Evaluating Domain Adaptation for Machine Translation Across Scenarios

Thierry Etchegoyhen[1], Anna Fernández Torné[2], Andoni Azpeitia[1], Eva Martínez Garcia[1] and Anna Matamala[2]

[1]Vicomtech, Donostia / San Sebastián, Spain [2]Universitat Autònoma de Barcelona, Barcelona, Spain
[1]{tetchegoyhen, aazpeitia, emartinez}@vicomtech.org [2]{Ana.Fernandez.Torne, Anna.Matamala}@uab.cat

Abstract

We present an evaluation of the benefits of domain adaptation for machine translation, on three separate domains and language pairs, with varying degrees of domain specificity and amounts of available training data. Domain-adapted statistical and neural machine translation systems are compared to each other and to generic online systems, thus providing an evaluation of the main options in terms of machine translation. Alongside automated translation metrics, we present experimental results involving professional translators, in terms of quality assessment, subjective evaluations of the task and post-editing productivity measurements. The results we present quantify the clear advantages of domain adaptation for machine translation, with marked impacts for domains with higher specificity. Additionally, the results of the experiments show domain-adapted neural machine translation systems to be the optimal choice overall.

Keywords: Machine Translation, Domain Adaptation, Quality Evaluation, Productivity Evaluation, Subjective Evaluation

1. Introduction

Statistical machine translation (SMT) (Brown et al., 1990) has been the dominant approach to automated translation for the last two decades, with neural machine translation (NMT) (Bahdanau et al., 2015) quickly becoming the new main paradigm in academic research and the industry, on the basis of the improvements it provides across the board (Toral and Sánchez-Cartagena, 2017). The data-driven nature of both approaches conditions the quality of their output to the availability of large volumes of adequate training resources for a given domain. However, domain-specific resources are usually scarce, thus making proper domain adaptation as much a challenge as it is a goal in developing accurate machine translation (MT) systems.

Domain adaptation has been extensively explored within SMT, with numerous studies focusing on the selection of supplementary data (Axelrod et al., 2011; Gascó et al., 2012; Eetemadi et al., 2015), translation model combination (Foster and Kuhn, 2007; Sennrich, 2012) or the integration of external information (Bisazza et al., 2011), to cite only a few. In NMT, domain adaptation is a more recent endeavour, with fine tuning currently the main method to gear generic translation networks towards specific domains (Luong and Manning, 2015; Freitag and Al-Onaizan, 2016; Crego et al., 2016).

Progress in machine translation technology has also given rise to large generic machine translation systems, many of which are freely available online. The increasing translation quality they provide, in part due to growing user feedback and training data covering multiple domains, has made them popular alternatives even in cases where domain-adapted systems might be better suited, although this specific aspect has not been fully evaluated yet. So far, large online translation systems have been essentially compared to academic systems for news-related translation (see, e.g., (Toral et al., 2011; Bojar et al., 2016)), rather than to systems tuned for the kind of specific domains that are more typical in the translation industry.

Machine translation quality and usefulness can be evaluated under various modalities. Quality can be measured via automated metrics such as BLEU (Papineni et al., 2002) or TER (Snover et al., 2006), among others. Additionally, or alternatively, direct assessments of translation quality can be made by professional translators or native speakers, and usefulness can be assessed via measurements of productivity gains and losses when post-editing machine-translated text. Over the years, human evaluations along these lines have shown the usefulness of machine translation in various scenarios (Plitt and Masselot, 2010; Pinnis et al., 2013; Etchegoyhen et al., 2014; Koehn and Germann, 2014). With the rise of neural machine translation, recent studies have also centred on comparing statistical and neural machine translation in different scenarios (Zoph et al., 2016; Castilho et al., 2017b).

In this paper, we focus on evaluating the benefits of domain adaptation in three distinct scenarios involving different domains and language pairs, with varying degrees of domain specificity and available in-domain resources. Domain-adapted statistical and neural machine translation systems are compared to each other and to generic online systems, thus providing an evaluation of the main options in terms of automated translations. Alongside automated translation metrics, we present experimental results involving professional translators in terms of quality assessment, subjective evaluations of the task and post-editing productivity measurements.

In the remainder of this paper, we describe the corpora and machine translation systems that were prepared, the design of the human quality and productivity evaluations, and the results in terms of automated metrics, human subjective assessments on a wide range of aspects, and objective analyses of post-editing results on the tasks carried by professional translators.

2. Domain Adaptation Scenarios

In order to evaluate different real-life scenarios for machine-translated content, we selected the three domains described below for our experiments:

DOMAIN	LANGS	TRAIN	DEV	TEST
MTOOL	ES-DE	25,256	1,984	3×50
ELEV	ES-FR	106,521	1,996	3×50
INTORG	EN-ES	23,138	1,998	3×50

Table 1: In-domain corpora statistics (number of parallel segments)

LANG	OOD	CORPUS								
		OPSUBS	UN	EUROP	JRC	NEWSCOM	CCRAWL	TED	WIKI	TOTAL
ES-DE	Generic	550,000	99,575	533,900	543,594	201,091	0	0	0	*1,784,385*
ES-FR	Generic	500,000	500,000	500,000	500,000	191,080	0	0	0	*2,191,079*
	WCrawl	0	0	0	0	0	0	0	13,177	*13,177*
EN-ES	Generic	499,000	499,000	551,000	293,586	206,137	499,000	156,895	0	*2,468,292*
	Pool	0	8,079,790	1,604,400	697,557	207,137	0	157,895	0	*10,410,392*

Table 2: Out-of-domain corpora statistics (number of parallel segments)

- MTOOL: Industrial documentation of machine tool components and processes.
- ELEV: Installation and maintenance documentation of elevators.
- INTORG: Reports and press releases of international non-profit organisms.

