# Translating Literary Text between Related Languages using SMT

**Antonio Toral**
ADAPT Centre
School of Computing
Dublin City University
Dublin, Ireland
atoral@computing.dcu.ie

**Andy Way**
ADAPT Centre
School of Computing
Dublin City University
Dublin, Ireland
away@computing.dcu.ie

## Abstract

We explore the feasibility of applying machine translation (MT) to the translation of literary texts. To that end, we measure the translatability of literary texts by analysing parallel corpora and measuring the degree of freedom of the translations and the narrowness of the domain. We then explore the use of domain adaptation to translate a novel between two related languages, Spanish and Catalan. This is the first time that specific MT systems are built to translate novels. Our best system outperforms a strong baseline by 4.61 absolute points (9.38% relative) in terms of BLEU and is corroborated by other automatic evaluation metrics. We provide evidence that MT can be useful to assist with the translation of novels between closely-related languages, namely (i) the translations produced by our best system are equal to the ones produced by a professional human translator in almost 20% of cases with an additional 10% requiring at most 5 character edits, and (ii) a complementary human evaluation shows that over 60% of the translations are perceived to be of the same (or even higher) quality by native speakers.

## 1 Introduction

The field of Machine Translation (MT) has evolved very rapidly since the emergence of statistical approaches almost three decades ago (Brown et al., 1988; Brown et al., 1990). MT is nowadays a growing reality throughout the industry, which continues to adopt this technology as it results in demonstrable improvements in translation productivity, at least for technical domains (Zhechev, 2012). Meanwhile, the performance of MT systems in research continues to improve. In this regard, a recent study looked at the best-performing systems of the WMT shared task for seven language pairs during the period between 2007 and 2012, and estimated the improvement in translation quality during this period to be around 10% absolute, in terms of both adequacy and fluency (Graham et al., 2014).

Having reached this level of research maturity and industrial adoption, in this paper we explore the feasibility of applying the current state-of-the-art MT technology to literary texts, what might be considered to be the last bastion of human translation. The perceived wisdom is that MT is of no use for the translation of literature. We challenge that view, despite the fact that – to the best of our knowledge – the applicability of MT to literature has to date been only partially studied from an empirical point of view.

In this paper we aim to measure the translatability of literary text. Our empirical methodology relies on the fact that the applicability of MT to a given type of text can be assessed by analysing parallel corpora of that particular type and measuring (i) the degree of freedom of the translations (how literal the translations are), and (ii) the narrowness of the domain (how specific or general that text is). Hence, we tackle the problem of measuring the translatability of literary text by comparing the degree of freedom of translation and domain narrowness for such texts to documents in two other domains which have been widely studied in the area of MT: technical documentation and news.



*Proceedings of NAACL-HLT Fourth Workshop on Computational Linguistics for Literature*, pages 123–132,
Denver, Colorado, June 4, 2015. 

Furthermore, we assess the usefulness of MT in translating a novel between two closely-related languages. We build an MT system using state-of-the-art domain-adaptation techniques and evaluate its performance against the professional human translation, using both automatic metrics and manual evaluation. To the best of our knowledge, this is the first time that a specific MT system is built to translate novels.

The rest of the paper is organised as follows. Section 2 gives an overview of the current state-of-the-art in applying MT to literary texts. In Section 3 we measure the translatability of literary texts. In Section 4 we explore the use of MT to translate a novel between two related languages. Finally, in Section 5 we present our conclusions and outline avenues of future work.

## 2 Background

To date, there have been only a few works on applying MT to literature, for which we provide an overview here.

Genzel et al. (2010) explored constraining statistical MT (SMT) systems for poetry to produce translations that obey particular length, meter and rhyming rules. Form is preserved at the price of producing a worse translation, in terms of the BLEU automatic metric, which decreases from 0.3533 to 0.1728, a drop of around 50% in real terms. Their system was trained and evaluated with WMT-09 data[1] for French–English.

Greene et al. (2010) also translated poetry, choosing target realisations that conform to the desired rhythmic patterns. Specifically, they translated Dante's *Divine Comedy* from Italian sonnets into English iambic pentameter. Instead of constraining the SMT system, they passed its output lattice through a FST that maps words to sequences of stressed and unstressed syllables. These sequences are finally filtered with a iambic pentameter acceptor. Their output translations are evaluated qualitatively only.

Voigt and Jurafsky (2012) examined how referential cohesion is expressed in literary and non-literary texts, and how this cohesion affects translation. They found that literary texts have more dense reference chains and conclude that incorporating discourse features beyond the level of the sentence is an important direction for applying MT to literary texts.

