A human evaluation of English-Irish statistical and neural machine translation

Meghan Dowling **Sheila Castilho** **Joss Moorkens** **Teresa Lynn** **Andy Way**

ADAPT Centre, Dublin City University, Ireland
firstname.lastname@adaptcentre.ie

Abstract

With official status in both Ireland and the EU, there is a need for high-quality English-Irish (EN-GA) machine translation (MT) systems which are suitable for use in a professional translation environment. While we have seen recent research on improving both statistical MT and neural MT for the EN-GA pair, the results of such systems have always been reported using automatic evaluation metrics. This paper provides the first human evaluation study of EN-GA MT using professional translators and in-domain (public administration) data for a more accurate depiction of the translation quality available via MT.

1 Introduction

The Irish language enjoys the status of both the first official language of Ireland and an official European Union language. As a result of this status is there is a requirement for official public content to be made available in Irish in both Ireland[1] and the EU.[2] There is currently a derogation on the amount of Irish content published by the EU, due to be lifted at the end of 2021 (Publications Office of the European Union, 2011). At this point, the already high demand for professional Irish translators will increase significantly. With this demand for the production of Irish-language text, usually with English as the source language, it is important that any available EN→GA MT systems are robust and fit-for-purpose.

Despite MT having been established as a useful tool in the workflow of a professional translator, it is not yet the norm for Irish translators, whether freelance or within a translation company.[3] As a lesser-resourced and minority language, Irish faces a barrier to state-of-the-art technology shown to be effective for majority languages (European Language Resource Coordination, 2020).

While there has been research on improving EN→GA MT (Dowling et al., 2015; Arcan et al., 2016; Defauw et al., 2019; Dowling et al., 2019) to date there have been no publications describing a human evaluation (HE) study for EN→GA MT. This study aims to provide the first EN→GA MT HE study, investigating the measurable usefulness of EN→GA in a professional translation capacity. In an attempt to closely match the context in which EN→GA MT is intended to be used, professional translators will undertake post-editing (PE) tasks using MT output.

Another aim of this study is to provide a human-derived comparison of EN→GA statistical machine translation (SMT) and neural machine translation (NMT). In previous work, a preliminary comparison of EN→GA SMT and NMT showed that SMT fared better than NMT in terms of automatic metrics (Dowling et al., 2018). More recent publications (Defauw et al., 2019; Dowling et al., 2019) show a more positive picture for EN→GA NMT, but without a direct comparison to SMT. The SMT/NMT comparison presented in this paper will take into account both the quantitative metadata gathered during the study (time per seg-

[1] The Official Languages Act (2003) requires all official public information and services to be available in both Irish and English: http://www.irishstatutebook.ie/eli/2003/act/32/enacted/en/html

[2] Irish has been a full EU language since 2006.

[3] A recent study by Moorkens (2020) reported that "...few participants appear to use MT at present..."

ment, number of keystrokes, etc.) as well as the qualitative opinions and recommendations of the participants.

This paper is presented as follows: Section 2 describes related work in the areas of EN→GA MT, HE, etc. In Section 3 we provide details of the data and parameters used in our SMT and NMT systems, while Section 4 describes the methodology used in this HE study. We present our results in Section 5 and provide some conclusions and avenues for future work in Section 6.

2 Related work

As Irish is a poorly-resourced language (Judge et al., 2012), the quality of MT output has struggled to reach the same level of quality as well-supported languages. Several studies (Dowling et al., 2015; Arcan et al., 2016; Dowling et al., 2016) thus focused on improving EN→GA SMT, as discussed in Section 1. In previous work comparing EN-GA SMT and NMT, preliminary results suggest that SMT seems to outperform NMT (Dowling et al., 2018), although we show a number of examples where the score may be misleading and recommend that a HE study may be necessary to fully understand the quality of each system type. More recent studies (Defauw et al., 2019; Dowling et al., 2019) show the effects of adding artificially-created training data to EN-GA NMT.

In terms of Irish translators' attitudes to MT, Moorkens' (2020) extensive survey reports varying attitudes between translators based on terms of employment, with freelance translators appearing to be poorly disposed towards MT.

Koehn and Knowles (2017) include low–resource languages as one of the main challenges still present in MT research. Unable to exploit cutting-edge techniques that require huge resources, MT researchers must look to creative solutions to improve low–resource MT. Such approaches include the creation of artificial parallel data, e.g. through back-translation (Poncelas et al., 2018), exploiting out-of-domain data (c.f. Imankulova et al., (2019)) and using a better-resourced language as a pivot (Wu and Wang, 2007; Liu et al., 2018; Cheng, 2019).