All three domains are representative of the various domains typically handled by translation services providers, each one being characterized by its own specialised vocabulary and constructions, which range from highly specific, as is the case for the MTOOL domain, to more general, as with the INTORG domain.

In addition to choosing markedly distinct domains, we selected different language pairs for each evaluation scenario: Spanish-German for MTOOL, Spanish-French for ELEV, and English-Spanish for INTORG.

The INTORG scenario is meant to evaluate domain adaptation in the least favourable case, i.e. where freely available training resources are abundant: the topics and language found in the texts of international organisms are rather close to those available in the United Nations and Europarl corpora (Eisele and Chen, 2010; Koehn, 2005), for instance; English-Spanish is also the language pair with the most abundant available parallel corpora, see, e.g., the resources in the OPUS repository (Tiedemann, 2012).

The other two scenarios and language pairs were chosen as representative of cases with strong demand in terms of internationalisation and relatively limited training resources, which represents a rather typical state of affairs in the translation services industry.

Finally, the three selected domains vary in terms of volumes of available corpora, both in-domain and related out-of-domain. These disparities in terms of amounts of training data are rather typical in the development of domain-adapted machine translation systems, with scarce resources for highly specific domains a particularly common scenario. Additionally, out-of-domain data that could complement scarce in-domain data might be limited in volume for highly specific domains. The task of domain adaptation is thus dependent on both the available in-domain data and the amounts of exploitable out-of-domain data for a given domain. The corpora collected for the domains at hand are described in more detail in the next section.

3. Corpora

In-domain data were provided in the form of translation memories for all three domains, with an additional collection of Spanish technical manuals previously translated into French for the ELEV domain. Since the documents provided in the latter case were unpaired, document alignment was performed using an in-house file name matcher, exploiting strong file naming consistency, and sentences were then aligned with Hunalign (Varga et al., 2005). As shown in Table 1, training data were particularly scarce for the MTOOL and INTORG domains, with at most 25 thousand unique parallel segments.

From the original data, we extracted around 2000 segments as development sets per domain, to serve as either tuning sets for SMT systems or validation sets for NMT systems. As test sets, for each domain we extracted 3 sets of 50 sentence pairs which had to be representative of the domain in terms of average sentence length and vocabulary, and be coherent in sequence, i.e., sampling was not performed randomly on a per sentence basis. These conditions were meant to allow for human quality and productivity evaluations that centred on realistic translation scenarios, as described in Section 5..

To complement the scarce in-domain datasets, we compiled the out-of-domain data described in Table 2.[1] Distinct freely available corpora were selected depending on the language pair and domain. The *Generic* datasets were prepared mainly to serve as basis for the NMT models, to be further fine-tuned with in-domain data for domain adaptation. For each corpus selected to compose the generic multi-domain corpus, parallel sentence pairs were first sorted by increasing perplexity scores according to language models trained on the entire monolingual sides of each parallel corpus, where the score was taken to be the arithmetic mean of source and target perplexities. Subsets of the ranked corpora were then selected to compose the

[1] Unless described otherwise, all corpora were downloaded from the OPUS website (*op. cit.*).

final corpus, with an upper selection bound taken to be either the median average perplexity score or the top n pairs if selecting up to median perplexity would result in over representing the corpus.

For ES-FR, we also included a small corpus, *WCrawl*, created from Wikipedia with an in-house crawler targeting domain terminology and the STACC$_w$ comparable sentence aligner (Azpeitia et al., 2017). Finally, for EN-ES, we prepared a data pool based on the concatenation of all corpora relevant to the domain of international news and regulations.

4. Models

As mentioned in Section 1., we aimed to evaluate the benefits of domain adaptation in various scenarios. With the aforementioned paradigm shift towards neural machine translation, it became necessary to further evaluate the potential differences between domain-adapted SMT and NMT machine translation systems. Thus, for each domain adaptation scenario described in Section 2. we trained two such domain-adapted systems. Statistical MT systems were standard phrase-based models built with the Moses toolkit (Koehn et al., 2007), with phrases of maximum length 5 and n-gram language models of order 5 built with KenLM (Heafield, 2011). Neural MT systems follow the attention-based encoder-decoder approach (Bahdanau et al., 2015) and were built with the OpenNMT toolkit (Klein et al., 2017). Generic translations were obtained from Google Translate in June 2017, where, to the best of our knowledge, ES-EN translations were generated by their NMT system and by their phrase-based SMT engines for the other two language pairs.

Several techniques are available to perform domain adaptation in SMT and we selected the method that gave the best results on the evaluation sets for each scenario.

For the INTORG domain, the optimal approach involved ranking the out-of-domain *Pool* dataset using the relative frequency ratio approach (RFR) of (Etchegoyhen et al., 2017) and selecting the best 1,000,000 sentence pairs as supplementary data. A phrase table was then created from the selected data and combined with the in-domain phrase table with the fill-up method of (Bisazza et al., 2011). In the ELEV domain, a similar approach was used, applying RFR ranking on the *Generic* dataset merged with the crawled data and selecting the best 98,845 sentence pairs, corresponding approximately to the size of the in-domain. Since manually revised domain-specific terms were available for this domain, we also included 162 phrasal term translations as favoured translation options using the XML-markup functionality in Moses, a domain adaptation technique readily available for SMT modelling. Finally, for the MTOOL domain we combined the phrases from the entire *Generic* dataset, via fill-up as well.

For all of our NMT models, domain adaptation was performed via fine-tuning (Luong and Manning, 2015; Freitag and Al-Onaizan, 2016; Crego et al., 2016), i.e. by further training the generic networks on the in-domain data.

