Procesamiento del Lenguaje Natural, Revista nº 55, septiembre de 2015, pp 169-172 recibido 24-03-2015 revisado 27-04-2015 aceptado 15-05-2015

Lexical semantics, Basque and Spanish in QTLeap: Quality Translation by Deep Language Engineering Approaches

QTLeap - Traducción de calidad mediante tratamientos profundos de ingeniería lingüística

Eneko Agirre, Iñaki Alegria, Nora Aranberri, Mikel Artetxe, Ander Barrena, António Branco (**), Arantza Díaz de Ilarraza, Koldo Gojenola, Gorka Labaka, Arantxa Otegi, and Kepa Sarasola
Ixa Taldea. Universidad del Pais Vasco /Euskal Herriko Unibertsitatea (UPV/EHU)
Manuel Lardizabal 1, -20018 Donostia
(**) Universidade de Lisboa, Departamento de Informática,Faculdade de Ciências
e.agrirre@ehu.eus

Resumen: El objetivo de este proyecto europeo FP7 es contribuir a la mejora en la calidad de la traducción automática mediante el uso de semántica, análisis sintático profundo y el uso de datos abiertos entrelazados.

Palabras clave: Traducción automática, Análisi profundo, Semántica, LOD

Abstract: The goal of this FP7 European project is to contribute for the advancement of quality machine translation by pursuing an approach that further relies on semantics, deep parsing and linked open data.

Keywords: Machine Translation, Deep language Engineering, Semantics, LOD ...

1 *Summary*

QTLeap project (Quality Translation by Deep Language Engineering Approaches) is a collaborative project funded by the European Commission (FP7-ICT-2013.4.1-610516) that aims to produce high-quality outbound Machine Translation (MT) using deep language engineering approaches to achieve higher quality translations (Branco and Osenova, 2014). The approach is based on deep processing and a transfer based architecture able to create hybrid systems. IXA Taldea is the partner responsible of the developments for semantics, Basque and Spanish in this project that is run by an European consortium with other seven partners: Bulgarian Academy of Sciences, Charles University in Prague, German Research Center for Artificial Intelligence, Higher Functions Lda., Humboldt University in Berlin, University of the Basque Country, University of Groningen and University of Lisbon. The project started in November 1st, 2013, and has a duration of 36 months.

The incremental advancement of research on Machine Translation has been obtained by encompassing increasingly sophisticated statistical approaches and fine grained linguistic features that add to the surface level alignment on which these approaches are ultimately anchored. The goal of this project is to contribute for the advancement of quality MT by pursuing an approach that further relies on semantics and opens the way to higher quality translation. We build on the complementarity of the two pillars of language technology, symbolic and probabilistic, and seek to advance their hybridization. We explore combinations of them that amplify their strengths and mitigate their drawbacks, along the development of three MT pilot systems that progressively seek to integrate deep language engineering approaches.

The construction of deep treebanks has progressed to be delivering now the first significant Parallel DeepBanks, where pairs of synonymous sentences from different languages are annotated with their fully-fledged grammatical representations, up to the level of their semantic representation.

The construction of Linked Open Data and other semantic resources, in turn, has progressed now to support impactful application of lexical semantic processing that handles and resolves referential and conceptual ambiguity.

ISSN 1135-5948

These cutting edge advances permit for the cross-lingual alignment supporting translation to be established at the level of deeper semantic representation. The deeper the level the less language-specific differences remain among source and target sentences and new chances of success become available for the statistically based transduction.

English is the common language for the MT systems to be built in the project, being as target or source for each one of the other 7 languages in the project: Basque, Bulgarian, Czech, Dutch, German, Portuguese and Spanish.

2 Architecture:TectoMT

All the partners use a MT transfer-based architecture, being TectoMT the architecture used for almost all the language pairs (Czech, Spanish, Portuguese, Basque, and Bulgarian). TectoMT is a highly modular, structural MT system implemented within the Treex NLP framework that allows for fast and efficient development of machine translation by exploiting a wide range of software modules already integrated in TectoMT, such as tools for sentence segmentation, tokenization, morphological analysis, POS tagging, shallow and deep syn- tax parsing, named entity recognition, anaphora resolution, tree-to-tree translation, natural language generation, word-level alignment of parallel corpora, and transfer based on deep syntactic (tectogrammatical) layer (Zeman et al., 2014). The tectogrammatical layer is based on the Prague Dependency Treebank. Figure 1 shows the architecture of the TectoMT system.

