Jornada de Seguimiento de Proyectos, 2009
Programa Nacional de Tecnologías Informáticas

OpenMT: Open Source Machine Translation Using Hybrid Methods

TIN2006-15307-C03

Kepa Sarasola, Iñaki Alegria *
Ixa Group. Euskal Herriko Unibertsitatea

Núria Castell, Lluís Màrquez †
TALP, Universitat Politècnica de Catalunya

Nerea Areta, Xabier Saralegi ‡
Elhuyar

Abstract

The main goal of the OpenMT project is the development of open source machine translation architectures based on hybrid models and advanced syntactic–semantic processors. These architectures combine the three main Machine Translation (MT) frameworks, *Rule-based* (RBMT), *Statistical* (SMT) and *Example–based* (EBMT), into hybrid systems. Defined architectures and results will be open source, allow for a rapid development and adaptation of new advanced machine translation systems for other languages. The project deals with four different languages: English, Spanish, Catalan and Basque. The translation systems developed for all language pairs will be internally evaluated using a rich set of linguistic metrics, and in different international evaluation campaigns.

Keywords: Machine Translation, MT Evaluation, SMT, RBMT, EBMT

1 Goals of the Project

The main goal of OpenMT is the development of Open Source Machine Translation Architectures based on hybrid models and advanced syntactic–semantic processors. More specifically:

1. Defining open source architectures for machine translation.
2. Combining three different MT paradigms into a hybrid system: Rule-Based (RBMT), Statistical (SMT) and Example-Based (EBMT).

*Email: {kepa.sarasola,i.alegria}@ehu.es
†Email: {castell,lluism}@lsi.upc.edu
‡Email: {nereaa,xabiers}@elhuyar.com

3. Integration of syntactic and semantic processing in machine translation. Among others, we will work with Semantic Role Labeling systems and Word Sense Disambiguation adapted to the selection of translation preferences.

4. Evaluation. A general open-source evaluation tool and an evaluation corpus for each language pair will be generated. The system will be evaluated merging different criteria and using different languages and domains.

Without abandoning the other goals, the main effort of the project so far has been devoted to the second and fourth. For the current year, the most important work in progress is the evaluation of the different translation prototypes generated during the project.

IXA (EHU) IXA (EHU) The OpenMT-EHU subproject will develop a MT hybrid system combining RBMT, SMT and EBMT technology. The possible combination will follow a semi-automatic process using modules for searching the most suitable translations for NP or PP segments (those modules will work at different level: lexical entry, phrase or phrase pattern and clause), and modules for reordering phrase candidates in the target sentence, but that will turn to single RBMT modules when the results from corpus are insufficient. An additional challenge is to enrich the RBMT hybrid technology with semantic information captured by machine learning from the EBMT systems.

TALP (UPC) The OpenMT-UPC subproject will develop the SMT technology needed for the combined MT systems produced in this project. The main challenge is to enrich the pure SMT technology with syntactic–semantic information captured by machine learning and also by the incorporation of linguistically-based transfer rules from the EBMT systems. Also, the UPC team will be in charge of leading the evaluation work package, including the improvement and adaptation of the IQMT framework for MT evaluation.

Elhuyar The participation of Elhuyar is oriented towards three main areas: 1) Basic infrastructure for Machine Translation systems: resources and tools; 2) Evaluation; 3) Exploitation and Dissemination

2 Project Progress and Achievements

In this section we present the main results obtained in OpenMT. The section is structured in six parts corresponding to the main work packages of the project.

2.1 Basic tools and resources

We have developed a web tool for managing and storing parallel corpora [51]. In addition to this, we have created several new Basque Spanish (es-eu) reference corpora by collecting translation memories. They are restricted to domains of newspapers, divulgative texts, technical manuals, environment and public administration.

On the other hand, in order to improve the recall of the dictionary of the RBMT system we have created specific domain bilingual terminology. To do that, we have used the Elexbi

terminology extractor, which has been revised and improved in this project [2]. In addition, we have carried out preliminary experiments on terminology extraction from bilingual comparable corpora. As the results are promising, and due to the scarcity of parallel corpora we could collect for Basque, this way become a promising strategy to explore [38, 39].

For the representation of the tagged bitext we propose a corpus structure for containing: (1) translation units and the linguistic information for each unit, and (2) the whole documents with their linguistic information. The corpus structure proposed may be seen as composition of several XML documents and is based on stand off annotation model. This structure permits to work with the corpus from two points of view: as an annotated corpus with linguistic information, as well as a translation memory [10, 11].