HE is a vital component of MT research (Castilho et al., 2018), with many of the major MT conferences including a translator track to encourage such publications. They are especially valuable in low-resource or minority contexts (e.g. Spanish-Galician MT (Bayón and Sánchez-Gijón, 2019), Russian-Japanese MT (Imankulova et al., 2019)) where the languages may be overlooked by global MT companies.

There have been comparisons of SMT and NMT since NMT first emerged in the field. The conference on machine translation (WMT) regularly feature both systems, with HE at the forefront (Bojar et al., 2016; Ondřej et al., 2017; Barrault et al., 2019). Castilho et al. (2017) describe an extensive comparison of SMT and NMT using both automatic metrics and HE. Mixed results overall highlight the need for language-specific HE studies.

Recently, Läubli et al., (2020) published a set of recommendations for performing HE of MT. They advocate for (1) the use of professional translators over novices, (2) translations to be evaluated on a document-level, (3) fluency to be evaluated in addition to adequacy, (4) reference translations not to be heavily edited for fluency and (5) the use of original source texts (rather than translated text as input). We take these recommendations into account when designing this HE study.

3 MT systems set-up

To compare SMT and NMT through HE it is first necessary to train a system of each type using the same training data. This section describes the data used in building both MT systems, their specific parameters and the automatic evaluation scores generated for each.

3.1 Data

Both SMT and NMT rely on large amounts of high-quality parallel data. This is especially true of NMT, a type of MT system that is highly data-driven. Although there are legal requirements regarding the creation of public Irish text (see Section 1) we may still describe Irish as a 'less-resourced' language. As mentioned previously, the derogation on the status of Irish has limited the amount of Irish content generated by the EU. Furthermore, the Irish Language Act (2003) does not enforce bilingual production of all public text and, until relatively recently, translation memories were not usually requested by public bodies when outsourcing translation work (Lynn et al., 2019).

Table 1 shows the sources and number of GA words of all datasets used to build the SMT and NMT systems. In line with previous work (Dowl-

Source	# words (GA)
DCHG	1,085,617
EU	439,262
Crawled	254,772
CnaG	21,365
Teagasc	32,908
UT	15,377
IT	57,314
Paracrawl	20,803,088
ELRC	415,648
ELRI	628,669
TOTAL	**23,754,020**

Table 1: Source and number of Irish words of data sources used to build the MT systems described in this paper

ing et al., 2019), in-domain data from the Department of Culture, Heritage and the Gaeltacht (DCHG) is used. This is supplemented with data from EU sources,[4] crawled data,[5] Conradh na Gaeilge[6] (CnaG), Teagasc,[7] University Times (UT) and the Irish Times (IT). The two latter sources contain monolingual Irish text only. As in Defauw et al., (2019), we include the Paracrawl corpus, a large corpus of webcrawled data (Esplà-Gomis et al., 2019). Further to this, we add two new corpora, referred to in Table 1 as ELRC and ELRI. ELRC refers to the European Language Resource Coordination,[8] an initiative led by the European Commission to gather language resources for all EU official languages. ELRI[9] is an initiative which focuses on the building and sharing of language resources within France, Ireland, Portugal and Spain (Etchegoyhen et al., 2018) (European Language Resource Infrastructure).

3.2 SMT parameters

When training the SMT system, we follow parameters identified in previous work. Moses (Koehn et al., 2007), the standard tool for building SMT systems, along with the data described in Section 3.1, is used to train our SMT model. KenLM (Heafield, 2011) is used to train a 6-gram language model using the GA portion of the parallel data, as well as the monolingual GA data. This wider-context language model (3-gram is the default) along with hierarchical reordering tables are used in an attempt to address the divergent word orders of EN and GA (EN having subject-verb-object and GA having verb-subject object word order.)

3.3 NMT parameters

As in other research on EN-GA NMT (Defauw et al., 2019; Dowling et al., 2018), we use OpenNMT (Klein et al., 2017) as the basis for training our NMT system. We implement a transformer-based approach (Vaswani et al., 2017), which has shown promising results for low-resource NMT with other language pairs (Lakew et al., 2017; Murray et al., 2019). We use parameters recommended by Vaswani et al., (2017).

3.4 Test data

1,500 sentences of gold-standard data,[10] with an average sentence length of 20 words per sentence, were held out from training data in order to perform automatic evaluation. This data contains extracts from DCHG sources such as official correspondence, public announcements, etc.

3.5 Automatic evaluation

Automatic evaluation metrics, while best used to track developmental changes in one particular MT system over time, can also be used to gauge differences in quality between two different MT systems. In this study we generate BLEU (Papineni et al., 2002), TER (Snover et al., 2009), CharacTER (Wang et al., 2016) and ChrF scores (Popović, 2015).