5. Human Evaluation

A field quasi-experiment, for which no random assignment of participants to treatment groups was applied, was conducted with 15 professional translators for the INTORG and ELEV domains, and 22 for the MTOOL domain. Participants performed the assigned tasks in a real-world environment, thus favouring external validity.

Following an approved ethical procedure, the experiment consisted of a remunerated assignment and a volunteer, optional part. The evaluation aimed to compare the three machine translation systems in the three domains previously described, and was performed taking three different aspects into account: quality, post-editing (PE) productivity, and attitude. We describe each aspect in turn below.

5.1. Quality Assessment

Quality was first assessed at the segment level. The quality of the raw MT segments was assessed by scoring their fluency and adequacy on a scale from 1 to 4 using the TAUS DQF on-line tool.[2] Fluency conveyed to what extent the translated segment flowed naturally with no grammatical or spelling mistakes and was considered genuine language by native speakers (Koehn and Monz, 2006). In turn, adequacy assessed the amount of information of the source segment that was actually present in the target one (Koponen, 2010). Comparing the three different translated versions of each source segment was also considered a valuable quality indicator, so that a ranking task was also conducted.

Quality was also measured at the document level, by means of post-questionnaires where participants were asked to give their subjective global perception of the texts in terms of the aforementioned fluency and adequacy, as well as PE necessity, PE easiness and PE effort, as defined below:

- FLUENCY: the overall level of fluency of the machine-translated text.
- ADEQUACY: the overall level of adequacy of the machine-translated text.
- NECESSITY: the need for post-editing, i.e. whether the machine-translated text required many modifications overall.
- EASINESS: the easiness of the post-editing task, i.e. whether the necessary edits were technically simple overall.
- EFFORT: the mental effort required by the post-editing task, i.e. whether the necessary edits were cognitively difficult overall.

5.2. Post-editing Productivity Measurement

While conducting the post-editing task, the TAUS DQF tool automatically calculated the time to edit, i.e. "the average number of words processed by the post-editor in a given timespan" (SPEED from now onwards) and the post-editing effort (hereafter, WORK), namely "the average percentage of word changes applied by the post-editor on the MT output provided" (TAUS, n.d.). SPEED and WORK were considered two relevant indicators of post-editing productivity.

[2] https://dqf.taus.net/

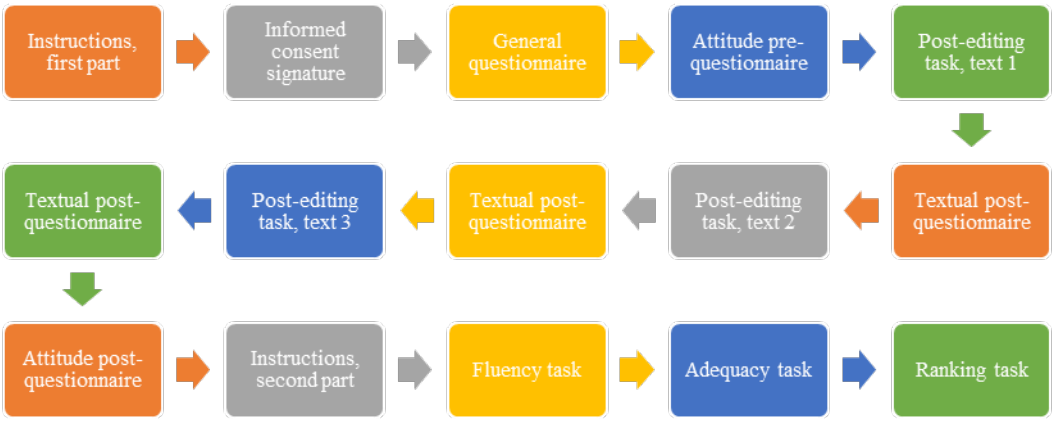


Figure 1: Experiment execution

DOMAIN	LANGS	MODEL	BLEU	METEOR	TER
MTOOL	ES-DE	SMT	19.830‡	35.260	69.378
		NMT	27.715† ∗	41.471	62.203
		GT	12.265	25.668	85.055
ELEV	ES-FR	SMT	62.524‡	74.627	25.550
		NMT	64.185†	76.062	23.100
		GT	18.857	37.955	63.050
INTORG	EN-ES	SMT	29.617	51.978	55.837
		NMT	32.726	54.467	50.620
		GT	33.024	55.068	50.646

Table 3: Results on automated machine translation metrics

5.3. Attitude Evaluation

Using scales from 1 to 10, participants were also asked to give their opinions on seven ratings in relation to their attitude towards MT and PE, namely the quality of raw machine-translated texts, the usefulness of MT for translators, their inclination to use MT as a text to depart from, their interest in PE, the boredom and the mental effort involved in PE, and the quality of post-edited machine-translated texts. These questions were presented twice, both before and after the PE task via pre- and post-questionnaires. This was aimed to see whether the actual post-editing task had in any way influenced their previous attitudes.

5.4. Experiment Execution

The experiment was divided in two parts to be performed at the participants' best convenience, as its expected length according to the pilot test was around 4 hours. Participants were informed via e-mail of the tasks to be carried out in each part. The instructions for the first part were to fill in a questionnaire on participant demographics and previous professional experience, and a pre-questionnaire on their attitude towards MT and PE.

Then, they were required to post-edit three texts. The order was balanced to minimise the fatigue and order-of-presentation effects, and was indicated individually to each participant on a separate email. After post-editing each text, they were requested to fill in the corresponding quality assessment post-questionnaire.