Additionally, since one of the objectives of the project is to cope with syntactic and semantic information inside MT systems, we have devoted some effort (in parallel with another related project) in order to improve linguistic analyzers for syntactic parsing and Semantic Role Labeling in the following languages: Catalan, English and Spanish. These work has resulted in very productive research [41, 34, 33, 42, 40, 32, 44, 45].

2.2 Improving the current RBMT and SMT systems

The design and development of an open transfer based MT architecture has been improved and the result has been tested for the Spanish–Basque and English–Spanish pairs. Following the reusing philosophy inherent to the open-source software development some of the components (modules, data formats and compilers) are inherited from previous open–source projects (FreeLing and Apertium). The tools could be adapted to translating between other languages with few resources [4]. Translating from English into Basque has demonstrated that the RBMT architecture is enough flexible to translate into Basque from different languages [3]. A first version of this prototype has been developed during the project. In the near future we want to translate from Basque into other languages reusing the resources and to test the architecture with other languages (for example Quechua)

SMT prototypes have been built from parallel corpora for translating among Spanish and English into Basque [28, 43]. A translation scheme based on morphemes instead of words has been introduced in order to be able to deal with the particular agglutinative nature of Basque. Comparing RBMT and SMT systems, the automatic metrics used in evaluation indicate that the data-driven system outperforms the rule-based system on the in-domain data. On the contrary, the subjective evaluation indicates that the rule-based system outperforms the data-driven approach for both corpora.

Regarding the UPC basic SMT systems, we have carried out the following improvements: (1) We have extended the system and significantly improved its results by combining shallow-syntactic translation models based on linguistic data views (i.e., redefining the translation unit so it may contain additional linguistic information beyond the lexical level and build translation models with these enriched translation units); (2) We have explored the domain adaptation issue by porting an English-to-Spanish phrase-based SMT system trained on the political domain to the domain of dictionary definitions.

All previous research is based upon two previous works [21, 20]. The most up to date revisions are reported in [18].

2.3 Combining EBMT and RBMT

In this task we applied Spanish-to-Basque MultiEngine Machine Translation to a specific domain to select the best output from three single MT engines [6, 8]. Guided by some previous results, we decided to follow a hierarchical strategy, applying: first, EBMT translation patterns, second, SMT (if its confidence score is higher than a fixed threshold) and, finally, RBMT.

An important improvement in translation quality (according to BLEU) is observed, in connection with the improvements obtained by other systems. We obtain 193.55% relative increase for BLEU when comparing the EBMT+SMT combination with the SMT system alone, and 15.08% relative increase when comparing EBMT+SMT combination with the EBMT single strategy. Those improvements would be difficult to obtain by single–engine systems. RBMT contribution seems to be very small with automatic evaluation, but we expect that HTER–based evaluation will show better results.

Despite of working on a specific domain, we think that our translation system is a major advance in the field of language tools for Basque. The restriction of using a domain–specific corpus is imposed by the absence of large and reliable Spanish–Basque parallel corpora.

2.4 Hybrid architecture based on SMT

We performed two experiments to verify the improvements obtained in other languages by using statistical post editing (SPE) [12]. Our experiments differ from other similar works in: the use of a morphological component in both RBMT and SMT translations, and the limited size of the available corpora. Our results are coherent with previous literature, showing large improvements when using the RBMT+SPE approach on a restricted domain. Specifically, we obtain 200% improvement on BLEU scores for a RBMT+SPE approach working with Matxin RBMT system, when comparing to raw translation output. The relative improvement with respect to the single SMT system is 40%. These results are also consistent when using more general corpora, but showing smaller quality improvements.

Experiments using rules in order to correct the results from the SMT system are in progress.

2.5 Advanced Semantic Processing for Machine Translation

The research and advances on linguistic tools for semantic annotation has been reported in section 2.1. Regarding the inclusion of proper semantic information in MT we have made substantial advances in two directions: 1) The inclusion of discriminative phrase selection models in SMT, which are inspired in Word Sense Disambiguation; and 2) the usage of lexical–semantic resources (e.g., WordNet) to improve coverage and quality of SMT when applied across domains. Finally, the inclusion of Semantic Role Labeling information (in the form of semantic dependencies) is left for the third year of research.

Discriminative phrase selection models In [23, 24], we have explored the application of discriminative learning to the problem of phrase selection in Statistical Machine Translation (Spanish–English language pair). Instead of relying on Maximum Likelihood estimates for the construction of translation models, we suggest using local classifiers which are able to take further advantage of contextual information. These classifiers are trained using Word Sense Disambiguation techniques, just by recasting *word senses* as *possible translations* for a given

word/phrase. The integration of dedicated discriminative phrase translation models into the statistical framework requires further study.

