Tapadóir: Developing a Statistical Machine Translation Engine and Associated Resources for Irish

Meghan Dowling[1], Lauren Cassidy[2], Eimear Maguire[3], Teresa Lynn[1], Ankit Srivastava[1], John Judge[1]

[1]ADAPT CENTRE, Dublin City University, Ireland
[2] Trinity College Dublin, Ireland [3]Université Paris Diderot
[1]meghan.dowling@dcu.ie,{asrivastava,tlynn,jjudge}@computing.dcu.ie
[2]cassidl2@tcd.ie [3]eimear.maguire@etu.univ-paris-diderot.fr

Abstract

Tapadóir (from the Irish *tapa* 'fast' and the nominal suffix *-óir*) is a statistical machine translation (SMT) project, funded by the Irish government. This work was commissioned to help government translators meet the translation demands which have arisen from the Irish language's status as an official EU and national language. The development of this system, which translates English into Irish (a morphologically rich, low-resourced minority language), has produced an interesting set of challenges. These challenges have inspired a creative response to the lack of data and NLP tools available for the Irish language and have also resulted in the development of new resources for the Irish linguistic and NLP community. We show that our SMT system out-performs Google Translate™ (a widely used general-domain SMT system) as a result of steps we have taken to tailor translation output to the user's specific needs.

1 Introduction

The Irish language is both the first official language of the Republic of Ireland and an official EU language. Despite this, as outlined by a recent META-NET study of the status of language technology resources for European languages (Judge et al., 2012), there is a significant lack of NLP resources available for Irish.

The lack of resources for Irish is particularly evident in the field of machine translation. While there is current research underway on a rule-based MT system[1], until now there has been limited data available for the development of a statistical machine translation (SMT) system. Such data-driven systems rely on parallel (bilingual) data upon which a translation system can learn and predict translations of previously unseen text.

Recently. at a national level, significant momentum has built up in the revitalisation of the Irish language. This arises from the recognition of Irish as an official EU language and from the implementation of the Official Languages Act[2] for the Irish language. Under this Act, all official documents and public services should be accessible in either English or Irish. As a result, there has been increased pressure on Irish government institutions to improve the provision of Irish language resources. The Tapadóir SMT project has thus grown both from the government's need to provide timely and accurate translations of such documents and the growth of the Irish NLP community.

At a European level, Irish is a target language for EU Parliament proceedings and documents. However, the volume of English–Irish parallel data available is significantly lower than other EU language pairs. For example, DCEP (Digital Corpus of the European Parliament)[3] contains only 14 documents[4] for Irish as opposed to >100,000 documents in English and other languages. Similarly, DGT-TM (Directorate General for Translation, Translation Memories)[5] contains 52,000 Irish translation units as opposed to over 6 million translation units in English, for example. These differences are due in part to a derogation[6] in place for Irish whereby only key EU legislation must be translated into Irish, i.e. only documents covered by co-decision between the European Parliament and the Council of Ministers.

This paper describes this work and resources and is divided as follows. In Section 2, we discuss the motivation for the development of this SMT system. Section 3 provides an overview of the Irish language and the various linguistic challenges that it presents for machine translation. Section 4 discusses the data requirements for a domain-specific SMT system, describes the development of our baseline resources and system, and presents our baseline results. It also discusses the improvements in scores following various approaches we took to add new training data and tweak the system settings. Finally, in Section 5, we discuss our ongoing work and future plans for enhancing the translation system.

2 Motivation

For the continued survival or revitalisation of a minority language such as Irish, it is crucial that it remains in use and therefore up to date with linguistic resources and technology. As Irish is the official and national language

[1]As yet, there is no publication available. Information received though personal communication with Elaine Uí Dhonnchadha, School of Linguistic, Speech and Communication Sciences, Trinity College Dublin.

[2]http://www.oireachtas.ie/documents/bills28/acts/2003/a3203.pdf

[3]https://ec.europa.eu/jrc/en/language-technologies/dcep

[4]In fact, only 3 of these proved to be suitable for training data. The remaining 11 were revision documents with minor updates

[5]https://ec.europa.eu/jrc/en/language-technologies/dgt-translation-memory

[6]Due to expire 2021.

of Ireland, the Irish government has a responsibility to provide all official documents in both English and Irish. More specifically, within government departments such as the Department of Arts, Heritage and the Gaeltacht (DAHG), internal and external reports and communications are conducted bilingually.