As a last step, they were asked to fill in a post-questionnaire on their attitude containing the same questions as the pre-questionnaire. Once they had finished the first part, they were asked to perform the fluency and adequacy evaluation tasks, and the rank comparison task (see Section 5.1.) for the same 150 segments.

The overall process is summarised in Figure 1.

5.5. Participants

Participants were selected following an a priori non-probabilistic purpose sampling (Bryman, 2012) based on subjects who met the following criteria: they had to be professional translators in the considered specific language pair and also native speakers of the target language.

A total of 52 participants took part in the experiment, distributed as follows: 11 female and 4 male EN-ES professional translators whose age ranged from 26 to 48 in the INTORG domain, 17 female and 5 male ES-DE professional translators ranging from 33 to 67 years old in the MTOOL domain, and 14 female and 1 male ES-FR professional translators whose ages ranged from 27 to 64 in the ELEV domain. Data from one participant in the MTOOL domain were not recorded due to technical problems. All participants but one in the MTOOL experiment had reached first cycle university studies.

6. Results

We first present results in terms of automated metrics, followed by a condensed representation of the human evaluation outcomes. A summary of all results is then added and discussed.

TASK	MEASURE	GT	NMT	SMT
SEGMENT-LEVEL QUALITY	ADEQUACY	3.56 ±0.70 † ‡	3.25 ±0.85 ∗	2.67 ±1.02
	FLUENCY	3.00 ±0.86 † ‡	2.75 ±0.97 ∗	2.00 ±0.93
	RANKING	1.39 ±0.61 † ‡	1.75 ±0.70 ∗	2.36 ±0.72
DOCUMENT-LEVEL QUALITY	FLUENCY	7.25 ±1.36 ‡	6.90 ±1.92 ∗	5.15 ±1.95
	ADEQUACY	8.67 ±0.78 †	7.90 ±0.94	7.54 ±1.56
	NECESSITY	5.83 ±2.08 ‡	7.00 ±1.90	7.92 ±1.18
	EASINESS	6.33 ±2.27	5.45 ±2.77	5.38 ±1.89
	EFFORT	6.90 ±2.31	7.64 ±2.62	7.38 ±1.04
PRODUCTIVITY	SPEED	1505 ±907.91 ‡	1236 ±560.60	957 ±375.83
	WORK	14.91 ±16.38 † ‡	19.76 ±16.91 ∗	27.28 ±18.56

Table 4: INTORG mean results and standard deviations for all human assessments

TASK	MEASURE	GT	NMT	SMT
SEGMENT-LEVEL QUALITY	ADEQUACY	2.55 ±0.96	3.25 ±0.86 † ∗	2.56 ±1.02
	FLUENCY	1.91 ±0.91	2.92 ±0.96 † ∗	1.91 ±1.06
	RANKING	2.04 ±0.72 ‡	1.49 ±0.72 † ∗	2.14 ±0.76
DOCUMENT-LEVEL QUALITY	FLUENCY	3.33 ±1.15	5.00 ±1.65 † ∗	3.31 ±2.17
	ADEQUACY	4.83 ±2.21	6.92 ±1.44 †	5.54 ±2.54
	NECESSITY	8.67 ±0.78	7.08 ±1.98 †	8.46 ±1.45
	EASINESS	4.92 ±2.97	5.25 ±2.67	4.62 ±3.07
	EFFORT	8.25 ±2.18	7.41 ±2.19	8.31 ±1.89
PRODUCTIVITY	SPEED	1018 ±500.40	1207 ±630.81	996 ±483.26
	WORK	37.56 ±21.51	20.49 ±20.82 † ∗	37.14 ±20.21 ‡

Table 5: MTOOL mean results and standard deviations for all human assessments

6.1. Automated metrics

We computed the performance of each model on the same test sets used for the human evaluations, in terms of BLEU (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005) and TER (Snover et al., 2006). For a closer comparison between automated and human evaluations, all machine-translated files were evaluated on cased detokenised output. Results are shown in Table 3.[3]

The first noticeable result is the strong benefit of domain adaptation for the MTOOL and ELEV domains, with large scoring differences on all metrics using either an SMT or an NMT domain-adapted system over generic GT engines. The only scenario where this result is not confirmed involves the INTORG domain. In this case, as previously described, the domain is the least restricted of the three, covering world news and events for which large amounts of training data are freely available. This wide scope domain demonstrates the convergence of various systems when in-domain data is not a marked provider of the most relevant information. Results from this domain also show the competitive scores achievable with comparatively small amounts of training data when compared to generic engines trained on significantly larger amounts of data. Overall, domain adaptation appears to be a necessary step to optimise translation quality, despite recent progress in the development of large generic natural machine translation systems.

A second important result is the effectiveness of current neural machine translation for narrow domains. Recent work had shown the need of large amounts of training data for NMT modelling, showing that SMT performed comparatively better in low resource scenarios (Zoph et al., 2016). Our experiments feature two narrow domains, with low amounts of parallel training data and high domain specificity, for which fine-tuned NMT models achieved the best results.[4] This outcome was obtained using simple fine tuning over generic models, an approach which has some inherent limitations such as need to restrict the adapted models to the vocabulary of the existing generic network. Domain modelling is thus limited in this approach, with domain-specific vocabulary handled via additional mechanisms such as unknown source word copies. Improved methods of domain adaptation for NMT are thus likely to provide gains to an already strong baseline for narrow domains.

Finally, in terms of automated metrics, SMT performed well in two out of three domains, reaching statistically comparable results to the ones obtained with domain-adapted

[3] Statistical significance was computed for the BLEU metric on the merged files for each domain via bootstrap resampling (Koehn, 2004). † indicates statistical significance, at $p < 0.05$, between NMT and GT; ‡ between SMT and GT; and ∗ between NMT and SMT.