The same techniques have been applied to the English–Arabic pair, obtaining comparable improvements [13, 14]. These results give more support to the appropriateness of the approach.

Usage of external lexical–semantic resources In [16], we propose a method for improving phrase–based Statistical Machine Translation by enriching the original translation model with information derived from a multilingual lexical knowledge base. Translation probabilities are estimated using a set of simple heuristics based on WordNet topology and local context. During decoding, these probabilities are softly integrated so they can interact with other statistical models. We have applied this type of domain–independent translation modeling to several translation tasks obtaining a moderate but significant improvement in translation quality. As an extension, we plan to exploit the information in the MCR so as to assist the above mentioned discriminative Phrase Selection engine, by providing additional features.

2.6 Evaluation and demonstration

The effort in this work package has focused, by now, in the development of a framework for MT automatic evaluation (IQMT, [19]), including a set of heterogeneous metrics operating at different linguistic levels. The IQMT manual was first published in [17]. The complete software and documentation is freely available through the IQMT website (see link [49]). Additional research has been devoted to the combination of metrics and the assessment of the quality of metric sets (i.e., *meta-evaluation* of metrics). The demonstration interface in the web is left for the third year of the project. In some more detail:

- In [22], we suggest using metrics for MT evaluation which take into account linguistic features at more abstract levels, as an alternative to the widespread lexical-based metrics (e.g., BLEU). We provide experimental results showing that metrics based on syntax and semantics are able to produce more reliable system rankings, specially when the systems under evaluation are of a different nature. In [18, 25] the set of metrics is enriched by considering also semantic features at the discourse level.

- Combining different metrics into a single measure of quality seems the most direct and natural way to improve over the quality of individual metrics. In [27], we study the behaviour of a non-parametric metric combination scheme, in which the above mentioned metrics are combined without having to adjust their relative importance.

- The previous linguistic-based metrics and their combinations have been evaluated in several international campaigns, including the ones conducted by the ACL Workshop on Machine Translation 2007 and 2008 [22, 25] and the NIST Metrics MATR workshop.

- Other usages of the IQMT metric set that we have explored include: A novel approach for parameter adjustment in Empirical Machine Translation, based on metric combinations with maximum descriptive power [30]; The study of the viability of performing heterogeneous automatic MT error analyses [26], based on the linguistic information computed by the metrics on the source, target and reference sentences.

Evaluation for the MT systems developed in the project using these metrics and tools are in progress during the third year of the project.

3 Results and Indicators

3.1 IXA (EHU)

Results, relevance and production The concrete objectives of the subproject were: (1) To improve the current open source RBMT system and the inclusion of English–Basque pair of languages; (2) To develop more sophisticated tools for the exploitation of bitexts; (3) To design and build a hybrid system combining techniques of RBMT and EBMT; (4) To apply machine learning and WSD technologies for the construction of a semantic module (in collaboration with the UPC group); and (5) To test the SMT based system for Basque (translating from Spanish and English).

The goals have been achieved, but the RBMT system for the English–Basque pair is postponed a few months. We want to highlight the relevance and originality of the work around SMT and hybrid systems, where we have achieved very satisfactory results (better than initially expected). In the next paragraph goals and production (mainly published papers, but also demos and web services) are linked:

- Improving the current open source RBMT system and inclusion of the English–Basque pair of languages. References: [3, 35]. Demo [46]
- Development of advanced tools for the exploitation of bitexts. References: [10, 11]
- Testing the SMT based system for Basque. References: [1, 28, 29, 43]. Demo [46].
- Design and construction of a hybrid system, combining RBMT, SMT and EBMT techniques. References: [6, 8, 12]. Demo [46]
- Dissemination. References: [4, 5, 7, 36][47][48]

Among the references we want to highlight the papers published in AMTA and MT-Summit conference proceedings [6, 28, 43]. Those are the main conferences in the MT area.

Technology transfer Together with the results from Elhuyar, most of the modules and technology are being transferred to Eleka (`www.eleka.net`) and to the OpenTrad consortium (`www.opentrad.org`) in the framework of two PROFIT projects.

Being Ixa group a partner of Eleka, a proposal based on this technology was presented to a public call for a major contract in es-eu MT opened by the autonomous Basque Government (Eusko Jaurlaritza) in 2008. Our proposal was evaluated as the second best choice (among four proposals). The winner was Lucy Software, an historicaly important German company, eventhough they didn offer any statistical component.