However, the current demand for translated content exceeds the capacity of the government's translation departments. To date, the DAHG translation team has used the commercial translation memory (TM) toolset SDL Trados[7] to facilitate the re-use of previous translations in order to speed up their workflow. The benefits of TM tools are, however, limited as they can only assist translators in translating text similar to that to which they have previously been exposed. Previously unseen text therefore poses problems to translators who produce translations of a varied and changing nature. In this type of translation environment, MT systems have proven highly beneficial in improving the efficiency and speed of translation output (Federico et al., 2012). Thus, this project facilitates the integration of a SMT system into the DAHG translation workflow.

3 Irish

Irish (Gaeilge) is an Indo-European language, belonging to the Celtic language branch. It is the national language of Ireland, with 1.7 million reported L2 speakers in the Republic of Ireland. However, only 77,000 people use it in their daily lives outside the education system. [8]

The Irish language presents some interesting linguistic features that can impact the quality of machine translation systems, particularly when paired with English – a relatively uninflected subject-verb-object (SVO) language. For example:

VSO word-order Unlike the majority of other Indo-European languages, the syntax of Irish follows a relatively strict Verb-Subject-Object (VSO) word order. This gives rise to interesting challenges, both for the language learners of Irish, and for machine translation processing of a divergent word order language pair such as English and Irish.

Copula Irish has two forms of the verb 'to be'. The substantive verb *bí* 'to be' predicates more temporal qualities, existence, location, possessive, and so on (Stenson, 1981). On the other hand, a separate copula form is used for permanent, unchanging situations, in constructions that link two nouns such as identity constructions or classification constructions. Both forms determine differing syntactic structures.

Morphology Translating into a morphologically rich language is more difficult than vice versa as the linguistic information simply is not present in the source language (Minkov et al., 2007; Fraser et al., 2014). In this context, it is important to note that Irish is regarded as a language rich in morphology, which, through data sparsity, impacts developments in Irish NLP (Lynn et al., 2013).

For example, one of the ways in which nouns inflect in Irish is to convey case. Irish contains four nominal cases: nominative, vocative, dative and genitive[9]. Each noun belongs to one of five declensions, with various rules to indicate morphological changes according to their grammatical case. In addition, Irish adjectives inflect for number, gender and case. Irish verbs also inflect for tense, mood, person and gender, featuring both analytic and synthetic forms. Common to Celtic languages, initial mutation is a frequent type of Irish inflection, where the initial phoneme of a word is altered through lenition (*boird* → *b**h**oird*) and eclipsis (*bord* → ***m**bord*).

4 Data and Experiments

The success of a SMT system relies on high quality, well-aligned bilingual data. With this type of data, a system can learn, through machine learning methods, how to predict translations for previously unseen text. The more data available to the system, the higher the accuracy score of the translated output. Therefore, the challenge for developing SMT systems for minority and low-resourced languages lies in the scarcity of available parallel data.

The following describes our data collection in establishing a baseline translation system and the subsequent data collection and tuning to build our current translation system.

4.1 Baseline System

4.1.1 Training data

At the outset, an insufficient amount of bilingual data was available to build an accurate SMT system. The existing publicly available data was:

(i) **Parallel English–Irish corpus of legal texts (Paradocs)** A parallel English–Irish corpus of legal texts containing over 98,000 sentences. [10] The language used in Paradocs is very technical and unnatural, containing much legal jargon. The nature of this text limits its suitability to developing a SMT system for legal documents only.

(ii) **CCGB** (Corpas Comhthreomhar Gaeilge–Béarla) is a bilingual corpus crawled from the web.[11] This data is a raw collection of English–Irish data that would require cleaning and alignment preprocessing steps to ensure suitability for accurate machine translation.

Domain-specific parallel data In order for a SMT system to perform well on a specific domain, it is important that the training data contains documents from that domain. DAHG provided us with their translation memory (TMX)[12] files, from which parallel text was easily extracted and aligned on a segment level. In addition, they provided us with documents that had been translated by

[7] http://www.translationzone.com/trados.html

[8] Source:http://www.cso.ie/en/census/census2011reports/ The total population of Ireland for 2011 was just over 4.5 million.

[9] The dative case only applies to pronouns, and the "common" case generally refers to nominative, accusative and dative cases.

[10] http://www.gaois.ie/crp/en/

[11] http://borel.slu.edu/corpas/index.html

[12] A standard exchange format for TM files.

their translation team prior to the introduction of a TM tool into their workflow. Our domain-specific data set is therefore representative of the project's use case – reports, staff notices, communications, annual reports, and so on, written in a formal tone and at times, a high register.