[4] Note that, for the INTORG and ELEV domains, our NMT models were trained on more data than their SMT counterparts, the latter being built following a standard set-up where in-domain and out-of-domain data are not merged and only a portion of the out-of-domain data is selected. Additional experiments not reported here showed that using the entire out-of-domain dataset for SMT did not provide significant improvements over the approach reported here.

TASK	MEASURE	GT	NMT	SMT
SEGMENT-LEVEL QUALITY	ADEQUACY	2.40 ±0.96	3.56 ±0.68 †	3.43 ±0.77 ‡
	FLUENCY	1.99 ±1.06	3.06 ±0.90 †	2.91 ±0.98 ‡
	RANKING	2.33 ±0.68	1.47 ±0.63 †	1.47 ±0.68 ‡
DOCUMENT-LEVEL QUALITY	FLUENCY	2.90 ±1.83	7.40 ±1.74 †	7.10 ±1.83 ‡
	ADEQUACY	5.10 ±2.42	8.10 ±0.93 †	8.10 ±0.93 ‡
	NECESSITY	8.60 ±1.88	5.40 ±2.60 †	5.60 ±2.74 ‡
	EASINESS	5.20 ±2.49	7.00 ±2.12	6.10 ±1.62
	EFFORT	8.10 ±1.27	7.10 ±2.47	5.80 ±2.22 ‡
PRODUCTIVITY	SPEED	881 ±294.85	1462 ±582.93 †	1477 ±485.63 ‡
	WORK	39.44 ±21.02	11.12 ±13.56 †	10.72 ±13.54 ‡

Table 6: ELEV mean results and standard deviations for all human assessments

NMT, although with absolute scores consistently below those achieved with NMT. It is worth noting that the SMT results were obtained with domain adaptation techniques that have been extensively researched and time-tested over the years. Thus, it is unlikely that different domain adaptation methods for statistical machine translation would provide significant gains overall, which in turn places domain-adapted NMT as the currently optimal approach in terms of automated metrics.

6.2. Human evaluation

In this section, we first present the results in terms of metrics, both objective and subjective, and then summarise the eventual changes in perception of MT and PE for the translators who participated in the evaluation.

6.2.1. Metrics

A statistical analysis was carried out using IBM SPSS V. 20, setting the significance level at 0.05. For qualitative data such as adequacy, fluency and ranking at segment level, chi-square tests were used to compare the distribution of the assessments by language pair. Data relating to textual quality were considered discrete numerical variables, so that a Mann-Whitney U test was used for the comparison of groups. For the continuous numerical variables PE speed and work, normality of distributions was assessed by using the Shapiro-Wilk test. As data were not normally distributed, the Bonferroni-corrected Mann-Whitney U test was used for multiple comparisons.

Intra-class correlation coefficient (ICC) estimates and their 95% confidence intervals were used as inter-rater reliability indexes, which for all three domains resulted in excellent levels of reliability (all above 0.92) for the quality assessment variables at the segment level.

In the INTORG domain, the means of all aspects assessed in the case of GT indicated better results than those obtained with either NMT or SMT. Thus, the means were lower in ranking, necessity, effort and work, and higher in the other categories where higher scores indicate better results. In turn, all NMT means but one (effort) showed better results than those of SMT, as shown in Table 4.

Examining post-editing edit distances, shown in Figure 2(a), confirms the ranking of the systems in this domain. Thus, in the case of GT more than 37% of the segments had an edit distance of 0 and 22% an edit distance of 1. For these same two distances, SMT featured 11.5% and 18.8% of the segments, respectively, whereas for NMT the proportions were 19.9% and 24.3%, respectively. Additionally, GT exhibits a gradual reduction in the percentage of segments as the number of edits increases, whereas for NMT the number of segments increases between distances of 0 and 1, and SMT has a higher percentage of segments with an edit distance of 2, beyond 20%.

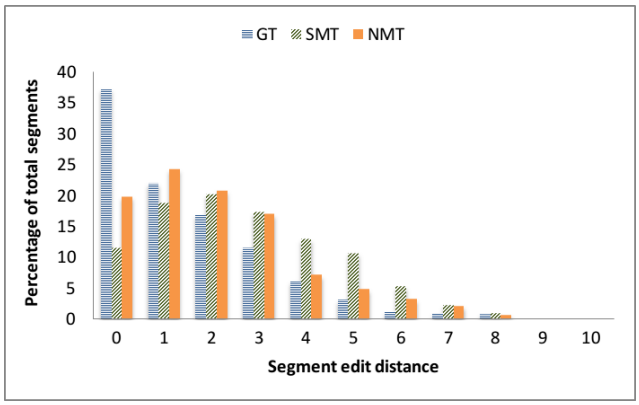
The conclusions are different in the MTOOL domain. As shown in Table 5, the leading position for all means in this case was for NMT, with GT and SMT presenting very close means in all items assessed. The distribution of results in terms of edit distances, shown in Figure 2(b), illustrates the differences between systems in this narrow domain. The first noticeable result is the extremely large difference for edit distances of 0, with 34.42% for NMT as opposed to 9.56% for GT and 11.43% for SMT. The NMT system cumulates more than 60% of the segments in the lowest edit distances, from 0 to 2, as opposed to 27% for GT and 25% for SMT. In terms of human evaluation, domain-adapted NMT was thus the optimal system in the MTOOL domain, with GT and ST showing comparable results.