The MT technology has been encapsulated in the AnHitz project [52]. AnHitz is a project promoted by the Basque Government to develop language technologies for the Basque language. AnHitz is a collaborative project between five participants, each of them with expertise in a different area. Besides to Ixa and Elhuyar the participants are: VICOMTech

(www.vicomtech.org), an applied research center working in the area of interactive computer graphics and digital multimedia; Robotiker-Teknalia (www.robotiker.com), a technology center specialized in information and telecommunication technologies and The Aholab Signal Processing Laboratory of the University of the Basque Country (aholab.ehu.es), specialized in speech technologies.

The MT technology developed at Ixa was integrated in a CLQA system named Ihardetsi [53] which was included in the QA-CLEF evaluation 2008.

Person training Three students participate in OpenMT-Ixa. One thesis has been presented by Aingeru Mayor (2007) [35] and Gorka Labaka is finishing in 2009. Gorka Labaka won the Albaycin evaluation for MT (jth2008.ehu.es/albayzin_VL.html) integrated in the JTH2008 conference [29].

Doctors from IXA are Ph.D. advisors of participants from Elhuyar: Antton Gurrutxaga (DEA), Xabier Saralegi and Igor Leturia. Their theses are planned to be finished during 2010.

The MATMT workshop, "Mixing Approaches to Machine Translation" [5][48], organized among the 3 partners in OpenMT, was very successful. More than 50 persons came from different places of Spain, Europe and Japan. 12 papers were sent and 8 were accepted. Invited talks were given by M. Federico, P. Koehn and A. Way. They were very useful for the main aim of the corpus: discussing different methods for hybridization in MT.

Dissemination of the results has been carried out and material from the project used in the HAP official master (https://ixa.si.ehu.es/master) and in some conferences and dissemination courses (UEU, www.ueu.org)

International Collaboration We have worked with Andy Way's group at the Dublin University. Kepa Sarasola and Gorka Labaka have been in Dublin two times each, and as a result of the collaboration two papers have been generated [43, 28].

A network of excellence on MT, *Achieving Machine Translation and Multilingual Text Processing Technologies*, was designed during the MATMT workshop in San Sebastian. We are finishing its definition now, and it is being coordinated by H. Sommers from DCU university. The consortium includes besides Ixa group the DCU (Dublin), UPC (Barcelona), Edinburgh, Aachen and Charles (Praga) universities, Xerox company (Grenoble) and Bruno Kressler (Trento) Foundation.

A STREP project was proposed in 2007 in the call from the European Community. It was called EurOpenTrans (Large Scale Open Source Machine Translation for low density European Languages). The set of partners included, besides the Ixa group: DCU (Dublin), UPC (Barcelona), Edinburgh and Charles (Praga) universities, Alpinion (Ljubljana) and Translan (Dublin) companies and Elhuyar Foundation. The project was evaluated with 9.0 points when the threshold was 10.0.

Finally, see the explanation in section 3.3 regarding the TransBlog project.

Coordination. Project management A wiki (http.//ixa2.si.ehu.es/openmt) and some periodical meetings have been the main tools for the coordination and management of the project. *Coordination meetings*: (1) Donostia 19/20-XII-2006, (2) Barcelona 9-X-2007, (3) Donostia 14-II-2008, (4) Barcelona: 2-X-2008. Technical meetings: (1) Lexical selection,

Donostia 19-IV-2007; (2) Lexical selection, Barcelona 6-VII-2007, (3) MT-hybridization, Donostia 14-II-2008.

3.2 TALP(UPC)

Results, relevance and production Find below the list of goals of the subproject together with the associated production:

1. *To improve and further develop some of the current linguistic analyzers for the MT task.* As reported in Section 2.1, we have extensively achieved the objectives in this point. Related publications are [32, 33, 41, 42, 44, 45]. Regarding resources, we created the JointParser parser for joint analysis of syntactic and semantic dependencies. A web demo is freely available at [50]. We highlight the following two publications: [41] is published in the Journal of Artificial Intelligence Research (JAIR), a high quality journal in the "Computer Science; Artificial Intelligence" area. The impact factor of the journal is 1.107 (SCI/SSCI). [45] is published in the Annual Conference of the ACL, the most prestigious conference in the Computational Linguistics area.

2. *Compilation and annotation of the test corpora for Catalan.* Currently we count with a bilingual Spanish–Catalan corpus based on the bilingual 'El Periódico' newspaper. This corpus has been syntactically analyzed (POS, lemmatization and chunking).