Jones and Irvine (2013) used existing MT systems to translate samples of French literature (prose and poetry) into English. They then used qualitative analysis grounded in translation theory on the MT output to assess the potential of MT in literary translation and to address what makes literary translation particularly difficult, e.g. one objective in literary translation, in contrast to other domains, is to preserve the *experience* of reading a text when moving to the target language.

Very recently, Besacier (2014) presented a pilot study where MT followed by post-editing is applied to translate a short story from English into French. In Besacier's work, post-editing is performed by non-professional translators, and the author concludes that such a workflow can be a useful low-cost alternative for translating literary works, albeit at the expense of sacrificing translation quality. According to the opinion of a professional translator, the main errors had to do with using English syntactic structures and expressions instead of their French equivalents and not taking into account certain cultural references.

Finally, there are some works that use MT techniques in literary text, but for generation rather than for translation. He et al. (2012) used SMT to generate poems in Chinese given a set of keywords. Jiang and Zhou (2008) used SMT to generate the second line of Chinese couplets given the first line. In a similar fashion, Wu et al. (2013) used transduction grammars to generate rhyming responses in hip-hop given the original challenges.

This paper contributes to the current state-of-the-art in two dimensions. On the one hand, we conduct a comparative analysis on the translatability of literary text according to narrowness of the domain and freedom of translation. This can be seen as a more general and complementary analysis to the one conducted by Voigt and Jurafsky (2012). On the other hand, and related to Besacier (2014), we evaluate MT output for literary text. There are two differences though; first, they translated a short story, while we do so for a longer type of literary

[1] http://www.statmt.org/wmt09/translation-task.html

text, namely a novel; second, their MT systems were evaluated against a post-edited reference produced by non-professional translators, while we evaluate our systems against the translation produced by a professional translator.

## 3 Translatability

The applicability of SMT to translate a certain text type for a given pair of languages can be studied by analysing two properties of the relevant parallel data.

- Degree of freedom of the translation. While literal translations can be learnt reasonably well by the word alignment component of SMT, free translations may result in problematic alignments.
- Narrowness of the domain. Constrained domains lead to good SMT results. This is due to the fact that in narrow domains lexical selection is much less of an issue and relevant terms occur frequently, which allows the SMT model to learn their translations with good accuracy.

We could say then, that the narrower the domain and the smaller the degree of freedom of the translation, the more applicable SMT is. This is, we assert, why SMT performs well on technical documentation while results are substantially worse for more open and unpredictable domains such as news (cf. WMT translation task series).[2]

We propose to study the applicability of SMT to literary text by comparing the degree of freedom and narrowness of parallel corpora for literature to other domains widely studied in the area of MT (technical documentation and news). Such a corpus study can be carried out by using a set of automatic measures. The perplexity of the word alignment can be used as a proxy to measure the degree of freedom of the translation. The narrowness of the domain can be assessed by measuring perplexity with respect to a language model (LM) (Ruiz and Federico, 2014).

Therefore, in order to assess the translatability of literary text with MT, we contextualise the problem by comparing it to the translatability of other widely studied types of text. Instead of considering the translatability of literature as a whole, we root the study along two axes:

- Relatedness of the language pair: from pairs of languages that belong to the same family (e.g. Romance languages), through languages that belong to the same group (e.g. Romance and Germanic languages of the Indo-European group) to unrelated languages (e.g. Romance and Finno-Ugric languages).
- Literary genre: novels, poetry, etc.

We hypothesise that the degree of applicability of SMT to literature depends on these two axes. Between related languages, translations should be more literal and complex phenomena (e.g. metaphors) might simply transfer to the target language, while they are more likely to require complex translations between unrelated languages. Regarding literary genres, in poetry the preservation of form might be considered relevant while in novels it may be a lesser constraint.

The following sections detail the experimental datasets and the experiments conducted regarding narrowness of the domain and degree of translation freedom.

### 3.1 Experimental Setup

In order to carry out our experiment on the translatability of literary texts, we use monolingual datasets for Spanish and parallel datasets for two language pairs with varying levels of relatedness: Spanish–Catalan and Spanish–English.

Regarding the different types of corpora, we consider datasets that fall in the following four groups: novels, news, technical documentation and Europarl (EP).

We use two sources for novels: two novels from Carlos Ruiz Zafón, *The Shadow of the Wind* (published originally in Spanish in 2001) and *The Angel's Game* (2008), for Spanish–Catalan and Spanish–English, referred to as novel1; and two novels from Gabriel García Márquez, *Hundred Years of Solitude* (1967) and *Love in the Time of Cholera* (1985), for Spanish–English, referred to as novel2.