In the ELEV domain, NMT had the lead again, as shown in Table 6. However, in this case SMT means were closer to NMT's means, which left GT as the worst performing system. This ranking of the systems is again illustrated by the distribution of edits shown in Figure 2(c). For both SMT and NMT, the dominant edit distance was 0, with 45.9% and 46.5%, respectively, whereas for GT the most frequent edit distances were 5 and 6, with 17.7% and 18.5%, respectively. In this domain, which exhibits a comparable domain-specificity to MTOOL but larger amounts of in-domain data, domain adaptation with NMT appears to be the optimal option, but the SMT system shows a comparable strong performance, with only the generic GT system performing markedly worse on all metrics.

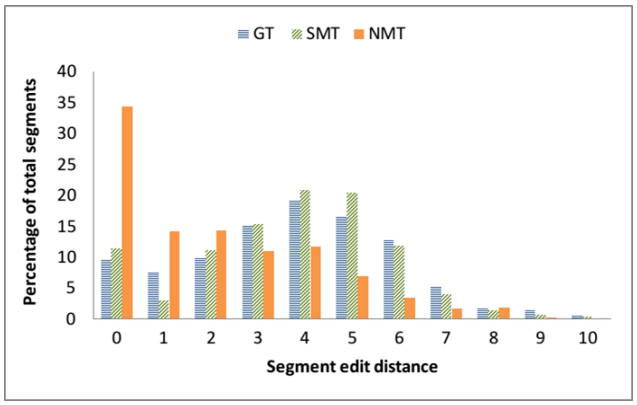
6.2.2. Attitude

We now summarise the results regarding the translators' changes in attitude before and after completing the tasks.

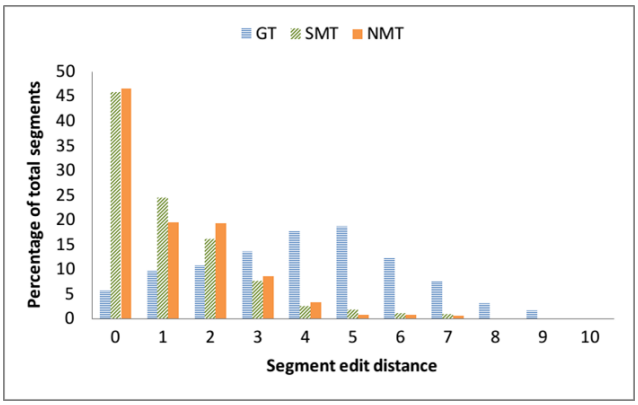
For INTORG, there was a positive evolution regarding aspects such as quality and utility of machine translation, as well as inclination towards using machine-translated texts as a starting point for translation. On the other hand, interest in post-editing lowered slightly, while boredom and



(a) Edit distances in the INTORG domain



(b) Edit distances in the MTOOL domain



(c) Edit distances in the ELEV domain

Figure 2: Edit distances per domain

MEASURE	INTORG	MTOOL	ELEV
ADEQUACY	r = 0.10, $p > 0.05$	r = -0.09, $p > 0.05$	r = -0.08, $p > 0.05$
FLUENCY	r = 0.36, $p < 0.001$	r = 0.48, $p < 0.001$	r = 0.61, $p < 0.001$
RANKING	r = -0.20, $p < 0.001$	r = -0.30, $p < 0.001$	r = -0.68, $p < 0.001$
SPEED	r = 0.04, $p > 0.05$	r = 0.05, $p > 0.05$	r = -0.26, $p < 0.001$
WORK	r = -0.51, $p < 0.001$	r = -0.50, $p < 0.001$	r = -0.69, $p < 0.001$

Table 7: Correlations between BLEU and human assessments

TASK	MEASURE	DOMAIN								
		ELEV			MTOOL			INTORG		
		GT	NMT	SMT	GT	NMT	SMT	GT	NMT	SMT
SEGMENT-LEVEL QUALITY	ADEQUACY	3 †	1 ‡	2 ∗	3 †	1	2 ∗	1 †	2 ‡	3 ∗
	FLUENCY	3 †	1 ‡	2 ∗	2 †	1 ‡	3 ∗	1 †	2 ‡	3 ∗
	RANKING	3 †	1 ‡	1 ∗	2 †	1 ‡	3 ∗	1 †	2 ‡	3 ∗
DOCUMENT-LEVEL QUALITY	FLUENCY	3 †	1 ‡	2 ∗	2 †	1	3 ∗	1 †	2 ‡	3 ∗
	ADEQUACY	3 †	1 ‡	1	3 †	1	2	1 †	2	3
	NECESSITY	3 †	1 ‡	2	3 †	1	2	1	2 ‡	3
	EASINESS	3	1	1	2	1	3	1	2	3
	EFFORT	3	2 ‡	1	2	1	3	1	3	2
PRODUCTIVITY	SPEED	3 †	2 ‡	1	2	1	3	1	2 ‡	3
	WORK	3 †	2 ‡	1	3 †	1	2 ∗	1 †	2 ‡	3 ∗
AUTOMATED METRICS	BLEU	3 †	1 ‡	2 ∗	3 †	1	2 ∗	1	2	3
	METEOR	3	1	2	3	1	2	1	2	3
	TER	3	1	2	3	1	2	1	2	3

Table 8: Summary of comparative results

the perception of the mental effort needed for post-editing increased. Inferential statistical analysis showed that there were statistically significant changes, with a p-value less than 0.05, only in the case of perception of machine translation quality and boredom.

In the MTOOL domain, the attitudes evolved negatively in all aspects except usefulness of machine translation and interest in post-editing. Perception of the quality of machine translation lowered, as did the interest in using machine-translated texts as input for translation. Results were statistically significant only in the case of perception of cognitive effort, which increased after completion of the task.