3. *To develop a flexible, efficient, and open source decoder for SMT.* This line was abandoned because, by the time the project started, the Moses decoder became available (`http://www.statmt.org/moses/`). Moses is open source and flexible enough for our purposes.

4. *Improvement of the current phrase-based SMT systems at UPC.* As explained in section 2.2, we have improved the UPC basic SMT system in multiple directions. Related publications are: [18, 21, 20]. From those, we highlight: [21], which is published in the joint ACL/COLING conference, one of the most prestigious conferences in the Computational Linguistics area; and [20], which is published in the ACL Workshop for Statistical Machine Translation. This is a very active forum for specific Statistical MT research and is becoming the biggest workshop of the ACL conference.

5. *To apply machine learning and WSD technologies for the construction of a semantic module expressing the translation preferences.* As reported in section 2.5 we have carried out extensive research around this point, with satisfactory results. Related publications are: [24, 23, 15, 14, 13] Additional work has been performed on the integration of lexical-semantic knowledge based into SMT [16].

   From all the above publications we want to highlight: [24], a 50-page chapter in a MIT Press book on Machine Learning for SMT (NIPS series); and [23], which is published in the aforementioned ACL Workshop on SMT.

6. *To improve and test the IQMT framework.* As explained in section 2.6 the work developed in these framework has been very successful and covered all objectives. Related publications are: [9, 17, 19, 22, 25, 26, 27, 30]. Regarding resources, the complete IQMT

software is freely available [49]. As for the quality of the publications, we highlight: [9], published at the COLING-ACL conference; [27], published at the IJCNLP conference (from the Asian Federation of Natural Language Processing); and [21, 25], published at the ACL Workshop on MT. Moreover, the IQMT linguistic metrics and their combinations ranked among the best in the evaluation tasks conducted at the ACL Workshops on MT of 2007 and 2008.

Technology transfer See the explanation in section 3.1 regarding the OpenTrad consortium.

Person training Three students have participated actively in the project: Jesús Giménez, who defended his PhD thesis in July 2008 [18], and Cristina España and Xavier Lluís, who defended their Master Theses in February and September of 2008, respectively [13, 31]. Also, during 2008, Stefan Bott achieved a post-doc grant form the Catalan Government (*Beatriu de Pinós* program) for working with the UPC group in Machine Translation. Stefan Bott already joined the UPC team and he is actively working in topics related to the project.

International Collaboration See the explanation in section 3.1 regarding the MATMT-2008 workshop, the STREP project proposal and the Network of Excellence proposal. Also, we have been in contact with Xavier Carreras from the MIT CSAIL laboratory. During 2008 Xavier Carreras won a postdoc position (Ramón y Cajal) and will join the UPC group in Spring to work on Machine Translation topics.

Project management See the corresponding description in section 3.1.

3.3 Elhuyar

Results, relevance and production The participation of Elhuyar is oriented towards three main areas: (1) Basic infrastructure for Machine Translation systems: resources and tools; (2) Evaluation; and (3) Exploitation and Dissemination

In the first area, the main goal was to provide the basic resources and tools that are necessary in complex MT systems: text resources, multilingual lexical resources, tools for the construction and management of corpora, and tools for the semiautomatic extraction of lexical and terminological resources. The objectives have been achieved [2, 38, 39][51]. As basic resources we have collected several parallel corpora for several domains that have been exploited for the development and improvement of the MT different paradigms: SMT, RBMT and EBMT. We've also developed a tool for managing and storing parallel corpora. A bilingual lexicon from the administrative–text domain was created making use of the terminology extractor Elexbi. Moreover, due to the scarcity of parallel corpora we have started to carry out some works on terminology extraction from comparable corpora. We have designed a methodology which achieves promising results.

In the area of evaluation, Elhuyar's mission is twofold: (1) to elaborate a reference corpus for the semiautomatic evaluation of MT systems on the following language pairs: Spanish–Basque, English–Basque, and English–Spanish.; (2) Elhuyar accomplishes the role of a third-party evaluator of the MT technology developend in the project. This work is programmed for the third year.

As for exploitation and dissemination, Elhuyar's first goal is the integration of the MT systems developed by the working groups in a user-friendly prototype. Secondly, Elhuyar is in charge of disseminating: (1) the results of the project through a web site; (2) the collaboration in different media; (3) the organization of an international Workshop; and (4) Scientific and technical production.

The international scientific workshop was successfully organized (see MATMT-2008 in section 3.1). The integration and dissemination work is planned for the third year.