General-domain parallel data In order to expand the parallel data set, we used the ILSP web-crawler[13] (Papavassiliou et al., 2013) to crawl web-pages containing the same information in both English and Irish. We found that websites providing public reference material were the best sources of parallel text for our purposes. [14]

However, the crawling task proved difficult at times as the translations provided on many websites were often only a summarisation of the source text instead of a direct translation. In addition, while the crawler relies on consistency in webpage labelling that clearly indicates the content's language, we found this was not the case for many Irish websites. Despite this, more than 10,000 suitable sentence pairs were collected from web sources and added to the training data.

All of the data collected from both the web and DAHG was pre-processed before being added to the training data. This stage involved full cleaning (removal of formatting such as XML or HTML tags) and accurate manual alignment. This data collection and curation exercise created a high quality parallel dataset of English–Irish text suitable for multiple text types.

4.1.2 Test Data

For our results to be indicative of how the system will actually be used, it is important that a suitable test set be developed. This test data should be domain-specific, with similar language and level of formality.

To meet these requirements a random selection of 1500 sentences from the domain-specific translation memories was held out for testing purposes.

4.1.3 Establishing a baseline score

The Tapadóir baseline system was built using Moses (Koehn et al., 2007) [15]. Initially the engine was trained on the TMX data from the DAHG, the Paradocs legal corpus, CCGB and the additional data that was crawled from the web, described in Section 4.1.1. The training data was tokenised and truecased in preparation for Moses. Using the cleaned data, a language model and translation model were built, and used to train the SMT engine.

This initial baseline was then tested on the test data described in Section 4.1.2. The results of automated testing (using BLEU (Papineni et al., 2002) and TER (Snover et al., 2006)) of various configurations of training data are shown in Table 1. As the results show the best performing configuration for the baseline system is trained on the DAHG translation memories and the CCGB web corpus, with the legal Paradocs corpus used only to enhance the language model.

System Training Data	BLEU	TER
CCGB + TM	39.36	0.479
CCGB + TM + Crawled	39.20	0.479
CCGB + TM + Paradocs	38.93	0.485
CCGB + TM + Paradocs + Crawled	38.80	0.484
CCGB + TM + (Paradocs)	**39.44**	**0.476**
CCGB + TM + Crawled + (Paradocs)	39.25	0.477

Table 1: BLEU and TER evaluations for the Baseline system trained on various combinations of the data available. Brackets indicate that the data was used to train the language model, but not the translation model.

In order to get some perspective on our results, we did some simple benchmarking against the Google Translate engine.[16] To achieve this, we used Google Translate to translate the test set. We then scored the translated output using the same gold standard translations and automatic metrics that we have used to measure the performance of the Tapadóir MT engines. The results in Table 3 show our system were found to score better than Google Translate according to both BLEU and TER evaluations.

4.2 Current system

Our current system is trained[17] on the data described in Section 4.1.1 (excluding CCGB and Paradocs, see Section 4.2.1) and the newly acquired data described in Section 4.2.1, and presented in Table 2. It also utilises the modifications outlined in Section 4.2.2. We show that our system continues to outperform Google Translate, reaching an increase of almost 10 BLEU points (see Table 3) when tested with our domain-specific test data.

4.2.1 Additional data considerations

The results of the baseline experiments (see Table 1) show that varying the amount and type of data in the training set has considerable impact on the performance. Thus, it became clear that additional high quality training data could have the greatest impact. Therefore our next step was further data collection.

Domain-specific parallel data At this stage of development, further manual translation had taken place at DAHG, creating additional TM data. This increased the size, coverage and quality of the domain-specific training data. This has added an extra 13,500 sentence pairs to our training data, and with Tapadóir now an integral part in the governmental translators' work,[18] this is set to increase on an ongoing basis.

Two additional translation memories, DCEP and DGT-TM (see Section 1) were made available by the Joint Research Centre of the European Commission in February

[13] Maligna was used to align segments.

[14] The ILSP crawler 'seed file' contained the following links: `www.education.ie`, `www.oideas-gael.ie`, `www.revenue.ie`, `www.udaras.ie`, `www.ahg.gov.ie`, `www.foras.ie`, `www.cnag.ie`.

[15] Version 2.0 was used for our baseline experiments, with a 6-gram language model and a maximum segment length of 6

[16] http://translate.google.com

[17] Moses version 3.0 was used to train the current system.

[18] With the performance of Tapadóir engines surpassing that of free online services, DAHG deployed the baseline engine in support of their translation workflows.

2015. Despite a significant portion of this data containing repetitions, this provided us with over 29,000 well aligned, relevant sentence pairs.

