011-270082, 2011-2012, http://www.paths-project.eu)
- **KYOTO**: knowledge yielding ontologies for transition-based organizations (STREP, ICT-2007-211423, 2007-2010, http://www.kyoto-project.eu)

COORDINACIÓN
Evaluación de la coordinación

- Creación de sitio web público del proyecto.
  http://ixa.si.ehu.es/openmt2
- Creación de sitio web privado del proyecto.
- Cuatro reuniones de coordinación
- Estancia de 6 meses del IP de la UPC en Donostia.
  => Creación de un prototipo híbrido de traducción
  => Dos publicaciones conjuntas
- Organización conjunta de un workshop internacional.
  "Using Linguistic Information for Hybrid Machine Translation"
  Barcelona 2011.
  http://ixa2.si.ehu.es/lihmt2011/

COORDINACIÓN

Publicaciones conjuntas

- **Hybrid Machine Translation Guided by a Rule–Based System**.
  C. España, G. Labaka, L. Marquez, A. Diaz de Ilarraza, K. Sarasola
  MT-summit. China 2011.

- **Divergences between manual and automatic evaluation.**
  C. España, G. Labaka, L. Marquez, A. Diaz de Ilarraza, K. Sarasola
  Submitted to the LIHMT Workshop. Barcelona 2011.

Previsión de desarrollos futuros

- Combining and hybridizing MT paradigms. **(%30)**
  - Experimenting with new features to improve the results of the hybrid system.
  - Use of manual evaluation to optimize hybridization.
- Pre-edition, post-edition and system improving based on collaboration with a web2.0 community. **(%50)**
  - Ongoing Wikiproject (openmt2 - eu.wikipedia) to collect 100.000 words corpus with hand made corrections for MT outputs. (December 2011)
  - Experimentation with statistical postedition.
- Advanced evaluation for MT **(%50)**
  - Extension of the Asiya suite with evaluation measures for Basque.
  - Asiya Interface (graphical and web-based) for error analysis.
  - Adequacy-based confidence estimation measures by projecting syntactic/semantic structures.

OPENMT-2:

Traducción automática híbrida y evaluación avanzada

http://ixa.si.ehu.es/openmt2
(TIN2009-14675-C03-01 y 02)

Kepa Sarasola
Ixa Group. Euskal Herriko Unibertsitatea
kepa.sarasola@ehu.es

Lluís Màrquez
TALP, Universitat Politècnica de Catalunya
lluism@lsi.upc.edu