We use two sources of news data: a corpus made of articles from the newspaper *El Periódico*[3] (re-

[2] http://www.statmt.org/wmt14/translation-task.html

[3] http://www.elperiodico.com/

ferred to as news1) for Spanish–Catalan, and news-commentary v8[4] (referred to as news2) for Spanish–English.

For technical documentation we use four datasets: DOGC,[5] a corpus from the official journal of the Catalan Goverment, for Spanish–Catalan; EMEA,[6] a corpus from the European Medicines Agency, for Spanish–English; JRC-Acquis (henceforth referred as JRC) (Steinberger et al., 2006), made of legislative text of the European Union, for Spanish–English; and KDE4,[7] a corpus of localisation files of the KDE desktop environment, for the two language pairs.

Finally, we consider the Europarl corpus v7 (Koehn, 2005), given it is widely used in the MT community, for Spanish–English.

All the datasets are pre-processed as follows. First they are tokenised and truecased with Moses' (Koehn et al., 2007) scripts. Truecasing is carried out with a model trained on the caWaC corpus for Catalan (Ljubešić and Toral, 2014) and News Crawl 2012[8] both for English and Spanish.

Parallel datasets not available in a sentence-split format (novel1 and novel2) are sentence-split using Freeling (Padró and Stanilovsky, 2012). All parallel datasets are then sentence aligned. We use Hunalign (Varga et al., 2005) and keep only one-to-one alignments. The dictionaries used for Spanish–Catalan and Spanish–English are extracted from Apertium bilingual dictionaries for those language pairs.[9,10] Only sentence pairs for which the confidence score of the alignment is $>= 0.4$ are kept.[11] Although most of the parallel datasets are provided in sentence-aligned form, we realign them to ensure that the data used to calculate word alignment perplexity are properly aligned at sentence level. This is to avoid having high word alignment perplexities due, not to high degrees of translation freedom, but to the presence of misaligned parallel data.

[4] http://www.statmt.org/wmt13/training-parallel-nc-v8.tgz

[5] http://opus.lingfil.uu.se/DOGC.php

[6] http://opus.lingfil.uu.se/EMEA.php

[7] http://opus.lingfil.uu.se/KDE4.php

[8] http://www.statmt.org/wmt13/translation-task.html

[9] http://sourceforge.net/projects/apertium/files/apertium-es-ca/1.2.1/

[10] http://sourceforge.net/projects/apertium/files/apertium-en-es/0.8.0/

[11] Manual evaluation for English, French and Greek concluded that 0.4 was an adequate threshold for Hunalign's confidence score (Pecina et al., 2012).

### 3.2 Narrowness of the Domain

As previously mentioned, we use LM perplexity as a proxy to measure the narrowness of the domain.

We take two random samples without replacement for the Spanish side of each dataset, to be used for training (200,000 tokens) and testing (20,000 tokens). We train an LM of order 3 and improved Kneser-Ney smoothing (Chen and Goodman, 1996) with IRSTLM (Federico et al., 2008).

For each LM we report the perplexity on the testset built from the same dataset in Figure 1. The two novels considered (perplexities in the range $[230.61, 254.49]$) fall somewhere between news ($[359.73, 560.62]$) and technical domain ($[127.30, 228.38]$). Our intuition is that novels cover a narrow domain, like technical texts, but the vocabulary and language used in novels is richer, thus leading to higher perplexity than technical texts. News, on the contrary, covers a large variety of topics. Hence, despite novels possibly using more complex linguistic constructions, news articles are less predictable.

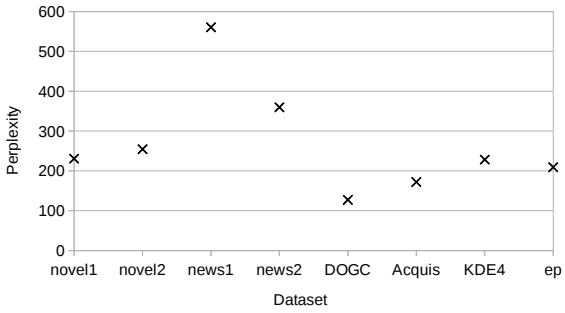


Figure 1: LM perplexity results

### 3.3 Degree of Translation Freedom

We use word alignment perplexity, as in Equation 1, as a proxy to measure the degree of translation freedom. Word alignment perplexity gives an indication of how well the model fits the data.

$$\log_2 PP = -\sum_s \log_2 p(e_s|f_s) \qquad (1)$$

The assumption is that the freer the translations are for a given parallel corpus, the higher the per-

plexity of the word alignment model learnt from such dataset, as the word alignment algorithms would have more difficulty to find suitable alignments.