Through extensive experimentation and testing, neither the CCGB nor the ParaDocs corpora proved sufficiently domain-specific to be included in the translation model. An increase in BLEU from 39.44 to 40.34 was observed when CCGB was replaced by DCEP and DGT-TM, and Paradocs was used only in training the language model.

Corpus	Size (lines)	Size (words)
DAHG (baseline)	29,000	67418
DAHG (additional)	13,500	68691
CCGB	6,000	113889
Crawled (cleaned)	10,000	183,999
Crawled (uncleaned)	55,000	1,062,942
DCEP & DGT-TM	29,000	439,262
ParaDocs	89,000	1,526,498

Table 2: Current data sets collected for Irish↔English MT. Word counts given for the English files only.

4.2.2 System Setting Modifications

Tuning Tuning is a process whereby a source language text is translated, and compared to the gold standard target language reference. Using this comparison, the parameters of the system are adjusted, re-weighting the different features in order to produce a translation as close as possible to the reference. The closer the tuning set is to the type of text the system is intended to translate, the better the results should be when the system is in use. By introducing a tuning phase to the system training the BLEU score on our test set increased from 39.44 to 39.69. Tuning was performed on a held-out section of 3000 sentence pairs from the acquired domain-specific TM data (see Section 4.1.1.

Hierarchical-based model Reordering table(s) are features of phrase-based translation models. We experimented with changing the reordering table used from phrase-based orientation to hierarchical (Galley and Manning, 2008). These are seamlessly integrated into a standard phrase-based MT system. A hierarchical reordering model is better able to handle larger ordering differences, by treating adjacent phrases as a single unit. This is particularly suitable to divergent order language pairs such as English and Irish, and is reflected in the improved BLEU scores (39.44 to 39.63).

	BLEU	TER
Google	33.91	0.506
Tapadóir (baseline)	39.44	0.476
Tapadóir (current system)	43.08	0.463

Table 3: Comparison of results for Google, our baseline system and our current system

5 Ongoing and Future Work

In this section, we discuss the ongoing data gathering, cleaning and curation of our corpora, and the number of other strands of research we are exploring to improve the system's performance and deal with issues specific to the Irish language.

Additional general-domain parallel data Our web crawling was scaled up to provide more general domain data. This data was then preprocessed (aligned and cleaned) and added to the bilingual corpus. However, this out-of-domain data was found to be significantly more noisy than the TM data. Without strict authoring rules to adhere to, the translations were often indirect, creative or even missing. As a result, significant time was spent assessing the quality of the crawled data and preprocessing usable data. 55,000 parallel sentence pairs of crawled data have been collected, and will be added following QA testing.

Source Side Re-ordering Source-side reordering (SSRO) is a pre-processing technique for the text in the source sentence. SSRO moves constituent words and phrases in the source sentence so that the words appear in an order more like that of the target language. This transformed sentence is then translated. The effect of SSRO on translation productivity has already been observed on language pairs with divergent word order such as Japanese and English (Zhechev, 2012). Rules were applied to move the syntactic heads of the English sentence to the position they would take in an equivalent Japanese sentence, resulting in a significant performance increase.

Following this vein, we investigated the use of SSRO in our pipeline. The reordering rules focused on linguistic differences between English and Irish, changing the word order from SVO to VSO, and dealing with copula and infinitive verb constructions. The work to date on SSRO has improved the performance of our baseline system on the test set to a BLEU score of 39.52 with no effect on TER. However, it was observed (see Section 4.2.2) that switching Moses from the default phrase-based translation model (baseline system) to a hierarchical one was more effective than our SSRO modules, bringing the BLEU score to 39.63 and TER to 0.474. Combining the SSRO with a hierarchical model did not, as expected, improve the results further. Instead it had a negative effect, the cause of which is still under investigation at the time of writing. We provide an example of English–Irish SSRO in Example 1. [19]

Example 1

SRC: The timeframe can [be_extended]
RO: can The timeframe [be_extended]
REF: 'Is féidir an tráthchlár [a_shíneadh]'

Automated PE We are currently developing a new automated post-editing module, which can be applied to MT output. This module corrects common MT mistakes and encodes rules which enforce language specific constraints such as removing or correcting orthographic impossibilities. It uses Irish surface orthography rather than deeper morphological analysis to correct morphological errors

[19] SRC = Source, RO = Reordered Source and REF = Translation Reference.

which appear in the MT output. While the corrections made by this module are minor and do not always bring the translation in line with the reference translation, they do improve grammaticality and readability, reducing the repetitiveness of the post-editor's work. [20] While the aim was to correct errors and improve readability and grammar, initial tests show an improvement in BLEU score (from 43.08 to 43.18) in our current best configuration.

