Tapadóir

The Department of Arts, Heritage and the Gaeltacht, Govt. of Ireland
MT development for translation workflow

List of partners
CNGL/ADAPT, Ireland
National Centre for Language Technology, Ireland
Dublin City University, Ireland
Department of Arts, Heritage and the Gaeltacht, Government of Ireland

Project duration: July 2014 — January 2016

Summary

Tapadóir (from the Irish "tapa" – fast) is a statistical machine translation project which has just completed its pilot phase. The heart of the project is the development of an English–Irish translation system, intended for integration into the workflow of a professional translator at an Irish government department. In practice, this means statistical machine translation from a highly-resourced majority language (English) to an under-resourced minority language (Irish) with significant linguistic differences. A secondary aim is the production of English–Irish parallel corpora suitable for future translation tool and NLP developers.

There is high demand for Irish-language translated texts within Irish government departments, and this MT integration aims to increase the speed of translation to meet this demand. Tapadóir currently out-performs (based on BLEU score) Google Translate on data from our use case domain (official government documents and reports). The official European Commission machine translation service, MT@EC, rate their English-to-Irish MT system as suitable for gist translation, but below useful editable quality, the standard required by the client. While MT@EC also build custom pilot projects based on existing user data, the client's data is limited. Therefore, further data collection constitutes a large proportion this project's remit.

English-to-Irish translation holds a number of challenges. From an NLP perspective, Irish is very much under-resourced, and much of the project so far has focused on corpus development. The target language is also morphologically much richer than the source (e.g. initial mutations, synthetic verb forms, case), and the resulting data sparsity further compounds the these translation challenges. Linguistically, the language pair word order is divergent (Subject-Verb-Object vs. Verb-Subject-Object), with other word order differences at lower levels, such as adjectives following nouns, and the genitive noun following its possessed object in Irish.

To cope with this, we are currently developing source-side reordering rules to address word-order divergence, and we are exploring ways to overcome the morphological discrepancies. Our aim is to use various methods to provide useful machine translation output for an unusual and challenging language pair. Rather than aiming to investigate the general effectiveness of particular methods, we are attempting to find the best practical combination for this resource-poor and linguistically challenging use-case. We expect that our work will be of use to developers of MT systems for other under-resourced languages.

The Tapadóir MT engine will be deployed for in-house use by the Irish Department of Arts, Heritage and the Gaeltacht. However, we hope to make freely available the resources gathered/created over the course of its development, for the sake of future Irish-language projects.