For each parallel dataset, we randomly select a set of sentence pairs whose overall size accounts for 500,000 tokens. We then run word alignment with GIZA++ (Och and Ney, 2003) in both directions, with the default parameters used in Moses.

For each dataset and language pair, we report in Figure 2 the perplexity of the word alignment after the last iteration for each direction. The most important discriminating variable appears to be the level of relatedness of the languages involved, i.e. all the perplexities for Spanish–Catalan are below 10 while all the perplexities for Spanish–English are well above this number.

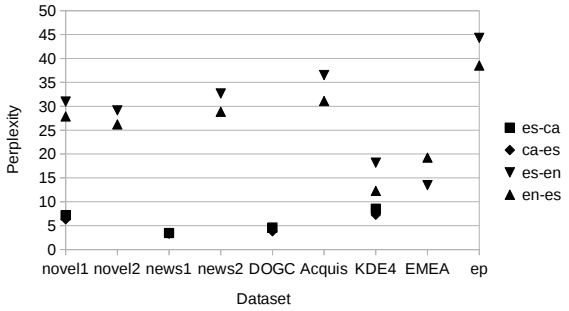


Figure 2: Word alignment perplexity results

## 4 MT for Literature between Related Languages

Encouraged by the results obtained for the translatability of novels (cf. Figures 1 and 2), we decided to carry out an experiment to assess the feasibility of using MT to assist with the translation of novels between closely-related languages. In this experiment we translate a novel, *The Prisoner of Heaven* (2011) by Carlos Ruiz Zafón, from Spanish into Catalan. This language pair is chosen because of the maturity of applied MT technology, e.g. MT is used alongside post-editing to translate the newspaper *La Vanguardia* (around 70,000 tokens) from Spanish into Catalan on a daily basis (Martín and Serra, 2014). We expect the results to be similar for other languages with similar degrees of similarity to Spanish, e.g. Portuguese and Italian.

| Type | Dataset | # sentences | Avg length |
|---|---|---|---|
| TM | News | 629,375 | 22.45 |
| | | | 21.49 |
| | Novel | 21,626 | 16.95 |
| | | | 15.11 |
| LM | News1 | 631,257 | 22.66 |
| | caWaC | 16,516,799 | 29.48 |
| | Novel | 22,170 | 17.14 |
| Dev | News | 1,000 | 22.31 |
| | | | 21.36 |
| | Novel | 1,000 | 16.92 |
| | | | 15.35 |
| Test | Novel | 1,000 | 17.91 |
| | | | 15.93 |

Table 1: Datasets used for MT

The translation model (TM) of our baseline system is trained with the news1 dataset while the LM is trained with the concatenation of news1 and caWaC. The baseline system is tuned on news. On top of this baseline we then build our domain-adapted systems. The domain adaptation is carried out by using two previous novels from the same author that were translated by the same translator (cf. the dataset novel1 in Section 3.1). We explore their use for tuning (+inDev), LM (concatenated +inLM and interpolated +IinLM) and TM (concatenated +inTM and interpolated +IinTM). The testset is made of a set of randomly selected sentence pairs from *The Prisoner of Heaven*. Table 1 provides an overview of the datasets used for MT.

We train phrase-based SMT systems with Moses v2.1 using default parameters. Tuning is carried out with MERT (Och, 2003). LMs are linearly interpolated with SRILM (Stolcke et al., 2011) by means of perplexity minimisation on the development set from the novel1 dataset. Similarly, TMs are linearly interpolated, also by means of perplexity minimisation (Sennrich, 2012).

### 4.1 Automatic Evaluation

Our systems are evaluated with a set of state-of-the-art automatic metrics: BLEU (Papineni et al., 2002), TER (Snover et al., 2006) and METEOR 1.5 (Denkowski and Lavie, 2014).

Table 2 shows the results obtained by each of the systems built. For each domain-adapted system

| System | BLEU | diff | TER | diff | METEOR | diff |
|---|---|---|---|---|---|---|
| baseline | 0.4915 | | 0.3658 | | 0.3612 | |
| +inDev | 0.4939 | 0.49% | 0.3641 | -0.47% | 0.3628 | 0.46% |
| +inDev+inLM | 0.4948 | 0.67% | 0.3643 | -0.41% | 0.3633 | 0.59% |
| +inDev+IinLM | 0.5045 | 2.64% | 0.3615 | -1.18% | 0.3669 | 1.59% |
| +inDev+inTM | 0.5238 | 6.57% | 0.3481 | -4.85% | 0.3779 | 4.61% |
| +inDev+IinTM | 0.5258 | 6.98% | 0.3510 | -4.04% | 0.3795 | 5.06% |
| +inDev+inLM+inTM | 0.5297 | 7.77% | 0.3433 | -6.17% | 0.3811 | 5.51% |
| +inDev+IinLM+IinTM | 0.5376 | 9.38% | 0.3405 | -6.92% | 0.3847 | 6.50% |
| inDev+inTM+inLM | 0.4823 | -1.87% | 0.3777 | 3.24% | 0.3594 | -0.49% |

