FACT: The US and the European Union represent the major sources of Foreign Direct Investment for Costa Rica.

CHALLENGE: To make information accessible to foreign investors (currently pursuing only English as the target language).

PROPOSAL: A in-domain, corpus-based machine translation and post-editing could represent a feasible and cost-effective solution to cope with the lack of information in English, following a GILT production model.

DATA COLLECTION
Corpora retrieved from the Opus Corpus
WC: 53,246,174
[DOT, ECB, News]

ENGINE TRAINING
KantanMT and MMT served as the hosting platforms to carry out the training phase. Only KantanMT succeeded through the final stage of the case.

QUALITY ASSESSMENT
> BLEU: 49%, TER: 45%, F-Measure: 68%
> Human-based Evaluation

MEASURING
BLEU: 57%, TER: 43%, F-Measure: 70%
The corpora from the Opus Corpus works as a starting point, but further customizations to the engine are necessary to produce a more proficient engine.

REFINEMENT
Crawling methods to gather more data.
WC: 54,406,825
MC: 49,672
[WIPO, CR Electoral Code, CR Constitution, glossary, run-time glossary and ignore words list]

COSTA RICA'S E-GOVERNMENT MT ENGINE

Final results:
✓ Proficient in vocabulary related to the public sector.
✓ Corroborated improvement.
✓ Automated metrics do not correlate with human quality expectations.

Recommendations for further studies:
→ To run a DQF productivity test to assess the reduction in translation cost.
→ To improve the amount of training data for the LM and TM.
→ To apply rule-based approaches and post-editing corrections.

Researcher and translator: Gloriana Coozza Garro. For more information, visit http://gcocozza.com