Technology transfer See the explanation in section 3.1 regarding the Anhitz project, the technology transfer to Eleka, and the application for major contract by the autonomous Basque Government.

Person training Two participants of the team (Xabier Saralegi and Igor Leturia) are currently finishing their official Masters. Another participant (Antton Gurrutxaga) obtained the PhD research diploma (DEA) an is currently finishing his research for the PhD thesis.

International Collaboration See the explanation in section 3.1 regarding the MATMT-2008 workshop and the EurOpenTrans STREP project proposal.

Elhuyar plans to submit a project proposal (TransBlog) in the European Work Programme 2009-2010, Challenge 2 call. We have already set up the consortium, and at this moment we are working in the proposal which will include the application of technologies developed in OpenMT in a framework for translating and increasing visibility of Blogs.

Project management See the corresponding description in section 3.1.

References

[1] E. Agirre, A. Díaz de Ilarraza, G. Labaka G. and K. Sarasola. Uso de información morfológica en el alineamiento Español-Euskara. In *Proceedings of the 12th Annual Conference of the SEPLN (Sociedad Española para el Procesamiento del Lenguaje Natural)*, 2006.

[2] I. Alegria, A. Gurrutxaga, X. Saralegi, S. Ugartetxea. ELeXBI, A Basic Tool for Bilingual Term Extraction from Spanish-Basque Parallel Corpora. In *Proceedings of the 12th EURALEX International Conference*. Torino. 2006.

[3] I. Alegria, A. Díaz de Ilarraza, G. Labaka, M. Lersundi, A. Mayor and K. Sarasola. Transfer-based MT from Spanish into Basque: reusability, standardization and open source. In *Proceedings of the CicLing-2007 Conference*, LNAI Series 4394, 374–384, 2007.

[4] I. Alegria, X. Arregi, X. Artola, A. Díaz de Ilarraza, G. Labaka, M. Lersundi, A. Mayor and K. Sarasola. Strategies for sustainable MT for Basque: incremental design, reusability, standardization and open-source. In *Proc. of the IJCNLP-08 Workshop on NLP for Less Privileged Languages*, 59–64. Hyderabad, 2007.

[5] I. Alegria, L. Màrquez and K. Sarasola (editors). *Mixing Approaches to Machine Translation*, Proceedings of the MATMT-2008 Workshop. Euskal Herriko Unibertsitatea (ISBN 978-612-2224-7), 2008.

[6] I. Alegria, A. Casillas, A. Diaz de Ilarraza, J. Igartua, G. Labaka, M. Lersundi, A. Mayor and K. Sarasola. Spanish-to-Basque MultiEngine Machine Translation for a Restricted Domain. In *Proceedings of the 8th Conference of the Association for Machine Translation in the Americas (AMTA-2008)*. Hawaii, USA, 2008.

[7] I. Alegria, A. Díaz de Ilarraza, G. Labaka, M. Lersundi and K. Sarasola. Itzulpen automatikoa: aukerak, arazoak eta erronkak. *Bat Soziolinguistika*, 66, 107–122, 2008.

[8] I. Alegria, A. Díaz de Ilarraza, J. Igartua, G. Labaka, B. Laskurain, M. Lersundi, A. Mayor, K. Sarasola, A. Casillas and X. Saralegi. Mixing Approaches to MT for Basque: Selecting the best output from RBMT, EBMT and SMT. In *Proceedings of the MATMT2008 Workshop: Mixing Approaches to Machine Translation*. Donostia-San Sebastian, 2008.

[9] E. Amigó, J. Giménez, J. Gonzalo and L. Màrquez. MT Evaluation: Human–like vs. Human Acceptable. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics, COLING-ACL 2006*, Sydney, Australia, 2006.

[10] A. Casillas, A. Díaz de Ilarraza, J. Igartua, R. Martinez and K. Sarasola. Compilation and Structuring of a Spanish-Basque Parallel Corpus. In *Proc. of the LREC-2006*, 2006.

[11] A. Díaz de Ilarraza, J. Igartua, K. Sarasola, A. Sologaistoa, A. Casillas and R. Martinez. Spanish-Basque Parallel Corpus Structure: Linguistic Annotations and Translation Units. In *Proceedings of TSD 2007 Conference*, Plzen, Czech Republic, 2007.

[12] A. Díaz de Ilarraza, G. Labaka and K. Sarasola. Statistical Post-Editing: A Valuable Method in Domain Adaptation of RBMT Systems. In *Proceedings of the MATMT2008 Workshop: Mixing Approaches to Machine Translation*, Donostia–San Sebastian, 2008.