3 Deep Language Analysis for Spanish and Basque

Deep Language Analysis for Spanish and Basque will be implemented in our group via Ixa-pipes (Agerri, Bermudez, and Rigau, 2014). IXA-pipes is a modular set of Natural Language Processing tools (or pipes) which provide easy access to NLP technology for several languages. It offers robust and efficient linguistic annotation to both researchers and non-NLP experts with the aim of lowering the barriers of using NLP technology either for research purposes or for small industrial developers and SMEs. The ixa-pipes tools can be used or exploit its modularity to pick and change different components (Morpholgy, POS tagger, Chunker, Coreference, Named Entities Recognizer...). The tools are developed by the IXA NLP Group of the University of the Basque Country.

4 Semantic ways for MT improving

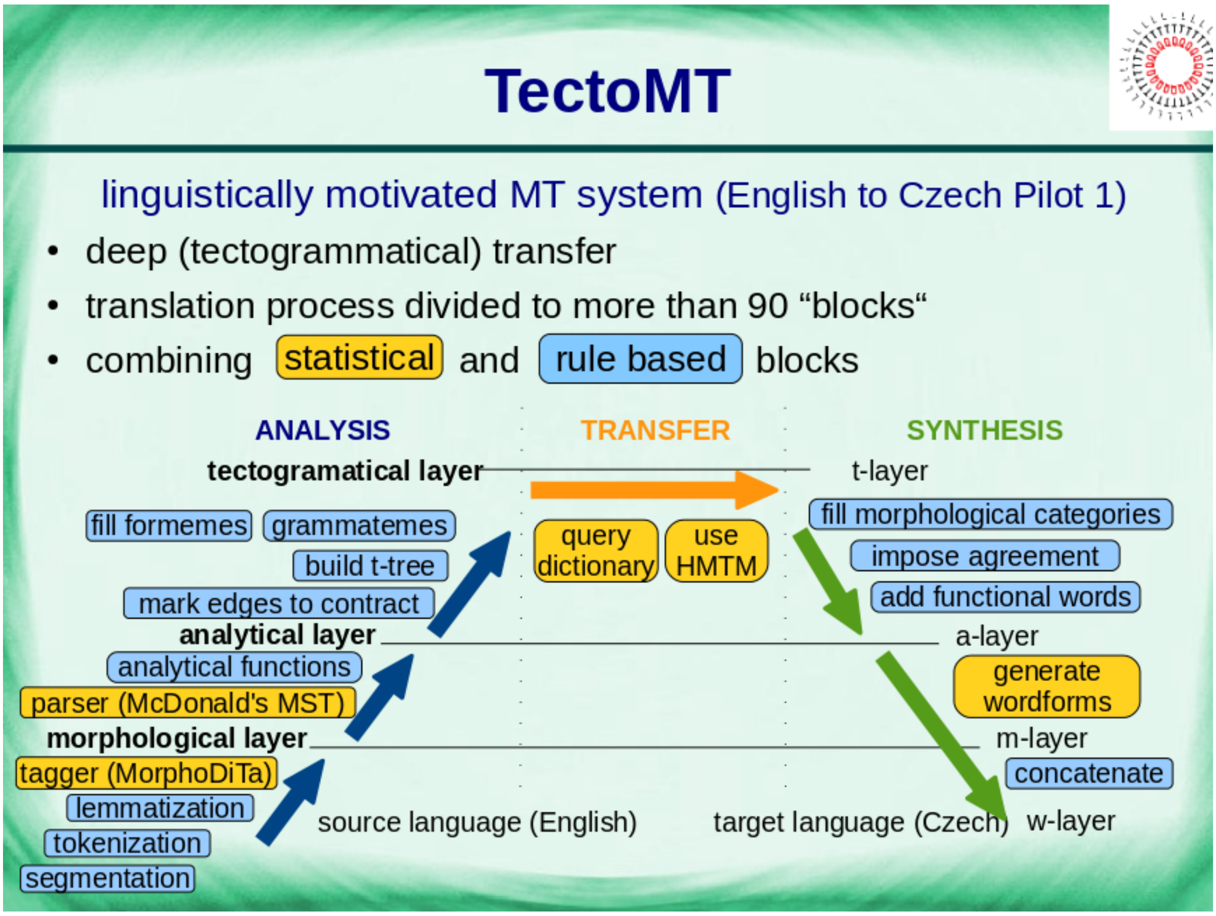
The overall goal of the work package on lexical semantics lead by Ixa Taldea is to enhance MT with advanced crosslingual methods for the resolution of referential and lexical ambiguity by pursuing the following objectives:

- to provide for the assembling and curation of the data sets and processing tools available to support the resolution of referential and lexical ambiguity,
- to leverage the resolution of referential and lexical ambiguity by means of advanced crosslingual named entity and word sense resolution methods,
- to proceed with the intrinsic evaluation of the solutions found in the previous task,
- to contribute for high quality machine translation by using semantic linking and resolving.

5 Real user scenario

QTLeap project successfully achieved the first year milestone last November related with the identification of a real user scenario that will show the suitability to use machine translation ('Leveraging practical multilingual helpdesk services with machine translation'). Namely in an IT helpdesk service provided by HF, Higher Functions - Intelligent Information Systems Ltd, a Portuguese SME, which is a partner of the consortium.

With this service, if a user of an IT device or service needs to solve a problem, he/she can ask a question for help through a chat channel. If there is already a similar question in the database, the associated response is immediately delivered to the user. This process helps to minimize the human operation, which becomes needed only in those cases when there is no similar question-answer pair already available in the database. The application of the machine translation system extends this support service by allowing the use of the seven languages in the project to ask a question to the helpdesk service. The



TectoMT

linguistically motivated MT system (English to Czech Pilot 1)

- deep (tectogrammatical) transfer
- translation process divided to more than 90 “blocks“
- combining statistical and rule based blocks

ANALYSIS

TRANSFER

SYNTHESIS

tectogramatical layer

t-layer

fill formemes

grammatemes

build t-tree

mark edges to contract

query dictionary

use HMTM

fill morphological categories

impose agreement

add functional words

analytical layer

a-layer

analytical functions

parser (McDonald's MST)

generate wordforms

morphological layer

m-layer

tagger (MorphoDiTa)

concatenate

lemmatization

tokenization

segmentation

source language (English)

target language (Czech)

w-layer

Figure 1: TectoMT architecture (Popel, 2014)

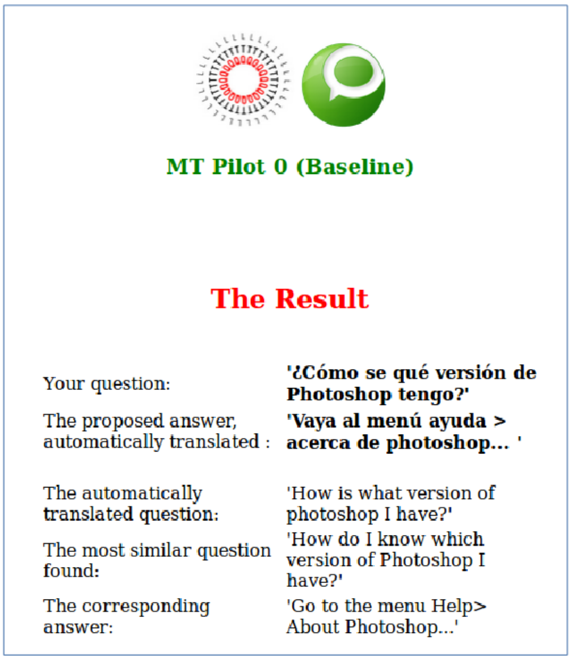
users can ask a question in their own languages, which is translated into the language of the questions and answers that are stored in the database. With the retrieved answer also translated back to the language of the original question, the helpdesk thus returns the response in the language of the user. Examples of the questions the user could ask are the following:

- Tengo un iPad. ¿Cómo abro una nueva pestana en Safari?
- ¿Cómo se qué versión de Photoshop tengo?
- En YouTube, ¿cómo puedo buscar vídeos en HD?
- ¿Qué es una IP estática?
- He instalado Linux. ¿Cómo puedo listar los archivos en una carpeta?