	BLEU↑	TER↓	ChrF↑	CharacTER↓
SMT	45.13	43.51	66.26	0.29
NMT	**46.58**	**40.85**	**67.21**	**0.28**

Table 2: Automatic evaluation scores for the SMT and NMT systems used to generate MT output, rounded to 2 decimal places. The best score in each column is highlighted in bold.

With automatic evaluation, the source side of the test data (EN) is translated using the MT system. BLEU and TER both compute scores by comparing words in the MT output to those in the GA por-

[4] Parallel texts from two EU bodies: the Digital Corpus of the European Parliament (DCEP) and Directorate General for Translation, Translation Memories (DGT-TM)

[5] Crawled from various sources including Citizens Information, an Irish government website that provides information on public services

[6] Conradh na Gaeilge is a public organisation tasked with the promotion of the Irish language

[7] The state agency providing research, advisory and education in agriculture, horticulture, food and rural development in Ireland.

[8] `http://www.lr-coordination.eu/`

[9] `http://www.elri-project.eu/`

[10] Professionally translated data within the same domain (from the DCHG corpus).

tion of the test data. CharacTER and chrF, however, compute a score based on a character-level comparison, which can be more accurate for inflected languages.

Table 2 shows the BLEU, TER, CharacTER and ChrF scores for the SMT and NMT systems. These scores can then be compared to the results provided through HE. Both BLEU and ChrF are precision based, with higher scores indicating higher precision and, in theory, higher quality. This is indicated with a ↑ in Table 2. TER (translation error rate) and CharacTER (TER on character level) are error-based metrics. Accordingly, a lower score represents a lower error rate, indicated with a ↓ in Table 2.

It can be seen from Table 2 that the NMT system achieves a better score across all four metrics, whether calculated on a word or character level.

4 Study methodology and set-up

MT HE can take many forms, e.g. ranking of MT output, annotation of incorrect parts of speech or post-editing of MT output. A HE study can also be carried out by providing the translators with MT output and asking them to post-edit it. One benefit of this method is that subjectivity can be decreased – data gathered through translator post-editing (e.g. time spent per segment, number of keystrokes, etc.) is used to assess the MT system rather than the participant being required to give a judgement per word/segment. It is also faster than error annotation and requires less training, particularly if the translators already have experience of post-editing MT output. It is also the method which is closest to the situation in which MT is intended to be used, and as a result translator opinions of the post-editing tasks can also be elicited. For these reasons, we see post-editing as the HE method that best suits the needs and intended outputs of this study. This section describes the set-up and methodology of the PE task and related survey.

4.1 PET tool and guidelines

Post-editing tool (PET) (Aziz et al., 2012) was chosen as the software with which to collect data for this study as it is freely available online and specifically designed for use in HE studies of MT. We configure PET with the default parameters and compose guidelines and instructions for the participants. For example, participants were permitted to use dictionaries while editing the output, but were not permitted to use another MT tool. The guidelines were written in Irish, the target language of this study.

4.2 Pilot study

Prior to the main study, we conducted a pilot study to ensure that the tool was set up correctly and to test the robustness of the guidelines. Two Irish linguists each post-edited 10 machine–translated sentences. We then updated the guidelines as per the feedback of both pilot study participants.

4.3 Data

Two subsets were extracted from the test data described in Section 3.1, each containing 100 EN sentences, and then translated with the SMT and NMT systems described in 3.2 and 3.3 respectively. With the merits of document-level translation raised in recent MT research (Toral et al., 2018; Werlen et al., 2018) and the importance of context in work using MT for dissemination, we choose to keep the sequence of sentences, rather than extract each of the 200 sentences individually at random.

Recent studies have shown that MT can have a negative impact on the linguistic richness of MT output (Vanmassenhove et al., 2019) and post-edited translations (Toral et al., 2018). To demonstrate the differences in linguistic richness between SMT and NMT, we calculate standardised type-token ratio (STTR) with the outputs.[11] Table 3 shows that, although a small difference can be seen between jobs for both systems, in average both MT systems have a very similar STTR.

System	Job 1	Job 2	Average
SMT	41.71	42.69	42.20
NMT	43.84	41.33	42.59

Table 3: Comparison of STTR between SMT and NMT outputs normalised per 1000 words

4.4 Participants

With EN-GA MT more likely to be used as a tool to help publish translated content in an official context rather than a gisting tool, it is important that the participants in this study match the profile of the intended user, namely a professional translator. To this end, we recruited participants with an ac-

[11] Type-token ratio normalised per 1,000 words.

creditation in EN-GA translation.[12] We recruited four accredited translators, referred to from now on as P1, P2, P3 and P4 respectively.