Factored Models The problems associated with translating a morphologically rich language are not unique to Irish. Other work (Koehn and Hoang, 2007) has shown that factored models built on parts-of-speech (POS) can be prove beneficial in these contexts. In these instances, a POS tagger is used to tag training data. A language model is then built which considers sequences of POS tags instead of sequences of tokens alone. This type of model can be used alongside the existing "traditional" language model to improve, for example, the sentence structure of the MT output. Currently we are using the Irish POS tagger (Dhonnchadha, 2009) to build a factored model for use in our MT system.

6 Conclusion

We have developed a robust English→Irish SMT system, which has reached the necessary benchmark to be deployed and integrated into existing translation workflows within an Irish government department. We have also shown how we are addressing the challenge of data sparsity and divergent linguistic structures across the language pair. The modifications described in Section 4.2 show the progress of the system through incremental improvements in training data and system configurations.

A key enabler of this project is the availability of parallel data. We gathered data from a wide variety of sources and turned this raw data into high quality datasets suitable for SMT research. In addition, the deployment of the engine in a real world environment means that future translated content can be added to the training corpus for further fine-tuning, resulting in an ongoing improvement of Irish SMT resources.

In funding Tapadóir, the Irish government have therefore begun to acknowledge the importance of creating sufficient language technology resources for a less-resourced European language. At the time of writing the authors are in negotiation with the government regarding publishing of the project data. This would make these resources freely available to the community and have significant impact on the technical readiness of the Irish language in the framework proposed by the META-NET language white papers (Judge et al., 2012)).

7 References

Dhonnchadha, Elaine Uí, 2009. *Part-of-Speech Tagging and Partial Parsing for Irish using Finite-State Transducers and Constraint Grammar*. Ph.D. thesis, Dublin City University.

Federico, Marcello, Alessandro Cattelan, and Marco Trombetti, 2012. Measuring user productivity in machine translation enhanced computer assisted translation. In *Proceedings of AMTA'12*.

Fraser, Alexander M., Kevin Knight, Philipp Koehn, Helmut Schmid, and Hans Uszkoreit, 2014. Statistical Techniques for Translating to Morphologically Rich Languages (Dagstuhl Seminar 14061). *Dagstuhl Reports*, 4(2):1–16.

Galley, Michel and Christopher D Manning, 2008. A simple and effective hierarchical phrase reordering model. In *Proceedings of EMNLP'08*. ACL.

Judge, John, Ailbhe Ní Chasaide, Rose Ní Dhubhda, Kevin P. Scannell, and Elaine Uí Dhonnchadha, 2012. *The Irish Language in the Digital Age*. META-NET White Paper Series: Europe's Languages in the Digital Age. Springer.

Koehn, Philipp and Hieu Hoang, 2007. Factored translation models. In *Proceedings of the Joint Conference on EMNLP and CoNLL'07)*. Prague, Czech Republic: ACL.

Koehn, Philipp, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst, 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, ACL '07. Stroudsburg, PA, USA: ACL.

Lynn, Teresa, Jennifer Foster, and Mark Dras, 2013. Working with a small dataset – semi-supervised dependency parsing for Irish. In *Proceedings of SPMRL'13*. Seattle, Washington, USA: ACL.

Minkov, Einat, Kristina Toutanova, and Hisami Suzuki, 2007. Generating complex morphology for machine translation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*. Prague, Czech Republic: ACL.

Papavassiliou, Vassilis, Prokopis Prokopidis, and Gregor Thurmair, 2013. A modular open-source focused crawler for mining monolingual and bilingual corpora from the web. In *Proceedings of the Sixth Workshop on Building and Using Comparable Corpora*. Sofia, Bulgaria: ACL.

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu, 2002. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of ACL'02*. Stroudsburg, PA, USA: ACL.

Snover, Matthew, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul, 2006. A study of translation edit rate with targeted human annotation. In *In Proceedings of AMTA'06*.

Stenson, Nancy, 1981. *Studies in Irish Syntax*. Tübingen: Gunter Narr Verlag.

Zhechev, Ventsislav, 2012. Machine Translation Infrastructure and Post-editing Performance at Autodesk. In *AMTA 2012 Workshop on Post-Editing Technology and Practice (WPTP 2012)*. San Diego, USA: AMTA.

[20] We also plan to assess the suitability of the open-source grammar checker 'An Gramadóir' developed by Kevin Scannell for this purpose `http://borel.slu.edu/gramadoir/`.