[13] C. España–Bonet. *A proposal for an Arabic-to-English SMT*. Master Thesis, Universitat Politècnica de Catalunya (Artificial Intelligence Program), Barcelona, 2008.

[14] C. España–Bonet, G. Giménez and L. Màrquez. Discriminative Phrase-Based Models for Arabic Machine Translation. Submitted to the *International Journal of Computer Processing of Languages*, January 2009.

[15] C. España–Bonet, J. Giménez and L. Màrquez. The UPC-LSI Discriminative Phrase Selection System: NIST MT Evaluation 2008. In *Proceedings of the 2008 NIST Open Machine Translation Evaluation Workshop, MT08*, Washington, USA, 2008.

[16] M. García, J. Giménez and L. Màrquez. Enriching Statistical Translation Models using a Domain-independent Multilingual Lexical Knowledge Base. To appear in *Proceedings of the CicLing-2009 Conference*, Mexico City, Mexico, 2009.

[17] J. Giménez. *IQMT: A Framework for Automatic Machine Translation Evaluation based on Human Likeness*. Technical Report LSI-07-29-R, LSI, Universitat Politècnica de Catalunya, 2007.

[18] J. Giménez. *Empirical Machine Translation and its Evaluation*. Ph.D. Thesis, Universitat Politécnica de Catalunya, Barcelona, July, 2008.

[19] J. Giménez and E. Amigó. IQMT: A Framework for Automatic Machine Translation Evaluation. In *Proc. of the LREC-2006*. Genoa, 2006.

[20] J. Giménez and L. Màrquez. Low–cost Enrichment of Spanish WordNet with Automatically Translated Glosses: Combining General and Specialized Models. In *Proc. of the 21st International Conference COLING-ACL 2006*, Sydney, 2006.

[21] J. Giménez and L. Màrquez. The LDV–COMBO system for SMT. In *Proceedings. of the First Workshop on Statistical Machine Translation, WMT 2006*, New York, 2006.

[22] J. Giménez and L. Màrquez. Linguistic Features for Automatic Evaluation of Heterogeneous MT Systems. In *Proceedings of the Second Workshop on Statistical Machine Translation, WMT 2007*, Prague, Czech Republic, 2007.

[23] J. Giménez and L. Màrquez. Context–aware Discriminative Phrase Selection for Statistical Machine Translation. In *Proceedings of the Second Workshop on Statistical Machine Translation, WMT 2007*, Prague, Czech Republic, 2007.

[24] J. Giménez and L. Màrquez. Discriminative Phrase Selection for Statistical Machine Translation. In C. Goutte, N. Cancedda, M. Dymetman and G. Foster (eds.) *Learning Machine Translation*. NIPS Workshop Series, MIT Press, 2008.

[25] J. Giménez and L. Màrquez. A Smorgasbord of Features for Automatic MT Evaluation. In *Proceedings of the Third Workshop on Statistical Machine Translation, WMT 2008*, Columbus, Ohio, USA, 2008

[26] J. Giménez and L. Màrquez. Towards Heterogeneous Automatic MT Error Analysis. In *Proc. of the 6th International Conference on Language Resources and Evaluation, LREC 2008*, Marrakech, Morocco, 2008.

[27] J. Giménez and L. Màrquez. Heterogeneous Automatic MT Evaluation Through Non-Parametric Metric Combinations. In *Proceedings of the Third International Joint Conference on Natural Language Processing, IJCNLP 2008*, Hyderabad, India, 2008.

[28] G. Labaka, N. Stroppa, A. Way and K. Sarasola. Comparing Rule-Based and Data-Driven Approaches to Spanish-to-Basque Machine Translation. In *Proceedings of MT-Summit XI*, Copenhagen, Denmark, 2007.

[29] G. Labaka, A. Díaz de Ilarraza and K. Sarasola. Descripcin de los sistemas presentados por IXA-EHU a la evaluación ALBAYCIN. In *Proceedings of V Jornadas en Tecnología del Habla*, Red Temática de Tecnologías del Habla, 93–95. Bilbao, 2008.

[30] P. Lambert, J. Giménez, M. R. Costa–jussà, E. Amigó, R. E. Banchs, L. Màrquez and J. A. R. Fonollosa. MT System Development based on Human Likeness. In *Proc. of the IEEE/ACL 2006 Workshop on Spoken Language Technology*, Palm Beach, 2006.