Each participant was asked to post-edit 210 sentences: 10 practice sentences, 100 sentences translated using SMT and 100 sentences translated using NMT. The same source text was provided to all 4 translators. Figure 1 shows the distribution of MT output across participants. Two participants (P1 and P3) were presented with the SMT output using Job 1 data and the NMT output using Job 2 data (set-up A). The other two participants (P2 and P4) were asked to post-edit set-up B, consisting of Job 1 machine-translated using NMT and Job 2 machine-translated using SMT. Both set-up A and set-up B contain 10 practice sentences from a similar source (Dublin City Council) so that the translators could try out the PET environment and get used to the software without worrying about speed. The output files and associated metadata from the practice segment are not included in the results.

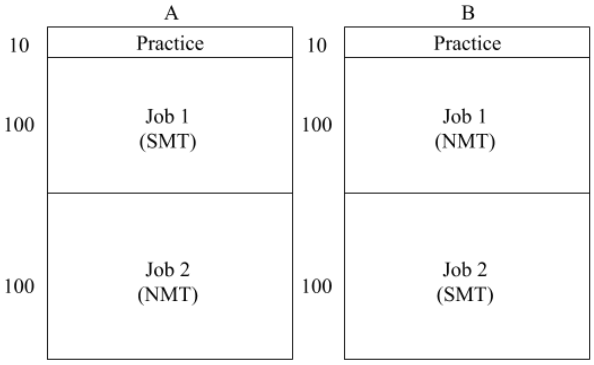

Figure 1: Distribution of MT output (not to scale)

4.5 Survey questions

A post-task survey was implemented to gather information about the participants' experience and their opinions of the two MT outputs. Participants were not informed whether the MT output was produced by an SMT or NMT system. The survey was distributed via Google sheets. The following information was gathered in the survey:

- months/years experience as a professional translator
- months/years experience post-editing MT in a professional capacity
- view of MT as a tool to be used by professional translators
- which system seems most fluent
- which system seems most accurate
- which system the translator would prefer to post-edit
- a text box for additional comments

5 Results and analysis

In this section we present the survey results and the results gathered via the PET output.

5.1 Survey results

The survey results show that all 4 participants are experienced translators. P1 has 25 years of experience, P2 5 years, P3 10 years part-time, and P4 13 years. Two of the participants' (P2 and P4) have experience post-editing (PE) MT in a professional capacity, with 3 years (P2) and 5 years (P4) of PE experience each (see Table 4).

Participant	Translator exp.	PE exp.
P1	25 years	N/A
P2	5 years	3 years
P3	10 years†	N/A
P4	13 years	5 years

Table 4: Table displaying the amount of experience (exp.) each participant has a professional EN-GA translator and, if relelvant, how much experience each has PE EN-GA output. A dagger (†) signifies that the experience is in a part-time capacity.

When asked for their views of MT as a tool to be used by professional translators, answers varied from positive ("It's a very useful tool") to cautious ("I think it depends very much on what the machine has been fed!"; "Improving constantly, but insufficient at present";) to negative ("It's not much use for English to Irish translation. It would take the same length of time to translate from scratch"). The positive but guarded responses came from participants with post-editing (PE) experience, whereas those without PE experience answered negatively. This may be an indication that there is a learning curve with PE before MT can be a valuable and useful addition to translation workflow.

[12] The Foras na Gaeilge seal of accreditation for translators. Details of translators with this accreditation who are available on a part- or full-time basis are published on the Foras na Gaeilge website: https://www.forasnagaeilge.ie/about/supporting-you/seala

Table 5 shows the survey results pertaining to differences between the two systems (SMT and NMT). The question "In general, which output did you perceive to be the most fluent-sounding?" is represented by the heading 'fluency'. 'Accuracy' is the heading used to represent the question "In general, which output did you did you perceive to be the most accurate in terms of semantics? (i.e. conveyed the meaning the best, fluency aside)." The final question dealing with SMT versus NMT, "Which output would you prefer to post-edit?" is represented with the heading 'prefer.' The participants were not aware which output was produced by which system; they were simply presented with two separate translation jobs.

Participant	fluency	accuracy	prefer
P1	NMT	NMT	No diff.
P2	No diff.	NMT	NMT
P3	No diff.	No diff.	No diff.
P4	NMT	NMT	NMT

Table 5: Survey responses relating to differences between SMT and NMT fluency, accuracy and participant preference.

5.2 PET results

Interestingly, none of the four participants gave SMT as an answer to any of these questions. This contradicts previous work comparing EN-GA SMT and NMT using automatic metrics (Dowling et al., 2018). It does, however, line up with the automatic metrics gathered during this study (BLEU, TER, ChrF and CharacTER scores suggested that the NMT output was of greater quality than that of SMT – see section 3 for more details).

