Embedding English to Welsh MT in a Private Company

Myfyr Prys
Cymen Cyf
Bangor University
Myfyr@Cymen.co.uk

Dewi Bryn Jones
Bangor University
d.b.jones@bangor.ac.uk

Abstract

This paper reports on a Knowledge Transfer Partnership (KTP) project that aimed to implement machine translation technology at a Welsh Language Service Provider, Cymen Cyf[1]. The project involved leveraging the company's large supply of previous translations in order to train custom domain-specific translation engines for its various clients. BLEU scores achieved ranged from 59.06 for the largest domain-specific engine to 48.53 to the smallest. A small experiment using the TAUS DQF productivity evaluation tool (Görög, 2014) was also run on the highest-scoring translation engine, which showed an average productivity gain of 30% across all translators. Domain-specific engines were ultimately successfully introduced into the workflow for two main clients, although a lack of domain specific data proved problematic for others. Various techniques such as domain-adaptation as well as improved tagging of previous translations may ameliorate this situation in the future.

1 Introduction

The translation industry in Wales has seen substantial growth over the past few decades, particularly in response to political pressures. Government legislation currently obliges all public sector bodies to produce bilingual versions of all public-facing documents, while sociocultural pressure has also influenced private businesses to invest in translation services. But the mounting demand for translation services presents challenges as well as opportunities for Welsh Linguistic Service Providers (hereafter LSPs). LSPs need to balance expenditure (on staff and equipment) with the capacity to deal with existing demands for services. Technology provides one answer to this challenge, as the work of a single translator can be extended.

A report by Bangor University's Language Technology Unit (Prys et al., 2009) found that using various kinds of translation technology could raise the economic productivity of the Welsh translation industry by 40% and could also prevent the undercutting of translation services by foreign providers leveraging new technology (2009: 23). The uptake of translation technology in Wales has been slow however, with various surveys (Prys et al., 2009 and Andrews 2010) reporting percentages of Welsh translators using translation environment technology as low as 49% and 50%, compared to the figures of 82% (Lagoudaki, 2006) and 65% (EU Commission, 2017) reported at the international level[2] and in the UK respectively. While low adoption rates for new technology may seem inevitable in the context of a lesser-resourced language, the Welsh Government has made the expansion of such tools an important part of its strategy to reach a million Welsh speakers by 2050 (Welsh Government, 2019: 34).

[1] Cymen Cyf have given their permission to be discussed as part of this study.

[2] The survey was completed by 874 respondents from 54 countries. The author does not provide information on the linguistic backgrounds of respondents, but does mention that the survey had to be completed in English, which could mean that results were biased towards "English-speaking professionals" (Lagoudaki, 2006: 6).

One tool which the Welsh Government has to promote specialist training and skills in the private sector is the Knowledge Transfer Partnership, or KTP. KTPs involve a partnership in which a university works together with a private business in order to transfer academic knowledge relating to a specific field. The project described in this paper involved a KTP between a Welsh University and a North Wales LSP, Cymen Cyf.

2 Cymen as an innovative Welsh LSP

Cymen was first established during the mid-eighties amid a wave of expansion in demand for English to Welsh translation (Andrews, 2015). The demographic profile of staff at Cymen fits the data reported by Prys et al. (2009), with a workforce which is primarily rural, female and educated to an advanced level. Staff are almost all recipients of further degrees in Welsh, which provides the fundamental skillset for the challenging task of English to Welsh translation. Cymen belies the typical image of a Welsh translation company, however, in that it has embraced a technological approach to translation. The use of translation memories[3] and termbases is well-established at the company, partly as a result of a KTP project in 2000 which led to the adoption of SDL's Translator's Workbench software, and later Trados SDL.

An analysis of Cymen's translation memories shows that at least 300,000 words are translated by the company's 16 translators each month using the Trados translation environment. Machine translation had not been implemented in the company until the advent of this project. Translation companies generally have two main options in this regard: to use a pre-existing paid service or to integrate some technical expertise into the company in order to implement an open source solution.