Table 2: Automatic evaluation scores for the MT systems built

| System | BLEU | diff | TER | diff | METEOR | diff |
|---|---|---|---|---|---|---|
| Google | 0.4652 | 15.56% | 0.4021 | -15.31% | 0.3498 | 9.98% |
| Apertium | 0.4543 | 18.34% | 0.3925 | -13.25% | 0.3447 | 11.60% |
| Lucy | 0.4821 | 11.51% | 0.3758 | -9.40% | 0.3550 | 8.35% |

Table 3: Automatic evaluation scores for third-party MT systems

we show its relative improvement over the baseline (columns diff). The use of in-domain data to adapt each of the components of the pipeline, tuning (+inDev), LM (+inLM and +IinLM) and TM (+inTM and +IinTM), results in gains across all the metrics. Additional gains are achieved when combining the different in-domain components. Interpolation, both for LM and TM, results in gains when compared to the systems that use the same data in a concatenated manner (e.g. +IinLM vs +inLM) except for the TM in terms of TER. The best system, with in-domain data used for all the components and interpolated TM and LM (+inDev+IinLM+IinTM), yields a relative improvement over the baseline of 9.38% for BLEU, 6.92% for TER and 6.5% for METEOR. Finally we show the scores obtained by a system that uses solely in-domain data (inTM+inLM+inDev). While its results are slightly below those of the baseline, it should be noted that both the TM and TL of this system are trained with very limited amounts of data: 21,626 sentence pairs and 22,170 sentences, respectively (cf. Table 1).

We decided to compare our system also to widely-used on-line third-party systems, as these are the ones that a translator could easily have access to. We consider the following three systems: Google Translate,[12] Apertium (Forcada et al., 2011)[13] and Lucy.[14] These three systems follow different approaches; while the first is statistical, the second and the third are rule-based, classified respectively as shallow and deep formalisms.

Table 3 shows the results of the third-party system and compares their scores with our best domain-adapted system in terms of relative improvement (columns diff). The results of the third-party systems are similar, albeit slighly lower, compared to our baseline (cf. Table 2).

We conducted statistical significance tests for BLEU between our best domain-adapted system, the baseline and the three third-party systems using paired bootstrap resampling (Koehn, 2004) with 1,000 iterations and $p = 0.01$. In all cases the improvement brought by our best system is found out to be significant.

Finally we report on the percentage of translations that are equal in the MT output and in the reference. These account for 15.3% of the sentences for the baseline and 19.7% for the best domain-adapted system. It should be noted though that these tend to be short sentences, so if we consider their percentage in terms of words, they account for 4.97% and 7.15% of the data, respectively. If we consider also the translations that can reach the reference in at most five character editing steps (Volk, 2009), then the percentage of equal and near-equal translations pro-

[12] https://translate.google.com

[13] http://apertium.org/

[14] http://www.lucysoftware.com/english/machine-translation/

duced by our best domain-adapted system reaches 29.5% of the sentences.

### 4.2 Manual Evaluation

To gain further insight on the results, we conducted a manual evaluation. A common procedure (e.g. conducted in the MT shared task at WMT) consists of ranking MT translations. Given the source and target sides of the reference (human) translations, and two or more outputs from MT systems, these outputs are ranked according to their quality, i.e. how close they are to the reference, e.g. in terms of adequacy and/or fluency.

In our experiment, we are of course not interested in comparing two MT systems, but rather one MT system (the best one according to the automatic metrics) and the human translation. Hence, we conduct the rank-based manual evaluation in a slightly modified setting; we do not provide the target of the reference translation as reference but as one of the MT systems to be ranked. The evaluator thus is given the source-side of the reference and two translations, one being the human translation and the other the translation produced by an MT system. The evaluator of course does not know which is which. Moreover, in order to avoid any bias with respect to MT, they do not know that one of them has been produced by a human.