The results gathered from PET provided us not only with the post-edited output, but also with the number of keystrokes, annotations, and seconds spent on each segment. We used this data to calculate the average seconds per segment, average keystrokes per segment, and the average unchanged segments per system per participant. These figures, as well as the human-targetered TER (HTER) scores (Snover et al., 2006), are displayed in Table 6. Where MT for dissemination is concerned, temporal effort, or time spent PE, is arguably the most important metric as payment is usually based on words translated. Two of the four participants in this study (P1 and P4) were more productive when working with NMT output. The difference for P4 was sizeable (an average of 48.53 seconds per segment for NMT compared to 193.06 for SMT), although it should be noted that P4 was required to repeat the PE task for the NMT job due to a technical error. It is likely that this led to a faster PE time for this job, and that other values for this job are also skewed. P2 and P3 were more productive using SMT, although for P2 the difference is negligible (an average of 119.86 seconds per segment for SMT PE in comparison to 120.59 for NMT).

HTER is a metric for evaluating MT output based on TER (see Section 3). Using HTER, a human translator post-edits MT output and the score is calculated using the post-edit distance and the length of the reference. A low HTER score should equate to low PE effort, although in practice, post-editors may delete and retype text rather than taking the shortest possible route from raw MT to PE.

In the case of P1, P2 and P4, HTER was lower for NMT than SMT. Results from P3 showed negligible difference between the HTER of both systems (a difference of 0.0004).

In the survey, P1 reported that the NMT output was more fluent-sounding and more accurate. This is reflected in the data. From Table 6 we can see that P1 was quicker, used fewer keystrokes, and left more segments unchanged when PE NMT output. P1 did, however, choose 'no difference' when asked which output they would prefer to PE.

P2 also voiced a preference for NMT output over SMT output, although reported 'no difference' in fluency. Scores generated from PET data indicated little/no difference in time, keystrokes, and unchanged segments, although the HTER score was markedly improved for NMT.

Although P3 answered 'no difference' to all three questions comparing SMT and NMT, this is not reflected in the time and keystrokes, which indicated more favourable results for SMT, nor in the unchanged segments for which NMT had a higher score. It is, however, reflected in the HTER scores which are almost identical for both outputs.

P4 reported NMT to be more fluent sounding, more accurate, and the output they would most prefer to post-edit. This is reflected in all metrics present in Table 6, where the results for the NMT output show a marked improvement over those of the SMT output, apart from the number of unchanged segments. However, as mentioned in Section 5.1, P4 had to repeat the entire PE task for the NMT output. This may have lead to a faster PE time with fewer keystrokes and, relatedly, a lower

participant	system	avg. time/seg.	avg. keys./seg.	avg unchanged segs.	HTER
1	SMT	102.4	91.47	0.12	0.33
1	NMT	89.16	89.16	0.2	0.28
2	SMT	119.86	207.09	0.11	0.52
2	NMT	120.59	205.61	0.12	0.43
3	SMT	173.15	90.44	0.17	0.36
3	NMT	207.21	139.9	0.2	0.36
4	SMT	193.06	100.49	0.1	0.43
4	NMT	48.53	48.73	0.18	0.24

Table 6: Table displaying the average (avg.) number of seconds (time) per segment (seg.), average number of keystrokes (keys.) per segments, average unchanged segments and HTER of each system for each participant.

HTER score. Overall, these results suggest HTER to be a more valuable indication than other metrics gathered.

5.3 PE output

With both the survey responses and figures generated using results from PET varying substantially from translator to translator we chose to take a closer look at the differences in PE output provided by the four participants. To identify potentially interesting sentences, we used compare-mt (Neubig et al., 2019), a tool designed to analyse MT output and provide the user with sentences which differ greatly. Although human-generated translations are not the intended input for compare-mt, it was still useful in identifying cases where the participants gave different translations.

Input:	If you have been allocated as a decision-maker..
SMT:	Má tá tú mar a déantóir cinntí..* *If you are a decision manufacturer..*
P1:	Más cinnteoir thú air.. *If you are a decision-maker for it..*
P3:	Má ainmníodh thú mar chinnteoir.. *If you are named as a decision-maker..*
NMT:	Má roghnaíodh mar chinnteoir thú.. *If you are chosen as a decision-maker..*
P2:	Má shanntar ról mar chinnteoir ort.. *If the role of decision-maker is assigned to you..*
P4:	Má roghnaíodh mar chinnteoir thú.. *If you are chosen as a decision-maker..*

Table 7: A portion of the PE output from P1, P2, P3 and P4. The EN data provided to the translators as input is also provided. The relevant MT output provided to translators is given above the participants' output. A gloss for each sentence is indicated in italics below each GA output. An asterix (*) indicates that the segment is not grammatically correct.