The first option is problematic for a variety of reasons: companies can quickly be locked in to services with little flexibility or control, and consequently may not be able to make the most of their translation engines. The second option, which involves integrating technical expertise into the company, has the advantage of leveraging free, open-source software with a flexible implementation. In practice this means that a company can create custom translation engines using their own data, while avoiding any potential data-protection concerns which may arise from having to hand over data to a multinational company. The aim of the KTP project was to realize the second option, using Cymen's existing archive of past translations to train domain-specific translation engines. Where possible, we also hoped to transfer the relevant expertise to the company's own staff.

3 Related work on MT

Previous attempts to create machine translation systems for the Welsh-English language pair are reported in Jones and Eisle (2006) and Tyers and Donnelly (2009). Jones and Eisle developed a baseline statistical machine translation system using Pharaoh (Koehn, 2004), a precursor to Moses SMT. They trained Welsh to English and English to Welsh engines on a 510,813 segment corpus extracted from the Record of Proceedings of the National Assembly for Wales. The authors report a BLEU[4] score of 40.22 for the Welsh to English engine and 36.17 for the English to Welsh engine.

Tyers and Donnelly (2009) developed a Welsh to English module for the rule-based machine translation (RBMT) system *Apertium*. The BLEU scores they report are relatively low as one might expect for an RBMT system, with a score of 15.68 for the Record of Proceedings of the National Assembly corpus. The authors argue, however, that such systems are crucial for lesser resourced languages like Welsh, drawing attention to the lack of publically available training corpora with open licensing. Beyond open source implementations, private companies such as Google and Microsoft provide English-Welsh translation engines that can be used within

[3] Translation memories are databases that store previous translations as segmented text. These segments can be retrieved and re-used to substantially speed-up repetitive translation work.

[4] BLEU (Papineni et al., 2001) is an algorithm that enables the automatic evaluation of a translation engine's output on a scale of 0 to 100, with higher scores indicating better translations. It works through comparing the engine's output with a reference translation of the same text, which is produced by a human translator.

translation software. A recent study by Screen (2018) offers evidence that using Google Translate within Trados for English to Welsh post-editing tripled translators' productivity, and cut typing by half. Although clearly effective, these services have some drawbacks including the need for payment and a lack of flexibility.

In terms of related language pairs, the ADAPT Centre team at Dublin City University have reported on their English to Irish Moses SMT engine *Tapadóir*, which was developed for use at a Government department responsible for Irish language affairs (Dowling et al., 2015). The team achieved an optimal BLEU score by combining in- and out of domain data, drawing on a mixture of publically available corpora, web-crawled data and domain-specific translation memory data. However, the relative sparsity of the available data and the comparative complexity of Irish morphology reportedly caused some problems. The team later developed an automatic post-editing module (APE) that allowed correction of certain repeated errors caused by these sparse data issues (Dowling et al., 2016). Dowling et al. (2018) reports on a comparison between the hybrid Moses SMT *Tapadóir* implementation and a newly developed NMT engine trained on the same data set. *Tapadóir* outperformed the baseline NMT engine by 8.75 BLEU points, although using byte pair encoding (Sennrich, 2016) with the NMT engine narrowed the gap slightly to 6.4 BLEU points. The authors argue that the poor performance of NMT in this case is largely due to the Irish language target exhibiting "many of the known challenges that NMT currently struggles with (data scarcity, long sentences and rich morphology)" (2018: 18).

Attempts to implement NMT for translation into another under-resourced and morphologically complex language, Basque, achieved more positive results in a recent study (Etchegoyhen et al., 2018) which found that an NMT system outperformed SMT by 4 BLEU points in Spanish to Basque machine translation. The best explanation for this discrepancy lies in the relative sizes of the corpora used for training. *Tapadóir* was trained on 108,796 parallel segments, while the MODELA engine was trained on 3,345,763 – a vast difference. Given that NMT is known to suffer from data scarcity, it seems clear that the greatest challenge facing lesser-resourced languages is the requisition of sufficient data suitable for training.