Two bilingual speakers in Spanish and Catalan, with a background in linguistics but without in-depth knowledge of MT (again, to avoid any bias with respect to MT) ranked a set of 101 translations. We carried out this rank-based evaluation with the Appraise tool (Federmann, 2012), using its 3-way ranking task type, whereby given two translations A and B, the evaluator can rank them as A>B (if A is better than B), A<B (if A is worse than B) and A=B (if both are of the same quality). Here we reproduce verbatim the evaluation instructions given to the evaluators:

*"Given the translations by two machine translation systems A and B, the task is to rank them:*
*- Rank A higher than B (A>B) if the output of system A is better than the output of system B*
*- Rank A lower than B (A<B) if the output of system A is worse than the output of system B*
*- Rank both systems equally (A=B) if the outputs of both systems are of an equivalent level of quality"*

The inter-annotator agreement, in terms of Fleiss' Kappa (Fleiss, 1971), is 0.49, which falls in the band of moderate agreement $[0.41, 0.60]$ (Landis and Koch, 1977).

Considering the aggregated 202 judgements, we have the breakdown shown in Table 4. In most cases (41.58% of the judgements), both the human translation (HT) and the MT are considered to be of equal quality. The HT is considered better than MT in 39.11% of the cases. Perhaps surprisingly, the evaluators ranked MT higher than HT in almost 20% of their judgements.

| Judgement | Times | Percentage |
|---|---|---|
| HT=MT | 84 | 41.58% |
| HT<MT | 39 | 19.31% |
| HT>MT | 79 | 39.11% |

Table 4: Manual Evaluation. Breakdown of ranks (overall)

We now delve deeper into the results and show in Table 5 the breakdown of judgements by evaluator. For around two thirds of the sentences, both evaluators agreed in their judgement: in 28.71% of the sentences both for HT=MT and for HT>MT, and in 9.9% of the sentences for HT<MT. They disagreed in the remaining one third of the data, the two main disagreements being between HT=MT and HT>MT (13.86%) and between HT=MT and HT<MT (11.88%). The remaining case of disagreement (between HT>MT and HT<MT) is encountered less frequently (6.93%).

| Judgement | Times | Percentage |
|---|---|---|
| HT=MT, HT=MT | 29 | 28.71% |
| HT<MT, HT<MT | 10 | 9.9% |
| HT>MT, HT>MT | 29 | 28.71% |
| Total | 68 | 67.33% |
| HT>MT, HT<MT | 7 | 6.93% |
| HT=MT, HT>MT | 14 | 13.86% |
| HT=MT, HT<MT | 12 | 11.88% |
| Total | 33 | 32.67% |

Table 5: Manual Evaluation. Breakdown of ranks (per evaluator)

We analyse the sets of sentences where both evaluators agree, for HT=MT, HT<MT and HT>MT. First, we report on their average sentence length in tokens, as shown in Table 6. We can conclude that

| | |
|---|---|
| Source | La habitación tenía un pequeño balcón que daba a la plaza. |
| Gloss | *The room had a small balcony facing the square.* |
| HT | La cofurna tenia un balconet que donava a la plaça. |
| MT | L'habitació tenia un petit balcó que donava a la plaça. |
| Discussion | Habitació (room) is the translation of habitación. |
| | Cofurna (hovel) has slightly different meaning. |
| Source | — ¿Adónde vas? |
| Gloss | — *Where are you going?* |
| HT | — ¿On vas? — hi vaig afegir. |
| MT | — ¿On vas? |
| Discussion | The snippet "hi vaig afegir" (I added) is not in the original. |

Table 7: Manual Evaluation. Examples of translations ranked as HT<MT

| Mode | Sentences | Tokens | Tokens/sent. |
|---|---|---|---|
| HT<MT | 10 | 127 | 12.7 |
| HT=MT | 29 | 278 | 9.59 |
| HT>MT | 29 | 657 | 22.66 |
| whole | 101 | 1,671 | 16.71 |

Table 6: Manual Evaluation. Avg sentence length per rank

MT results in translations of good quality for shorter sentences than the average, while HT remains the best translation for longer sentences.

We now look at each of these sets of sentences and carry out a qualitative analysis, aiming at finding out what types of sentences and translation errors are predominant.

For most of the HT=MT cases (22), both translations are exactly the same. In the remaining 7 cases, up to a few words are different, with both translations being accurate.

In most of the 10 sentences ranked as HT<MT, the translator has either added some content that is not in the original or has used words that have a slightly different meaning than the corresponding words in the original, while the MT translation is accurate. Table 7 provides examples of both cases.