Table 7 shows a shortened portion of a segment of PE output produced by P1, P2, P3 and P4. It can be seen, even to those who do not speak Irish, that all four translators chose to post-edit the MT input in a different way. In fact, there is no word which is repeated (with the same inflections) throughout all four translations. Despite all being correct, it stands to reason that automatic values generated for this output, such as HTER and number of keystrokes, would also differ. This highlights the limitations of such metrics, as well as the need for multiple references when generating automatic evaluation scores.

Similarly, in Table 8, all four participants chose slightly different translations of the source text. In this example, the importance of context can be seen. In the source text, the acronym for Freedom of Information (FOI) is not expanded. Despite this, only P3 chooses to use the equivalent Irish acronym – possibly due to both MT systems producing the expanded acronym (shown in bold). The three other translators (P1, P2 and P4) chose to preserve the expanded acronym in the GA PE sentence. It could be the case that, in Irish, the acronym is not as instantly recognised as its English counterpart. This is quite common, when an acronym is commonly used in one language but not in another. Without training data to reflect this, it is unlikely that an MT system would produce such an output. This inconsistent spelling-out of the acronym in the post-edited texts again indicates the importance of in-domain training data and of seeking the advice of professional translators when training MT systems.

6 Conclusions and Future Work

We have presented the first HE study for EN-GA SMT and NMT. We have shown that, while auto-

Source	to ensure.. in the **FOI** legislation..
SMT:	chun a chinntiú.. sa reachtaíocht um **Shaoráil Faisnéise**.. *to ensure.. in legislation surrounding the **Freedom of Information..***
P1:	cinntiú.. sa reachtaíocht um **Shaoráil Faisnéise**. *ensure.. in the legislation surrounding the **Freedom of Information..***
P3:	chun a chinntiú.. sa reachtaíocht **SF**. *to ensure.. in the **FOI** legislation..*
NMT:	féachaint.. sa reachtaíocht um **Shaoráil Faisnéise**.. *see.. in the legislation surrounding the **Freedom of Information..***
P2:	a fheacháint.. i reachtaíocht um **Shaoráil Faisnéise..** *to see.. in legislation surrounding the **Freedom of Information..***
P4:	féachaint.. sa reachtaíocht um **Shaoráil Faisnéise..** *see.. in the legislation surrounding the **Freedom of Information..***

Table 8: A portion of the PE output from P1, P2, P3 and P4. The EN data provided to the translators as the source text is also provided. The relevant MT output provided to translators is given above the participants output. A gloss for each sentence is indicated in italics below each GA output.

matic metrics can be useful in obtaining a rough idea of MT system quality, it does not always correlate with HE. Although in automatic metrics NMT was identified as the ‘better’ system and was the system translators deemed most accurate,[13] this did not consistently align with the scores generated using the PET output, nor the translators’ perceptions of fluency or the system which they would most prefer to post-edit.[14]

Overall, we can see that, even with just four participants, results can vary from translator to translator with both automatic metrics and those gathered as a direct result of PE. As a result, it is unreasonable to expect any one automatic metric to perfectly mirror HE.

Within this study, we have observed HTER as the metric which most closely matches our participants’ survey responses. However, it is important to note that with this study being limited to four participants we are unable to make definitive conclusions as to the best metric with which to guide EN→GA MT system development. As might be expected, the recommended approach would be to use HE wherever possible, and, in cases where this is not feasible, a combination of automatic metrics will provide the broadest snapshot of MT quality.

In terms of future work, we propose a similar study with more participants. We have seen that translators vary in MT PE approaches, experience and opinion. Accordingly, more participants would provide us with a more accurate picture of EN→GA MT quality and would provide us with a greater amount of data points to extrapolate from. We also suggest a more fine-grained evaluation of EN→GA MT output. In the survey portion of this study we elicit opinions of MT quality over 100-sentence documents in general. In the future it may be beneficial to examine specific differences between EN-GA SMT and NMT at the sentence-level, examining variations in errors in case, semantics, tense, etc.

On a final note, it is worth considering that with the derogation of EN→GA translation within the EU lifting in 2021, there is an urgent requirement for Irish language translation with too few translators available to satisfy the demand as well as a lack of Irish language resources. This means that we have a greater need than ever for EN-GA MT systems designed with the end-user in mind.

Acknowledgements The ADAPT SFI Centre for Digital Media Technology is funded by Science Foundation Ireland through the SFI Research Centres Programme and is co-funded under the European Regional Development Fund through Grant #13/RC/2106. We thank the reviewers for their useful comments, and the four participants for their invaluable contribution to this research; go raibh míle maith agaibh.

References

Arcan, Mihael, Caoilfhionn Lane, Eoin O Droighneáin, and Paul Buitelaar. 2016. IRIS: English-Irish machine translation system. In *The International Conference on Language Resources and Evaluation*, pages 566–572, Portoroz, Slovenia.