4 Data collection and preparation

The MT system implemented is based on the Welsh National Language Technologies Portal's (Prys and Jones, 2018) provision of Moses SMT (Jones et al., 2016).[5] This is a baseline implementation of Moses SMT with a simplified interface and the ability to run a Moses server instance from a Docker container image. Some advantages of the implementation are that it simplifies installation as well as subsequent training, tokenization and truecasing processes, streamlines the use of a Moses server API in third-party applications, and provides Docker containerization options. The machine translation provision was further expanded during the KTP project to include automatic tuning and evaluation of the Moses model using MERT[6] (Och, 2003) and BLEU (Papineni et al., 2002) respectively. Translation engines are trained using TMX files (an xml specification for transferring translation data between different localization software) extracted from Cymen's various translation memories.

TMXs were chosen as our main focus because they contain source and target segments already aligned, dispensing with the need for complicated alignment processes, and normally contain relatively clean data which has been carefully curated by the company. They were also convenient because the company's archive of previous translations was already largely available in this format.

Cymen's translation memory workflow revolves around a policy of assigning a TM to each regular client (although they also have some general domain memories, such as 'health' or 'education'). For instance, in the case of a fictional client named *Ideore*, the process would work as follows:

[5] https://hub.docker.com/r/techiaith/moses-smt

[6] MERT is a tuning algorithm which uses BLEU to find the optimal weights for various model features e.g. the language model, re-ordering model, and more. This process can significantly improve the quality of a translation engine.

1. Work from the client becomes frequent, so the company creates a dedicated *Ideore* translation memory and termbase.
2. These resources are consolidated into a project template file that facilitates the creation of *Ideore* projects by admin staff.
3. *Ideore* projects are allocated to specific translators based on availability and expertise, and the translation memory starts to fill.

In order to process a TMX file for training translation engines, certain pre-processing steps are necessary. Welsh and English segments are extracted from the TMX files and are stored in an SQL database, having been tagged for metadata such as domain (usually the client's name), language pair, date, and more. Different permutations of data can then be selected and exported to parallel text files for training and testing. Following this the data is randomized and split into three parts. Two held-out data sets are created - a 3,000 segment test set for evaluation with BLEU and a 2,000 segment tuning set for tuning with MERT. The language model is created from the target side of the training corpus. Finally, segments from both the training set and tuning set are removed from the main training corpus and language model to avoid skewing the evaluation and tuning steps.

5 Training the engines

We decided to set an arbitrary threshold of a million Welsh words before attempting to evaluate the baseline capability of engines trained on such data. Figure 1 below shows all of Cymen's TMs arranged by number of Welsh words.

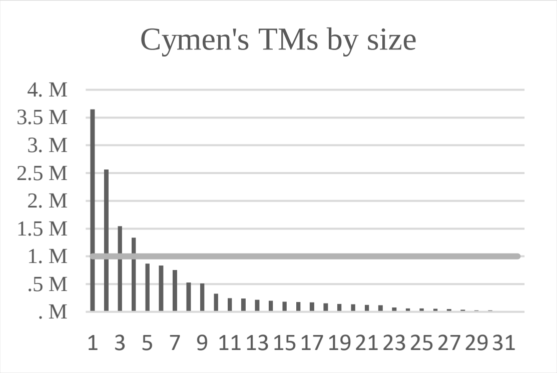

Figure 1. Each of Cymen's TMs arranged by number of Welsh words, from largest to smallest

As can be seen in the chart, only 4 of our TMs are currently large enough to satisfy this criterion. Table 1 displays BLEU scores for engines trained from these TMs client-specific TMX files, as well as the general-domain Cymen translation engine. As might be expected based on previous research (e.g. Koehn 2001), the scores seem to be related to the size of the corpus used for training as well as the specificity of the domain. The highest scoring engine is the domain-specific engine 1, which was trained on a 174,354 segment parallel corpus of data relating to a client in a technology-related domain. Although the size of the parallel corpus used for training is an obvious contributor to its relatively high score, the nature of the domain, which consists of a highly technical and repetitive register, also seems to be a factor.