Finally, regarding the translations ranked as HT>MT, the translation as produced by the MT systems has some errors, in most cases affecting just one or a few words. The most common errors are related to:

- OOVs, mainly for verbs that contain a pronoun as a suffix in Spanish. E.g. "escrutándola" (*scrutinising her*).
- Pronouns translated wrongly. E.g. "lo" (him) wrongly translated as "ho" (*that*) instead of "el".
- Word choice. Either the translation looks unnatural or its meaning is closely related to the original but it is not exactly the same.

## 5 Conclusions and Future Work

This paper has explored the feasibility of applying MT to the translation of literary texts. To that end, we measured the translatability of literary texts and compared it to that of other datasets commonly used in MT by measuring the degree of freedom of the translations (using word alignment perplexity) and the narrowness of the domain (via LM perplexity). Our results show that novels are less predictable than texts in the technical domain but more predictable than news articles. Regarding translation freedom, the main variable is not related to the type of data but to the level of relatedness of the pair of languages involved.

Furthermore, we explored the use of state-of-the-art domain adaptation techniques in MT to translate a novel between two closely-related languages, Spanish and Catalan. This is the first time that specific MT systems are built to translate novels. Our best domain-adapted system outperforms a strong baseline by 4.61 absolute points (9.38% relative) in terms of BLEU. We provided evidence that MT can be useful to assist with the translation of novels between closely-related languages, namely (i) the translations produced by our best system are equal to the ones produced by a professional human translator in almost 20% of cases, with an additional 10% requiring at most 5 character edits, and (ii) over 60%

of the translations are perceived to be of the same (or even higher) quality by native speakers.

As this is the first work where a specific MT system has been built to translate novels, a plethora of research lines remain to be explored. In this work we have adapted an MT system by learning from previous novels from the same author. A further step would be to learn from translators while they are translating the current novel using incremental retraining techniques. We would like to experiment with less related language pairs (e.g. Spanish–English) to assess whether the current setup remains useful. As pointed out by Voigt and Jurafsky (2012), and corroborated by our manual evaluation (some of the MT errors are due to mistranslation of pronouns), we would like to explore using discourse features. Finally, as the ultimate goal of this work is to integrate MT in the translation workflow to assist with the translation of literature, we would like to study which is the best way of doing so, e.g. by means of post-editing, interactive MT, etc, with real customers.

## Acknowledgments

This research is supported by the European Union Seventh Framework Programme FP7/2007-2013 under grant agreement PIAP-GA-2012-324414 (AbuMaTran) and by Science Foundation Ireland through the CNGL Programme (Grant 12/CE/I2267) in the ADAPT Centre (www.adaptcentre.ie) at Dublin City University.

## References

Laurent Besacier. 2014. Traduction automatisée d'une oeuvre littéraire: une étude pilote. In *Traitement Automatique du Langage Naturel (TALN)*, Marseille, France.

P. Brown, J. Cocke, S. Della Pietra, V. Della Pietra, F. Jelinek, R. Mercer, and P. Roossin. 1988. A statistical approach to language translation. In *Proceedings of the 12th Conference on Computational Linguistics - Volume 1*, COLING '88, pages 71–76, Stroudsburg, PA, USA. Association for Computational Linguistics.

Peter F. Brown, John Cocke, Stephen A. Della Pietra, Vincent J. Della Pietra, Fredrick Jelinek, John D. Lafferty, Robert L. Mercer, and Paul S. Roossin. 1990. A statistical approach to machine translation. *Comput. Linguist.*, 16(2):79–85, June.

Stanley F. Chen and Joshua Goodman. 1996. An empirical study of smoothing techniques for language modeling. In *Proceedings of the 34th Annual Meeting on Association for Computational Linguistics*, ACL '96, pages 310–318, Stroudsburg, PA, USA. Association for Computational Linguistics.

Michael Denkowski and Alon Lavie. 2014. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the EACL 2014 Workshop on Statistical Machine Translation*.

Marcello Federico, Nicola Bertoldi, and Mauro Cettolo. 2008. Irstlm: an open source toolkit for handling large scale language models. In *INTERSPEECH*, pages 1618–1621.

Christian Federmann. 2012. Appraise: An open-source toolkit for manual evaluation of machine translation output. *The Prague Bulletin of Mathematical Linguistics*, 98:25–35, September.

Joseph L. Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382.

Mikel L. Forcada, Mireia Ginestí-Rosell, Jacob Nordfalk, Jim O'Regan, Sergio Ortiz-Rojas, Juan Antonio Pérez-Ortiz, Felipe Sánchez-Martínez, Gema Ramírez-Sánchez, and Francis M. Tyers. 2011. Apertium: A free/open-source platform for rule-based machine translation. *Machine Translation*, 25(2):127–144, June.