Finally, for the ELEV domain all positive attitude indicators increased, except for quality of post-edited texts, which maintained the same score, and boredom, which also increased. Thus, after completion of the task, the perception of quality of machine translation increased, as did its perceived usefulness and the interest in using it as input. The required mental effort was also perceived as lower after completion of the task. The changes were not statistically significant in any of the aspects, though, with p-values above 0.05.

Although attitude changes were not statistically significant in most cases, they are in line with the results on the previously discussed metrics, with an overall increase in positive perception of the post-editing task in the domain with markedly better translations, namely ELEV, and mixed results for the other two domains.

6.3. Summary

As shown in Table 7, there are relevant correlations between segment-level BLEU and the human assessments for fluency, ranking and WORK, particularly in the ELEV domain. Thus, the higher the fluency and ranking results and the lower the WORK, the higher the BLEU metric obtained. Although BLEU usually shows higher correlations with human judgements at the document level than at the segment level, in these experiments the correlations were significant.

Results in terms of both automated metrics and human assessments show an almost perfect match, as seen in Table 8, which includes the position each engine occupies taking into account every aspect assessed and the statistical significance of the differences observed between each pair of systems. Thus, in the INTORG domain GT is almost unanimously considered the best system by all items assessed, while SMT is deemed the worst performing one. Likewise, NMT is the system obtaining the best results in the MTOOL scenario, while GT and SMT vie for the last position. In the ELEV domain, NMT obtains again the best results, although closely followed by SMT, which clearly leaves GT as the worst classified system.

Considering the different domains selected for the experiments, their specificity as well as the amount of available in-domain data, the results were not unexpected but the experiments performed provide a quantified view of the impact of domain adaptation. Thus, for the two domains that were more specific, domain-adapted systems in one form or another provided clear advantages that are reflected in all metrics, automated, based on human subjective evaluation, or based on objective post-editing metrics.

Overall, adapting neural machine translation systems to a specific domain proved the optimal approach, performing better where domain-specificity was higher, and competing with a large state-of-the-art generic translation system while being trained on only a relatively small amount of data overall. This result shows the progress of NMT in general, as it performed better than statistical machine translation systems even in the case of highly specific domains. Note also that the NMT systems performed better overall than the SMT ones in terms of adequacy as well, in contrast with the results described in (Castilho et al., 2017a), where neural models performed better than statistical ones in terms of fluency, but not always in terms of adequacy.

It is worth noting also that, as shown by the ELEV domain, SMT systems could remain competitive in domain adaptation scenarios, although it is likely that future more sophisticated domain-adaptation methods for NMT will likely extend the gap between the approaches.

7. Conclusions

We have described the evaluation of the benefits of domain adaptation for machine translation under different scenarios that involve unrelated domains and language pairs, with varying degrees of domain specificity and amounts of training data. Our protocols include domain-adapted statistical and neural machine translation systems, as well as a large generic online system, thus addressing the main options currently available in terms of automated translation.

The human evaluation, which involved professional translators, covered quality assessments, post-editing productivity measurements, as well as attitude evaluations. An in-depth statistical analysis of the results was provided, along with an evaluation of eventual changes in perception of the task by the participants.

Overall, results in terms of both automated metrics and human assessments show the benefits of domain adaptation, with marked gains across the board for domain-adapted systems on all metrics for the more specific domains. Although not unexpected, given that domain-specific knowledge could be expected to positively impact data-driven translation systems, these results have been substantiated on a wide range of aspects.

Finally, the reported experiments show the comparative effectiveness of domain-adapted neural machine translation across the board, confirming that the paradigm shift that has taken place in the field towards neural machine translation can be considered adequate as well for the highly-specific scenarios that are common in the translation industry.

8. Acknowledgements

This work was partially funded by the Spanish Ministry of Economy and Competitiveness via the AdapTA project (RTC-2015-3627-7). We would like to thank MondragonLingua Translation & Communication as coordinator of the project, the anonymous LREC reviewers for their helpful feedback and suggestions, and the translators who participated in the experiments. Anna Matamala is a member of Transmedia Catalonia, a research group funded by the Catalan government under SGR call (2017SGR113).

9. Bibliographical References

Axelrod, A., He, X., and Gao, J. (2011). Domain adaptation via pseudo in-domain data selection. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 355–362. Association for Computational Linguistics.

Azpeitia, A., Etchegoyhen, T., and Martinez García, E. (2017). Weighted set-theoretic alignment of comparable sentences. In *Proceedings of the 10th Workshop on Building and Using Comparable Corpora*, pages 41–45.

Bahdanau, D., Cho, K., and Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. In *Proceedings of the International Conference on Learning Representations*.

Banerjee, S. and Lavie, A. (2005). METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization, Ann Arbor, Michigan*, volume 29, pages 65–72.

Bisazza, A., Ruiz, N., Federico, M., and Kessler, F.-F. B. (2011). Fill-up versus interpolation methods for phrase-based smt adaptation. In *Proceedings of the International Workshop on Spoken Language Translation*, pages 136–143.

Bojar, O., Chatterjee, R., Federmann, C., Graham, Y., Haddow, B., Huck, M., Yepes, A. J., Koehn, P., Logacheva, V., Monz, C., 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*, volume 2, pages 131–198.

Brown, P. F., Cocke, J., Pietra, S. A. D., Pietra, V. J. D., Jelinek, F., Lafferty, J. D., Mercer, R. L., and Roossin, P. S. (1990). A statistical approach to machine translation. *Computational linguistics*, 16(2):79–85.

Bryman, A. (2012). *Social Research Methods*. Oxford University Press, Oxford.

Castilho, S., Moorkens, J., Gaspari, F., Calixto, I., Tinsley, J., and Way, A. (2017a). Is neural machine translation the new state of the art? *The Prague Bulletin of Mathematical Linguistics*, 108(1):109–120.