Translation Engine ID	Number of Welsh words	Number of segments	BLEU score
1	3.65 million	174,354	59.06
2	2.56 million	130,235	58.75
3	1.54 million	83,745	50.92
4	1.34 million	74,840	48.53
Cymen	65.3 million	3,985,674	54.23

Table 1. BLEU scores and corpus size for the five top translation engines

Engines three and four had substantially lower scores, reflecting the smaller corpora used for training. For comparison, we also trained a general domain corpus consisting of all of Cymen's combined data (named Cymen in table 1). Although trained on a comparatively large data set, translation engines 1 and 2 still outperform this engine in terms of BLEU score, which provides some indication of the value of using domain-specific engines.

6 Engine effectiveness

To gain a general idea of the effectiveness of our engines, we used the TAUS DQF evaluation tool (Görög, 2014) to carry out a productivity test on segments automatically translated by our highest-scoring domain-specific translation engine (ID 1, BLEU score 59.06 – see table 1 above). Eight translators were selected to translate 50 in-domain segments from a held-out data set, with a total of

905 words. The segments were randomly selected before being submitted to the TAUS DQF evaluation tool. TAUS DQF automatically shuffles the segments and presents half to be translated from scratch (i.e. without the machine-translated output) and half to be post-edited (with the machine-translated output). The tool then times the completion of each segment and generates a report based on the average completion time for both conditions.

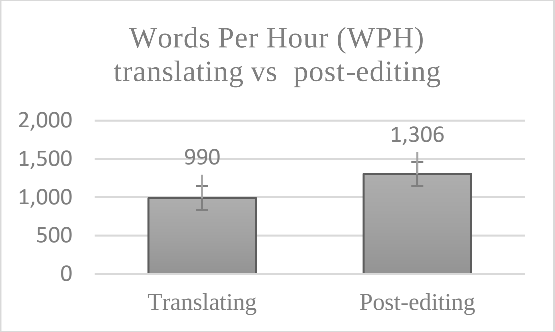

Figure 3. The average WPH of Cymen's translators while translating vs post-editing

The results (figure 3) show that the participants produced 1,372 words per hour on average while post-editing as opposed to 1,055 words per hour while translating from scratch. Individual results for the translators (figure 4) show quite considerable variation, although all translators performed more quickly in the post-editing condition.

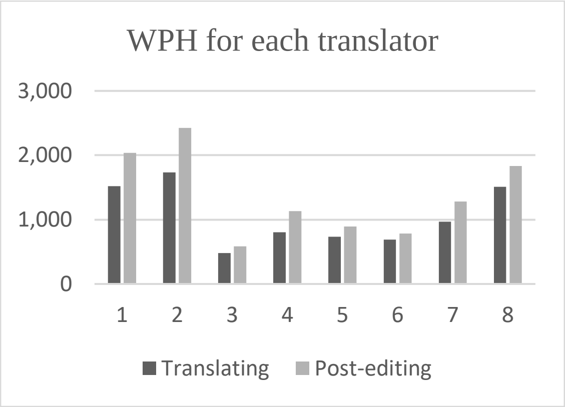

Figure 4. WPH per condition for each translator

The result of this analysis suggests that using a machine of this quality can increase productivity across all translators by 30%, while individual gains vary from a high of 41% to a low of 14%.

7 Implementing the engines

Once we were satisfied that the quality of the engines was of a satisfactory level, the engines needed to be embedded in the company's translation environment software. An app for integrating the engines was designed on the basis of an open-source C# solution available on GitHub[7]. In order to allow the selection of multiple engines, distinct engines are run from their own Docker containers, which can each then be selected from within the Trados interface.