Dmitriy Genzel, Jakob Uszkoreit, and Franz Och. 2010. "Poetic" Statistical Machine Translation: Rhyme and Meter. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, EMNLP '10, pages 158–166.

Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. 2014. Is machine translation getting better over time? In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 443–451, Gothenburg, Sweden.

Erica Greene, Tugba Bodrumlu, and Kevin Knight. 2010. Automatic analysis of rhythmic poetry with applications to generation and translation. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, EMNLP '10, pages 524–533, Stroudsburg, PA, USA. Association for Computational Linguistics.

Jing He, Ming Zhou, and Long Jiang. 2012. Generating chinese classical poems with statistical machine translation models.

Long Jiang and Ming Zhou. 2008. Generating Chinese Couplets Using a Statistical MT Approach. In *Proceedings of the 22nd International Conference on Computational Linguistics - Volume 1*, pages 377–384.

Ruth Jones and Ann Irvine, 2013. *Proceedings of the 7th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, chapter The (Un)faithful Machine Translator, pages 96–101. Association for Computational Linguistics.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, pages 177–180.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing , EMNLP 2004, A meeting of SIGDAT, a Special Interest Group of the ACL, held in conjunction with ACL 2004, 25-26 July 2004, Barcelona, Spain*, pages 388–395. ACL.

Philipp Koehn. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. In *Conference Proceedings: the tenth Machine Translation Summit*, pages 79–86, Phuket, Thailand. AAMT, AAMT.

R. J. Landis and G. G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159 – 174.

Nikola Ljubešić and Antonio Toral. 2014. caWaC - a Web Corpus of Catalan and its Application to Language Modeling and Machine Translation. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland.

Juan Alberto Alonso Martín and Anna Civil Serra. 2014. Integration of a machine translation system into the editorial process flow of a daily newspaper. *Procesamiento del Lenguaje Natural*, 53(0):193–196.

Franz Josef Och and Hermann Ney. 2003. A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*, 29(1):19–51.

Franz Josef Och. 2003. Minimum Error Rate Training in Statistical Machine Translation. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics*, pages 160–167.

Lluís Padró and Evgeny Stanilovsky. 2012. Freeling 3.0: Towards wider multilinguality. In *Proceedings of the Language Resources and Evaluation Conference (LREC 2012)*, Istanbul, Turkey, May. ELRA.

Kishore Papineni, 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.

Pavel Pecina, Antonio Toral, Vassilis Papavassiliou, Prokopis Prokopidis, and Josef van Genabith. 2012. Domain adaptation of statistical machine translation using web-crawled resources: a case study. In *Proceedings of the 16th Annual Conference of the European Association for Machine Translation*, pages 145–152.

Nicholas Ruiz and Marcello Federico. 2014. Complexity of spoken versus written language for machine translation. In *17th Annual Conference of the European Association for Machine Translation*, EAMT, pages 173–180, Dubrovnik, Croatia.

Rico Sennrich. 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.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and Ralph Weischedel. 2006. A Study of Translation Error Rate with Targeted Human Annotation. In *Proceedings of the Association for Machine Translation in the Americas*.

Ralf Steinberger, Bruno Pouliquen, Anna Widiger, Camelia Ignat, Tomaz Erjavec, Dan Tufis, and Dániel Varga. 2006. The jrc-acquis: A multilingual aligned parallel corpus with 20+ languages. *CoRR*, abs/cs/0609058.

Andreas Stolcke, Jing Zheng, Wen Wang, and Victor Abrash. 2011. SRILM at Sixteen: Update and Outlook. In *Proceedings of ASRU*.

Dániel Varga, László Németh, Péter Halácsy, András Kornai, Viktor Trón, and Viktor Nagy. 2005. Parallel corpora for medium density languages. In *Recent Advances in Natural Language Processing*, pages 590–596, Borovets, Bulgaria.

Rob Voigt and Dan Jurafsky, 2012. *Proceedings of the NAACL-HLT 2012 Workshop on Computational Linguistics for Literature*, chapter Towards a Literary Machine Translation: The Role of Referential Cohesion, pages 18–25.

Martin Volk. 2009. The automatic translation of film subtitles. A machine translation success story? *JLCL*, 24(3):115–128.

Dekai Wu, Karteek Addanki, and Markus Saers. 2013. Modelling hip hop challenge-response lyrics as machine translation. In *Machine Translation Summit XIV*, pages 109–116, Nice, France.

Ventsislav Zhechev. 2012. Machine Translation Infrastructure and Post-editing Performance at Autodesk. In *AMTA 2012 Workshop on Post-Editing Technology and Practice (WPTP 2012)*, pages 87–96, San Diego, USA.