[31] X. Lluís. *Joint Learning of Syntactic and Semantic Dependencies*. Master Thesis, Universitat Politècnica de Catalunya (Artificial Intelligence Program), Barcelona, 2008.

[32] X. Lluís and L. Màrquez. A Joint Model for Parsing Syntactic and Semantic Dependencies. In *Proc. of the 12th Conference on Computational Natural Language Learning (CoNLL-2008)*, Manchester, UK, 2008.

[33] L. Màrquez, L. Padró, M. Surdeanu, and L. Villarejo. UPC: Experiments with Joint Learning within SemEval Task 9. In *Proc. of the 4th SemEval*, 426–429, ACL-07, Prague, 2007.

[34] L. Màrquez, L. Villarejo, M. A. Martí and M. Taulé. SemEval-2007 Task 09: Multilevel Semantic Annotation of Catalan and Spanish. In *Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007)*, 42–47, ACL-07, Prague, Czech Republic, 2007.

[35] A. Mayor. *Matxin: Erregeletan oinarritutako itzulpen automatikoko sistema baten eraikuntza estaldura handiko baliabide linguistikoak berrerabiliz*. Ph. D. Dissertation. Euskal Herriko Unibertsitateko Donostiako Informatika Fakultatea, 2007.

[36] A. Mayor. Matxin: Erregeletan oinarritutako IA sistema baten eraikuntza estaldura handiko baliabide linguistikoak berrerabiliz. *Uztaro*, 66, 105–106, 2008.

[37] X. Saralegi, I. Alegria. Similitud entre documentos multilingues de carácter técnico en un entorno Web. In *Proc. of the SEPLN 2007*. Sevilla, 2007.

[38] X. Saralegi, I. San Vicente, A. Gurrutxaga. A utomatic Extraction of Bilingual Terms from Comparable Corpora in a Popular Science Domain. In *Proc. of the 6th LREC - Building and using Comparable Corpora workshop*. Marrakech, 2008.

[39] X. Saralegi, I. san Vicente, M. Lopez de Lacalle. Mining Term Translations from Domain Restricted Comparable Corpora. In *Proc. of the SEPLN 2008*. Madrid, 2008.

[40] M. Surdeanu, R. Johansson, A. Meyers, L. Màrquez and J. Nivre. The CoNLL-2008 Shared Task on Joint Parsing of Syntactic and Semantic Dependencies. In *Proc. of the 12th CoNLL*, Manchester, UK, 2008.

[41] M. Surdeanu, L. Màrquez, X. Carreras and P. R. Comas. Combination Strategies for Semantic Role Labeling. *Journal of Artificial Intelligence Research*, 29, 105–151, 2007.

[42] M. Surdeanu, R. Morante and L. Màrquez. Analysis of Joint Inference Strategies for the Semantic Role Labeling of Spanish and Catalan, In *Proc. of the 9th CICLing*, LNCS 4919, 206–218, Haifa, 2008.

[43] N. Stroppa, D. Groves, A. Way and K. Sarasola. Example-Based Machine Translation of the Basque Language. In *Prceedings of the 7th conference of the Association for Machine Translation in the Americas (AMTA)*, Boston, MA, 2006.

[44] B. Zapirain, E. Agirre, L. Màrquez. Sequential SRL Using Selectional Preferences. An aproach with Maximum Entropy Markov Models. In *Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007)*, 354–357, ACL-07, Prague, 2007.

[45] B. Zapirain, E. Agirre and L. Màrquez. Robustness and Generalization of Role Sets: PropBank vs. VerbNet. In *Proceedings of the 46th Annual Meeting of the Association of Computational Linguistics (ACL-08)*, 550–558, Columbus, Ohio, USA, 2008.

Links to Resources and Demos

[46] MT systems, en-eu and hybrids: `http://ixa2.si.ehu.es/openmt-demo`

[47] Website of the OpenMT project: `http://ixa.si.ehu.es/openmt`

[48] Website of the MATMT-2008 Workshop: `http://ixa2.si.ehu.es/matmt-2008/`

[49] Website of the IQMT Framework for MT Evaluation: `http://www.lsi.upc.edu/~nlp/IQMT/`

[50] Website of JointParser: `http://www.lsi.upc.edu/~xlluis/jointparser/`

[51] Website repository of TMs: `http://ixa.si.ehu.es/openmt/demoak_html`

[52] Demo AnHitz: `http://teknopolis.elhuyar.org/ikusi.asp`

[53] Demo Ihardetsi: `http://sisx04.si.ehu.es/IhardetsiWebDemo/IhardetsiBezeroa.jsp`