Before the engines could be fully integrated into translators' workflows, the company's management team trialed their performance personally for a probationary period. Based on this the management decided to start using engines 1 and 2 in every project for the relevant clients. As engines 3 and 4 were not considered of a high enough quality to use routinely, we decided not to implement them for the time being. Given that there is a clear link between the size of a domain-specific engine and its effectiveness, and that the company's store of data for each client is always growing, it was decided that we would attempt to retrain using these client's data at a later date. It was also decided that we would start gathering analytic data on the size and relative growth of translation memories on a monthly basis. This should allow the company to make informed decisions concerning whether data belonging to a particular client has reached a point where training an effective translation engine for it has become viable.

8 Reception by translators

In order to take possible resistance from translators to machine translation into account, we introduced the engines into their workflow gradually. Firstly, we waited until the output of the translation engines was of a relatively high quality - based on the management team's assessment - before introducing them into the daily workflows. This hopefully mitigated the prospect of poor machine translation irreparably damaging translators' feelings towards the technology. The practical implementation of the engines was also

[7] https://github.com/OpenNMT/Plugins/tree/master/SDL%20Trados%20Plugin

relatively subtle, with engine allocation happening at management level, meaning that translators did not need to take any actions themselves. Finally, a series of four workshops were held for all staff, where the KTP assistant was able to describe the basic principles of the technology. Particular emphasis was placed on the fact that machine translation is a post-editing tool, which cannot replace human translators. It was also stressed that productivity gains associated with less typing and additional time can make the translator's work less laborious and more comfortable.

In general the company's reception of the technology has been positive. The most obvious manifestation of this is that engines 1 and 2 are now used routinely for translating those clients' respective domain-specific data, which taken together represent a large proportion of the company's daily output. The general domain *Cymen* engine is also now used frequently for translating data for smaller clients that have data particularly suitable for machine translation. One feature that was repeatedly praised by the translators was the autosuggest capability[8], which prompts the user with suggested words or phrases extracted from the translation engine as they type. This was seen as particularly useful because translators were able to leverage useful elements of an engine's output even when the segment as a whole was not perfect.

9 Future research

Following the successful implementation of machine translation during the KTP, both partners are interested in extending the capabilities of the translation system. Obvious candidates for such improvements include neural and/or hybrid translation systems, which have not yet been reported in open source implementations for Welsh. However, the primary challenge facing Cymen is the lack of sufficient data for training domain-specific engines for the majority of its clients. Exploring domain adaptation techniques (e.g. Axelrod et al., 2011), which allow out-of-domain data to be leveraged for domain-specific engines, offers one way of dealing with the scarcity of data for some domains discussed above. Otherwise, the main way that Cymen can improve its translation engines is through the natural growth of its translation memory archive through daily translation work, which continues apace.

10 Conclusion

This paper has discussed the implementation of open-source machine translation software at a Welsh translation company. We have shown that leveraging a private company's archive of previous translations to train domain-specific translation engines is a relatively straightforward task, although the success of the endeavour is to some extent dependent on the company storing translations with some kind of metadata indicating domain. This shows the importance of educating the translation sector in Wales (and beyond) in the value of such data and the importance of storing in such a way that its usefulness for future MT tasks is maximized.

References

Andrews, T. (2015) Cyd-destun gwleidyddol a chymdeithasol cyfieithu yn y Gymru gyfoes. In: Prys, D & Trefor, R. (Eds). *Ysgrifau a Chanllawiau Cyfieithu*. [Online]. Carmarthen: Coleg Cymraeg Cenedlaethol. Available at: https://llyfrgell.porth.ac.uk/media/ysgrifau-a-chanllawiau-cyfieithu-delyth-prys-arobat-trefor-goln [Retrieved: 27/01/2016].

Dowling, M., Cassidy, L., Maguire, E., Lynn, T., Srivastava, A., and Judge, J. (2015). Tapadóir: Developing a statistical machine translation engine and associated resources for Irish. In *Proceedings of the The Fourth LRL Workshop: "Language Technologies in support of Less-Resourced Languages",* Poznan, Poland.