Aziz, Wilker, Sheila Castilho, and Lucia Specia. 2012. PET: a tool for post-editing and assessing machine translation. In *The International Conference on Language Resources and Evaluation*, pages 3982–3987, Istanbul, Turkey.

[13] Three of the four translators chose the NMT system as the most accurate output in the post-task survey, see Table 5.

[14] Two of the four chose NMT for both ’fluency’ and ’prefer’ in Table 5.

Barrault, Loïc, Ondřej Bojar, Marta R Costa-jussà, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, et al. 2019. Findings of the 2019 conference on machine translation (wmt19). In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 1–61, Florence, Italy.

Bayón, María Do Campo and Pilar Sánchez-Gijón. 2019. Evaluating machine translation in a low-resource language combination: Spanish-galician. In *Proceedings of Machine Translation Summit XVII Volume 2: Translator, Project and User Tracks*, pages 30–35, Dublin, Ireland.

Bojar, Ondřej, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Varvara Logacheva, Christof Monz, et al. 2016. Findings of the 2016 conference on machine translation. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 131–198, Berlin, Germany.

Castilho, Sheila, Joss Moorkens, Federico Gaspari, Rico Sennrich, Vilelmini Sosoni, Panayota Georgakopoulou, Pintu Lohar, Andy Way, Antonio Valerio Miceli Barone, and Maria Gialama. 2017. A comparative quality evaluation of pbsmt and nmt using professional translators. page 116–131.

Castilho, Sheila, Stephen Doherty, Federico Gaspari, and Joss Moorkens. 2018. Approaches to human and machine Translation Quality Assessment. In *Translation Quality Assessment: From Principles to Practice*, volume 1 of *Machine Translation: Technologies and Applications*, pages 9–38. Springer International Publishing, July.

Cheng, Yong. 2019. Joint training for pivot-based neural machine translation. In *Joint Training for Neural Machine Translation*, pages 41–54. Springer.

Defauw, Arne, Sara Szoc, Tom Vanallemeersch, Anna Bardadym, Joris Brabers, Frederic Everaert, Kim Scholte, Koen Van Winckel, and Joachim Van den Bogaert. 2019. Developing a neural machine translation system for Irish. In *Proceedings of the 2nd Workshop on Technologies for MT of Low Resource Languages*, pages 32–38, Dublin, Ireland.

Dowling, Meghan, Lauren Cassidy, Eimear Maguire, Teresa Lynn, Ankit Srivastava, and John Judge. 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"*, pages 314–318, Poznan, Poland.

Dowling, Meghan, Teresa Lynn, Yvette Graham, and John Judge. 2016. English to Irish machine translation with automatic post-editing. In *PARIS Inalco du 4 au 8 juillet 2016*, page 42, Paris, France.

Dowling, Meghan, Teresa Lynn, Alberto Poncelas, and Andy Way. 2018. SMT versus NMT: Preliminary comparisons for Irish. In *Association for Machine Translation in the Americas (AMTA)*, pages 12–20, Boston, USA.

Dowling, Meghan, Teresa Lynn, and Andy Way. 2019. Investigating backtranslation for the improvement of English-Irish machine translation. *TEANGA, the Journal of the Irish Association for Applied Linguistics*, 26:1–25.

Esplà-Gomis, Miquel, Mikel L Forcada, Gema Ramírez-Sánchez, and Hieu Hoang. 2019. Paracrawl: Web-scale parallel corpora for the languages of the eu. In *Proceedings of Machine Translation Summit XVII Volume 2: Translator, Project and User Tracks*, pages 118–119, Dublin, Ireland.

Etchegoyhen, Thierry, Borja Anza Porras, Andoni Azpeitia, Eva Martínez Garcia, Paulo Vale, José Luis Fonseca, Teresa Lynn, Jane Dunne, Federico Gaspari, Andy Way, et al. 2018. ELRI. European language resource infrastructure. page 351.

European Language Resource Coordination. 2020. *ELRC WHITE PAPER Sustainable Language Data Sharing to Support Language Equality in Multilingual Europe*. OVD Verlag Saarbrucken, Saarbrucken.

Heafield, Kenneth. 2011. Kenlm: Faster and smaller language model queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 187–197, Edinburgh, Scotland.

Imankulova, Aizhan, Raj Dabre, Atsushi Fujita, and Kenji Imamura. 2019. Exploiting out-of-domain parallel data through multilingual transfer learning for low-resource neural machine translation. In *Proceedings of Machine Translation Summit XVII Volume 1: Research Track*, pages 128–139, Dublin, Ireland.

Judge, John, Ailbhe Ní Chasaide, Rose Ní Dhubhda, Kevin P. Scannell, and Elaine Uí Dhonnchadha. 2012. *An Ghaeilge sa Ré Dhigiteach – The Irish Language in the Digital Age*. META-NET White Paper Series. Georg Rehm and Hans Uszkoreit (Series Editors). Springer. Available online at `http://www.meta-net.eu/whitepapers`.