Castilho, S., Moorkens, J., Gaspari, F., Sennrich, R., Sosoni, V., Georgakopoulou, Y., Lohar, P., Way, A., Miceli Barone, A. V., and Gialama, M. (2017b). A Comparative Quality Evaluation of PBSMT and NMT using Professional Translators. In *Proceedings of Machine Translation Summit XVI*, Nagoya, Japan.

Crego, J., Kim, J., Klein, G., Rebollo, A., Yang, K., Senellart, J., Akhanov, E., Brunelle, P., Coquard, A., Deng, Y., et al. (2016). Systran's pure neural machine translation systems. *arXiv preprint arXiv:1610.05540*.

Eetemadi, S., Lewis, W., Toutanova, K., and Radha, H. (2015). Survey of data-selection methods in statistical machine translation. *Machine Translation*, 29(3-4):189–223.

Eisele, A. and Chen, Y. (2010). MultiUN: A Multilingual Corpus from United Nation Documents. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation*, pages 2868–2872. European Language Resources Association (ELRA), 5.

Etchegoyhen, T., Bywood, L., Fishel, M., Georgakopoulou, P., Jiang, J., Van Loenhout, G., Del Pozo, A., Maucec, M. S., Turner, A., and Volk, M. (2014). Machine translation for subtitling: A large-scale evaluation. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2014)*, pages 46–53.

Etchegoyhen, T., Azpeitia, A., and Martinez García, E. (2017). Exploiting Relative Frequencies for Data Selection. In *Proceedings of the 16th Machine Translation Summit*, volume 1: Long Papers, pages 170–184.

Foster, G. and Kuhn, R. (2007). Mixture-model adaptation for SMT. In *Proceedings of the Second Workshop on Statistical Machine Translation*, StatMT '07, pages 128–135, Stroudsburg, PA, USA. Association for Computational Linguistics.

Freitag, M. and Al-Onaizan, Y. (2016). Fast domain adaptation for neural machine translation. *arXiv preprint arXiv:1612.06897*.

Gascó, G., Rocha, M.-A., Sanchis-Trilles, G., Andrés-Ferrer, J., and Casacuberta, F. (2012). Does more data always yield better translations? In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '12.

Heafield, K. (2011). Kenlm: Faster and smaller language model queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, WMT '11, pages 187–197, Stroudsburg, PA, USA. Association for Computational Linguistics.

Klein, G., Kim, Y., Deng, Y., Senellart, J., and Rush, A. M. (2017). Opennmt: Open-source toolkit for neural machine translation. In *Proceedings of the 20th Annual Conference of the European Association for Machine Translation*.

Koehn, P. and Germann, U. (2014). The impact of machine translation quality on human post-editing. In *Proceedingds of the Workshop on Humans and Computer-Assisted Translation*, pages 38–46.

Koehn, P. and Monz, C. (2006). Manual and automatic evaluation of machine translation between european languages. In *Proceedings of the Workshop on Statistical Machine Translation*, pages 102–121. Association for Computational Linguistics.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., et al. (2007). Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the ACL*, pages 177–180. Association for Computational Linguistics.

Koehn, P. (2004). Statistical significance tests for machine translation evaluation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 388–395.

Koehn, P. (2005). Europarl: A Parallel Corpus for Statistical Machine Translation. In *Proceedings of the 10th Machine Translation Summit*, pages 79–86.

Koponen, M. (2010). Assessing machine translation quality with error analysis. In *Electronic proceedings of the*

KäTu symposium on translation and interpreting studies, volume 4, pages 1–12.

Luong, M.-T. and Manning, C. D. (2015). Stanford neural machine translation systems for spoken language domains. In *Proceedings of the International Workshop on Spoken Language Translation*.

Papineni, K., Roukos, S., Ward, T., and 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*, pages 311–318. Association for Computational Linguistics.

Pinnis, M., Skadiņa, I., and Vasiļjevs, A. (2013). Domain adaptation in statistical machine translation using comparable corpora: case study for english latvian it localisation. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 224–235. Springer.

Plitt, M. and Masselot, F. (2010). A productivity test of statistical machine translation post-editing in a typical localisation context. *The Prague bulletin of mathematical linguistics*, 93:7–16.

Sennrich, R. (2012). Perplexity minimization for translation model domain adaptation in statistical machine translation. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 539–549. Association for Computational Linguistics.

Snover, M., Dorr, B., Schwartz, R., Micciulla, L., and Makhoul, J. (2006). A study of translation edit rate with targeted human annotation. In *Proceedings of Association for Machine Translation in the Americas*, pages 223–231.

Tiedemann, J. (2012). Parallel data, tools and interfaces in OPUS. In *Proceedings of the 8th Language Resources and Evaluation Conference*, pages 2214–2218.

Toral, A. and Sánchez-Cartagena, V. M. (2017). A multifaceted evaluation of neural versus phrase-based machine translation for 9 language directions. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 1063–1073, Valencia, Spain, April. Association for Computational Linguistics.

Toral, A., Gaspari, F., Kumar Naskar, S., and Way, A. (2011). Comparative evaluation of research vs. online MT systems. In *Proceedings of the 15th Conference of the European Association for Machine Translation*.

Varga, D., Németh, L., Halácsy, P., Kornai, A., Trón, V., and Nagy, V. (2005). Parallel corpora for medium density languages. In *Recent Advances in Natural Language Processing (RANLP 2005)*, pages 590–596.

Zoph, B., Yuret, D., May, J., and Knight, K. (2016). Transfer learning for low-resource neural machine translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1568–1575.