- Ich habe ein iPad. Wie kann ich in Safari einen neuen Tab öffnen?
- Wie kann ich auf YouTube nach Videos in HD suchen?
- Wie kann ich erfahren, welche Photoshop Version ich habe?
- Was ist ein statisches IP?
- Ich habe Linux installiert. Wie kann ich die Liste der Dateien in einem Ordner erstellen?

- iPad bat dut. Nola ireki dezaket Safari fitxa berri batean?
- Nola jakin dezaket Photoshop-en zein bertsio dudan?
- YouTube-n, nola bila ditzaket HD bideoak?
- Zer da IP estatiko bat?
- Linux instalatu dut. Nola zerrenda ditzaket fitxategiak karpeta batean?

Figure 2 shows an example of the work done by the initial translators for Spanish. Although the answer given to the user is not being to be perfect, it uses to be suitable enough to resolve the question. The evaluation of this helpdesk service expanded with



MT Pilot 0 (Baseline)

The Result

Your question:	**'¿Cómo se qué versión de Photoshop tengo?'**
The proposed answer, automatically translated :	**'Vaya al menú ayuda > acerca de photoshop... '**
The automatically translated question:	'How is what version of photoshop I have?'
The most similar question found:	'How do I know which version of Photoshop I have?'
The corresponding answer:	'Go to the menu Help> About Photoshop...'

Figure 2: Example of use of PCMEDIC, the real user scenario

the machine translation functionality showed that even with a low-quality machine translation system, it was already possible to achieve a very significant reduction of human operation of about 60% on average for each new language to be covered by the service. This technical advance leverages a great deal of advantages for this kind of business with regard to its extension to the single digital market, as well for its improvement in terms of productivity and resources optimization, with the consequent effective reduction of costs.

There is a common path of progression for each pair X<->EN, ensuring comparability of the research exercise: every pair is developed along pilots 0 to 3 reinforced by complementary strengths and backgrounds of the different partners with their systems, resources and technology.

6 Advisory Board

The direction of the project is informed by the advice on strategic issues from the Advisory Board of Potential Users. This Advisory Board includes industrial participants that are ready to contribute with their advice on the strategic course of the project activities, and are interested in the innovation potential of the results targeted at by the project and will be in the first row of the potential users that will take the lead to exploit their business potential. The members of this board are:

- CA Technologies Development Spain S.A (Spain)
- Eleka Ingeniaritza Linguistikoa SL (Basque Country, Spain)
- OMQ GmBH (Germany)
- Lingea s.r.o. (Czech Republic)
- Seznam.cz, a.s. (Czech Republic)
- Higher Functions, Lda (Portugal, also partner)

Information on QTLeap project and contact details:
Website: http://qtleap.eu/
Facebook: https://www.facebook.com/qtleap
Twitter: https://twitter.com/QTLeap

7 Aknowledgements

The research leading to these results was carried out as part of the QtLeap project funded by European Community (FP7-ICT-2013.4.1-610516)

References

Agerri, R., J. Bermudez, and G. Rigau. 2014. IXA pipeline: Efficient and Ready to Use Multilingual NLP tools. 9th Language Resources and Evaluation Conference (LREC2014), Reykjavik, Iceland. pages 26–36.

Branco, A. and P. Osenova. 2014. QTLeap - Quality Translation with Deep Language Engineering Approaches. Poster at EAMT2014, Dubrovnik.

Popel, M. 2014. MT Pilot 1: Entry-level Deep MT. Internal presentation in QTLeap project Meeting. Lisbon.

Zeman, D., O. Dušek, D. Mareček, M. Popel, L. Ramasamy, J. Štěpánek, Z. Žabokrtský, and J. Hajič. 2014. Hamledt: Harmonized multi-language dependency treebank. *Language Resources and Evaluation*, 48(4):601–637.