Dowling, M., Lynn, T., Graham, Y., and Judge, J. (2016). English to Irish machine translation with automatic post-editing. In *2nd Celtic Language Technology Workshop* (pp. 42-54), Paris, France.

Dowling, M., Lynn, T., Poncelas, A., & Way, A. (2018). SMT versus NMT: Preliminary Comparisons for Irish. In *Workshop on Technologies for MT of Low Resource Languages* (p.p. 12-20), Boston, USA.

European Commission Representation in the UK, Chartered Institute of Linguists and the Institute of Translation and Interpreting. (2017) *2016 UK Translator Survey - Final Report.* Available online at: https://ec.europa.eu/unitedkingdom/sites/unit

[8] This feature is part of Trados software

edkingdom/files/ukts2016-final-report-web_-_18_may_2017.pdf Retrieved: [28/06/2019]. Cymraeg 2050:

Görög, A. (2014). Quantifying and benchmarking quality: the TAUS Dynamic Quality Framework. *Tradumàtica*, (12), 443-454.

Jones, D.and Andreas E. (2006). Phrase-based statistical machine translation between English and Welsh. In *Strategies for developing machine translation for minority languages (5th SALTMIL workshop on Minority Languages),* LREC-2006 (p.p. 75-78), Genoa, Italy.

Jones, D.B., Prys, D., Ghazzali, S. and Robertson, P. (2016) Facilitating the Multilingual Single Digital Market: Case Studies in Software Containerization of Language Technologies. In *Proceedings of META-FORUM 2016*, Lisbon, Portugal.

Lagoudaki, E. (2006). Translation memories survey 2006: Users' perceptions around TM use. In *proceedings of the ASLIB International Conference Translating & the Computer* (Vol. 28, No. 1, pp. 1-29). London, UK.

Och, F. J. (2003). Minimum error rate training in statistical machine translation. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1* (pp. 160-167). Sapporo, Japan.

Papineni, K., Roukos, S., Ward, T., & Zhu, W. J. (2002). BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics* (pp. 311-318). Philadelphia, USA.

Prys, D. & Jones, D.B. (2018) *National Language Technologies Portals for LRLs: a Case Study.* In: Vetulani Z., Mariani J., Kubis M. (eds) Lecture Notes in Artificial Intelligence. Springer.

Prys, D. et al. (2009). Gwell Offer Technoleg Cyfieithu ar gyfer y Diwydiant Cyfieithu yng Nghymru: Arolwg Dadansoddol [Online]. Bangor: Language Technologies Unit, Canolfan Bedwyr. Available at: https://goo.gl/52ZYfj [Retrieved: 01/05/2019].

Screen, B. (2018). Defnyddio Cyfieithu Awtomatig a Chof Cyfieithu wrth gyfieithu o'r Saesneg i'r Gymraeg: Astudiaeth ystadegol o ymdrech, cynhyrchedd ac ansawdd gan ddefnyddio data Cofnodwyr Trawiadau Bysell a Thracio Llygaid. (Unpublished PhD Thesis). Cardiff University, Wales. Available online at: http://orca.cf.ac.uk/111362/

Sennrich, R., Haddow, B., & Birch, A. (2015). Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*.

Thierry Etchegoyhen, Eva Martínez, Andoni Azpeitia, Gorka Labaka, Iñaki Alegria, Itziar Cortes, Amaia Jauregi, Igor Ellakuria, Maite Martin eta Eusebi Calonge (2018) Neural Machine Translation of Basque EAMT 2018. Alicante.

Tyers, F., & Donnelly, K. (2009). apertium-cy - a collaboratively-developed free RBMT system for Welsh to English. *The Prague Bulletin of Mathematical Linguistics*, *91*, 57-66.

Welsh Government (2019) 2050: A million Welsh speakers. Annual report 2017–18. Available at: https://gov.wales/sites/default/files/publications/2019-03/cymraeg-2050-a-million-welsh-speakers-annual-report-2017-18.pdf Retrieved: [02/07/2019].