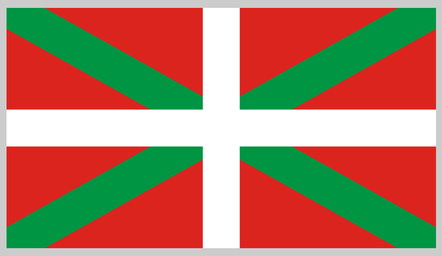
SMT and Hybrid systems of the QTLeap project in the WMT16 IT-task

Rosa Del Gaudio[1], Gorka Labaka[2], Eneko Agirre[2], Petya Osenova[3], Kiril Simov[3], Martin Popel[4], Dieke Oele[5], Gertjan van Noord[5], Luís Gomes[6], João Rodrigues[6], Steven Neale[6], João Silva[6], Andreia Querido[6], Nuno Rendeiro[6], António Branco[6]

rosa.gaudio@pcmedic.pt, {gorka.labaka, e.agirre}@ehu.eus,{petya, kivs}@bultreebank.org, popel@ufal.mff.cuni.cz, {d.oele, g.j.van.noord}@rug.nl,luisgomes@gmail.com, {joao.rodrigues, steven.neale, jsilva, andreia.querido, nuno.rendeiro, antonio.branco}@di.fc.ul.pt

[1]Higher Functions Sistemas Inteligentes, Lisbon, Portugal - [2]University of the Basque Country, UPV/EHU, San Sebastian, Spain - [3]Linguistic Modelling Department, IICT-BAS, Sofia, Bulgaria - [4]Charles University in Prague, Faculty of Mathematics and Physics,UFAL, Czechia - [5]Rijksuniversiteit Groningen, Groningen, The Netherlands - [6]Universidade de Lisboa, Departamento de Informática, Faculdade de Ciências

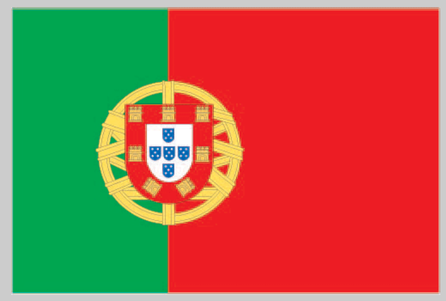


Introducion: the QTLeap Project

The QTLeap (http://www.qtleap.eu) project focuses on the development of an articulated methodology for machine translation that explores **deep language engineering approaches and sophisticated semantic datasets**.

In this paper, we present the systems developed by the University of Basque Country for **Basque** and **Spanish**, Charles University in Prague for **Czech**, by University of Groningen for **Dutch**, by University of Lisbon for **Portuguese** and by IICT-BAS of the Bulgarian Academy of Sciences for **Bulgarian**.

For each language two different systems were submitted: a phrase-based MT system built using Moses, and a system exploiting deep language engineering approaches, that in all the languages but Bulgarian was implemented using TectoMT.



Baseline : Moses

All the systems based on Moses have been trained on a phrase-based model by Giza++ or mGiza with "growdiag-final-and" symmetrization and "msd-bidirectional-fe" reordering. For the language pairs where big quantities of domain-specific monolingual data were available separate language models (domain-specific and generic) were interpolated against our ICT domain-specific development set.

QTLeap: TectoMT

The deep translation is based on the TectoMT system, an open-source MT system based on the Treex platform for general natural-language processing. TectoMT uses a combination of rule-based and statistical modules, with a statistical transfer based on HMTM at the level of a deep, so-called tectogrammatical, representation of sentence structure. The general TectoMT pipeline is language independent, and consists of analysis, deep transfer, and synthesis steps. .

Results

QTLeap system significantly better than the baseline Moses for 5 out of 6 languages

Systems	Basque		Bulgarian		Czech		Dutch		Spanish		Portuguese	
	BLEU	TrueSkill	BLEU	TrueSkill	BLEU	TrueSkill	BLEU	TrueSkill	BLEU	TrueSkill	BLEU	TrueSkill
Moses	8.3	-1.570	**16.6**	**5.262**	20.8	-0.616	**21.9**	-2.462	16.0	-1.926	13.7	-2.276
TectoMT	**10.3**	**+1.570**	-	-	**21.5**	**0.130**	19.0	**0.154**	**24.2**	**-0.809**	**15.2**	**-1.063**
DeepMoses	-	-	15.3	-5.262	-	-	-	-	-	-	-	-

Discussion

For **Dutch**, the Moses system outperforms theTectoMT only when considering the BLUE score. Better results, in terms of BLEU-score, were obtained in the opposite translation direction which indicates that more effort should be put into this translation direction.

Regarding **Bulgarian**, the current drop might be overcome by improving the WordNet information for Bulgarian, its mapping to the English WordNets as well as the processing pipelines. Also, the train of this system should use more more data and exploit other bilingual dictionaries.

.

The development is currently supported by the QTLeap 7th FP project No. 610516 (http://qtleap.eu).