Klein, Guillaume, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander M Rush. 2017. Opennmt: Open-source toolkit for neural machine translation. In *Proceedings of ACL 2017, System Demonstrations*, pages 67–72, Vancouver, Canada.

Koehn, Philipp and Rebecca Knowles. 2017. Six challenges for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 28–39, Vancouver, Canada.

Koehn, Philipp, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, et al. 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*, pages 177–180, Prague, Czech Republic.

Lakew, Surafel M, Mattia A Di Gangi, and Marcello Federico. 2017. Multilingual neural machine translation for low resource languages. In *Fourth Italian Conference on Computational Linguistics*, page 189, Rome, Italy.

Läubli, Samuel, Sheila Castilho, Graham Neubig, Rico Sennrich, Qinlan Shen, and Antonio Toral. 2020. A set of recommendations for assessing human–machine parity in language translation. *Journal of Artificial Intelligence Research*, 67:653–672.

Liu, Chao-Hong, Catarina Cruz Silva, Longyue Wang, and Andy Way. 2018. Pivot machine translation using chinese as pivot language. In *China Workshop on Machine Translation*, pages 74–85, Wuyishan, China.

Lynn, Teresa, Micheál Ó Conaire, and Jane Dunne, 2019. *Country Profile Ireland*, pages 92–97. German Research Center for Artificial Intelligence (DFKI), Saarbrucken, Germany.

Moorkens, Joss. 2020. Comparative satisfaction among freelance and directly-employed Irish-language translators. *The International Journal for Translation & Interpreting Research*, 12(1):55–73.

Murray, Kenton, Jeffery Kinnison, Toan Q Nguyen, Walter Scheirer, and David Chiang. 2019. Auto-sizing the transformer network: Improving speed, efficiency, and performance for low-resource machine translation. In *Proceedings of the 3rd Workshop on Neural Generation and Translation*, pages 231–240, Hong Kong.

Neubig, Graham, Zi-Yi Dou, Junjie Hu, Paul Michel, Danish Pruthi, and Xinyi Wang. 2019. compare-mt: A tool for holistic comparison of language generation systems. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 35–41, Minneapolis, Minnesota.

Ondřej, Bojar, Rajen Chatterjee, Federmann Christian, Graham Yvette, Haddow Barry, Huck Matthias, Koehn Philipp, Liu Qun, Logacheva Varvara, Monz Christof, et al. 2017. Findings of the 2017 conference on machine translation (wmt17). In *Second Conference onMachine Translation*, pages 169–214, Copenhagen, Denmark.

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318, Philadelphia, Pennsylvania.

Poncelas, Alberto, Dimitar Shterionov, Andy Way, Gideon Maillette de Buy Wenniger, and Peyman Passban. 2018. Investigating backtranslation in neural machine translation. In *Proceedings of the 21st Annual Conference of the European Association for Machine Translation*, pages 249–258, Alacant, Spain.

Popović, Maja. 2015. chrf: character n-gram f-score for automatic mt evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal.

Publications Office of the European Union. 2011. *Interinstitutional style guide*. Available online at `http://publications.europa.eu/resource/cellar/e774ea2a-ef84-4bf6-be92-c9ebebf91c1b.0018.03/DOC_2`.

Snover, Matthew, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the. Association for Machine Translation of the Americas*, pages 223–231, Cambridge, Massachusetts, USA.

Snover, Matthew G, Nitin Madnani, Bonnie Dorr, and Richard Schwartz. 2009. Ter-plus: paraphrase, semantic, and alignment enhancements to translation edit rate. *Machine Translation*, 23(2-3):117–127.

Toral, Antonio, Sheila Castilho, Ke Hu, and Andy Way. 2018. Attaining the unattainable? reassessing claims of human parity in neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 113–123, Brussels, Belgium.

Vanmassenhove, Eva, Dimitar Shterionov, and Andy Way. 2019. Lost in translation: Loss and decay of linguistic richness in machine translation. In *Proceedings of Machine Translation Summit XVII Volume 1: Research Track*, pages 222–232.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, California, USA.

Wang, Weiyue, Jan-Thorsten Peter, Hendrik Rosendahl, and Hermann Ney. 2016. Character: Translation edit rate on character level. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 505–510, Berlin, Germany.

Werlen, Lesly Miculicich, Dhananjay Ram, Nikolaos Pappas, and James Henderson. 2018. Document-level neural machine translation with hierarchical attention networks. *CoRR*, abs/1809.01576.

Wu, Hua and Haifeng Wang. 2007. Pivot language approach for phrase-based statistical machine translation. *Machine Translation*, 21(3):165–181.